

Monica Bianchini

Vincenzo Piuri

Sanjoy Das

Rabindra Nath Shaw *Editors*

Advanced Computing and Intelligent Technologies

Proceedings of ICACIT 2021



Springer

Lecture Notes in Networks and Systems

Volume 218

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, Turkey

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.

More information about this series at <http://www.springer.com/series/15179>

Monica Bianchini · Vincenzo Piuri · Sanjoy Das ·
Rabindra Nath Shaw
Editors

Advanced Computing and Intelligent Technologies

Proceedings of ICACIT 2021



Springer

Editors

Monica Bianchini
Department of Information Engineering
and Mathematics
University of Siena
Siena, Italy

Sanjoy Das
Regional Campus Manipur
Indira Gandhi National Tribal University
Imphal, Manipur, India

Vincenzo Piuri
Department of Computer Science
University of Milan
Milan, Italy

Rabindra Nath Shaw
School of Electrical and Electronic
Engineering
Galgotias University
Greater Noida, Uttar Pradesh, India

ISSN 2367-3370

ISSN 2367-3389 (electronic)

Lecture Notes in Networks and Systems

ISBN 978-981-16-2163-5

ISBN 978-981-16-2164-2 (eBook)

<https://doi.org/10.1007/978-981-16-2164-2>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022, corrected publication 2022

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

ICACIT 2021 Organizing Committee

Patrons

Suneel Galgotia, Chancellor, Galgotias University, India

Andrea Garulli, Department of Information Engineering and Mathematics,
University of Siena, Italy

Dhruv Galgotia, CEO, Galgotias University, India

R. Venkatesh Babu, PVC, Galgotias University, India

Munish Sabharwal, Dean, School of Computing Science & Engineering, Galgotias
University, India

General Chair

Monica Bianchini, Department of Information Engineering and Mathematics,
University of Siena, Italy

Honorary Chairs

Bhim Singh, Indian Institute of Technology, Delhi

Saad Mekhilef, Dean (Faculty of Engineering), University of Malaya, Malaysia

Conference Chair and Chairman, Oversight Committee

Rabindra Nath Shaw, Galgotias University, India

Conference Secretaries

Saravanan D., Galgotias University, India
Indu, Galgotias University, India

Technical Chairs

Nishad Mendis, Det Norske Veritas, Australia
Ankush Ghosh, TNU, India
Rohit Tripathi, Galgotias University, India

Publication Chairs

Vincenzo Piuri, FIEEE, Professor, University of Milan, Italy
Valentina E Balas, Aurel Vlaicu University of Arad, Romania
Sanjoy Das, IGNTU, India

Publicity Chairs

Prashant R. Nair, Amrita Vishwa Vidyapeetham, India
Priyanka Singh, Amity University, India

Springer/ICACIT Liaison

Aninda Bose, Senior Editor, Springer Nature

International Advisory Board

Ajay Gupta, IEEE Computer Society, USA
Mohammad S. Obaidat, Fellow of IEEE University of Sharjah, UAE
Fawnizu Azmadi Hussin, Chair, IEEE Malaysia Section
Dharmendra Sharma, Chairman, University Academic Board, University of Canberra
Maria Virvou, University of Piraeus, Greece
Marcin Paprzycki, Polish Academy of Sciences, Poland

Laxmi C. Jain, Co founder, KES International, UK
Celia Shahnaz, Chairperson, IEEE Bangladesh Section
Anna Esposito, Seconda Università di Napoli, Italy
Sanjeevikumar Padmanaban, Aalborg University Esbjerg, Denmark
Atiqur Rahman Ahad, University of Osaka, Japan
George T., University of Piraeus, Greece
Margarita N. Favorskaya, Reshetnev Siberian State University
Urszula Stańczyk, Silesian University of Technology (SUT), Gliwice, Poland
Valentina Balas, University of Arad, Romania
N. R. Pal, President, IEEE CIS
Yen-Wei Chen, Ritsumeikan University, Japan
Milan Simic, RMIT University, Australia

Technical Program Committee and Reviewers

Brijesh Iyer
Meenakshi Bhardwaj
Anshu Parashar
Maya L. Pai
Vinod Kumar
Vijayakumar Ponnusamy
Sreeja P. S.
Rajashree Narendra
Saru Kumari
Mallikharjuna Rao K.
R. Ramanathan
Rishibrind Kumar Upadhyay
Jasminder Sandhu
A. S. M. Touhidul Hasan
Arvind Dagur
Hussain Mahdi
Suneet Kumar
S. Srinivasulu Raju
Pinaki Chakraborty
Preeti Rai
Sandeep Mathur
Sachin Goyal
Amit Jain
Rajesh Prasad
Kapil Gupta
Nuparam Chauhan
Jitali Patel
Ankit Saxena

Prashant Nair
Vishwesh Laxmikant Akre
Kamlesh Lakhwani
Irfan Siddavatam
Malaya Nath
Era Johri
Abhishek Dubey
Arun K. Singh
Karthikeyan B.
Rohit Tripathi
Krishnananda Shet
Indrani Das
Shrawan Kumar
Jayshree Pande
Aditi Paul
Mahesh Pawar
Priyanka Singh
Shrikant Sonekar
Nuzhat Shaikh
Drmandeep Kaur
Keshav Gupta
Sanjeev Rana
Pranav Kumar Singh
Joe Louis Paul I.
Venugopala P. S.
Ashwini B.
Ebha Koley
Mohd Sadiq
Suraiya Jabin
Pawan Kumar
Dalia Nandi
S. Vijay Kumar
Sameena Naaz
Kedar Nath Sahu
Mohit Tiwari
Ashwini Kumar
Harshal Shah
Sunil Singh
Surbhi Gupta
Jay Kumar Jain
Vijayalakshmi Kakulapati
Kirthiga S.
Alka Leekha
Sanjeev Kumar
Jawahar P. K.

Shyamal Mondal
Prachi Garg
Krishna Kumar Singh
Sartajvir Singh
M. T. L. Gayatri
Md. Forhad Rabbi
Rohit Raja
Vinay Goyal
Prashantkumar Vats
Mansaf Alam
Parul Dawar
Rajendra Patil
Prem Chand
Jinesh M. K.
Hardeo Kumar Thakur
Abhijit Lahiri
Shailendra Dwivedi
Vandana Niranjan
Akshat Agrawal
Anup Kumarbarman
Sarita Kumari
Ahmed Ali Shah
Anjana Pandey

Preface

This book gathers selected high-quality research papers presented at the 2021 International Conference on Advanced Computing and Intelligent Technologies (ICACIT), held at NCR New Delhi, India, during March 20–21, 2021, jointly organized by Galgotias University, India, and Department of Information Engineering and Mathematics Università Di Siena, Italy. It discusses emerging topics pertaining to Advanced Computing, Intelligent Technologies and Networks including AI and Machine Learning, Data Mining, Big Data Analytics, High Performance Computing Network Performance Analysis, Internet of Things Networks, Wireless Sensor Networks, and others. Written by respected experts and researchers working on Computing and Intelligent Technologies, the book offers a valuable asset for researchers both from academia and industries involved in advanced studies.

We are thankful to all the authors that have submitted papers for keeping the quality of the ICACIT 2021 at high levels. The editors of this book would like to acknowledge all the authors for their contributions and the reviewers. We have received invaluable help from the members of the International Program Committee and the chairs responsible for different aspects of the workshop. We also appreciate the role of Special Session Organizers. Thanks to all of them, we had been able to collect many papers on interesting topics, and during the conference, we had very interesting presentations and stimulating discussions.

Our special thanks go to Janus Kacprzyk (Editor in Chief, Springer, Advances in Intelligent Systems and Computing Series) for the opportunity to organize this guest-edited volume.

We are grateful to Springer, especially to Dr. Thomas Ditzinger (Senior Editor, Applied Sciences and Engineering, Springer-Verlag), for the excellent collaboration, patience, and help during the evolution of this volume.

We hope that the volume will provide useful information to professors, researchers, and graduated students in the area of soft computing techniques and applications, and everyone will find this collection of papers inspiring, informative, and useful. We also hope to see you at a future ICACIT event.

Siena, Italy
Milan, Italy
Imphal, India
Greater Noida, India

Monica Bianchini
Vincenzo Piuri
Sanjoy Das
Rabindra Nath Shaw

Contents

Recent Trends in Electromyography Signal Processing of Neuromuscular Diseases: An Outlook	1
M. Emimal, W. Jino Hans, T. M. Inbamalar, and N. Mahiban Lindsay	
Conceptualization, Visualization, and Modeling of Ontologies for Elementary Kinematics	15
C. S. Nandhakishore, Gerard Deepak, and A. Santhanavijayan	
A Deep Neural Approach Toward Staining and Tinting of Monochrome Images	25
Debosmit Neogi, Nataraj Das, and Suman Deb	
Evaluation of IoT Based Automatic Headlight Dimmer Systems	37
Kanika Gandhi, Karanpartap Singh Aulakh, Jobanpreet Singh Thind, Gurpreet Singh Kharoud, and Seemu Sharma	
Enhanced Energy-Efficient Fuzzy Logic Clustering and Network Coding Strategy for Wireless Sensor Networks (EEE-FL-NC)	59
K. S. Fathima Shemim and Dr. Ulf Witkowski	
Malware Classification Using Automated Transmutation and CNN	73
Ritu Agarwal, Saurabh Patel, Sparsh Katiyar, and Sharad Nailwal	
Content-Based Image Retrieval Using Energy-Based Frequency Domain Features	83
Hillol Barman, Netalkar Rohan Kishor, U. S. N. Raju, Debanjan Pathak, and Sweta Panigrahi	
Decision Tree-Based Event Detection Framework for UWSN Routing to Optimize Energy Consumption During Transmission	99
Rakesh Kumar and Diwakar Bhardwaj	

Recognizing Child Unsafe Apps Through User Reviews on the Google Play Store	111
Ashwini Dalvi, Irfan Siddavatam, Viraj Thakkar, Aditya Vedpathak, and Abhishek Patel	
Brain Tumor Classification Using Deep Learning and Big Data Analytics	121
V. P. Arathi, Gayathri Suresh, T. H. Harikrishna, and A. G. Hari Narayanan	
Drone Stability Simulation Using ROS and Gazebo	131
Rajesh Kannan Megalingam, Darla Vineeth Prithvi, Nimmala Chaitanya Sai Kumar, and Vijay Egumadiri	
Knowledge Management Framework for the Supervision of IT Postgraduate Research in Sri Lanka	145
W. M. J. H. Fernando and M. P. A. W. Gamage	
RFM-Based Customer Analysis and Product Recommendation System	159
Rahul Krishnan and Prashant R. Nair	
Efficient Hardware Trojan Detection Using Generic Feature Extraction and Weighted Ensemble Method	165
Vaishnavi Sankar and M. Nirmala Devi	
An Approach for Offline Handwritten Character Shape Reconstruction Using Active Contour and Morphological Techniques	183
Anupam Garg, Amrita Kaur, and Anshu Parashar	
Stock Direction Prediction Using Sentiment Analysis of News Articles	195
Nipun Jain, Mohit Motiani, and Preeti Kaur	
Analysis and Prediction of COVID-19 Confirmed Cases Using Deep Learning Models: A Comparative Study	207
Trisha Sinha, Titash Chowdhury, Rabindra Nath Shaw, and Ankush Ghosh	
Machine Learning Technique to Prognosis Diabetes Disease: Random Forest Classifier Approach	219
Prajyot Palimkar, Rabindra Nath Shaw, and Ankush Ghosh	
A Low Cost and Enhanced Assistive Environment for People with Vision Loss	245
Tushar Vashisth, Ritika Khareta, Nishi Bhati, Venkata Chanakya Samsani, and Shalu Sharma	

A Comparative Study of Myocardial Infarction Detection from ECG Data Using Machine Learning	257
Aritra Chakraborty, Santanu Chatterjee, Koushik Majumder, Rabindra Nath Shaw, and Ankush Ghosh	
Construction of Effective Wireless Sensor Network for Smart Communication Using Modified Ant Colony Optimization Technique	269
Avishek Banerjee, Sudip Kumar De, Koushik Majumder, Victor Das, Debasis Giri, Rabindra Nath Shaw, and Ankush Ghosh	
A Framework for Personalizing Atypical Web Search Sessions with Concept-Based User Profiles Using Selective Machine Learning Techniques	279
Pradeep Bedi, S. B. Goyal, Anand Singh Rajawat, Rabindra Nath Shaw, and Ankush Ghosh	
Smart Luminaires for Commercial Building by Application of Daylight Harvesting Systems	293
S. B. Goyal, Pradeep Bedi, Anand Singh Rajawat, Rabindra Nath Shaw, and Ankush Ghosh	
Multi-objective Fuzzy-Swarm Optimizer for Data Partitioning	307
S. B. Goyal, Pradeep Bedi, Anand Singh Rajawat, Rabindra Nath Shaw, and Ankush Ghosh	
Efficient Deep Learning for Reforming Authentic Content Searching on Big Data	319
Anand Singh Rajawat, Kanishk Barhanpurkar, S. B. Goyal, Pradeep Bedi, Rabindra Nath Shaw, and Ankush Ghosh	
IoT as a Platform: For Smart Home Analysis and Monitoring of Fire Parameters	329
Sudip Suklabaidya and Indrani Das	
Death Prediction in the Current Pandemic Scenario and Cluster Classification Using Soft Computing Techniques	339
Loshima Lohi and Maya L. Pai	
Taala Classification in Carnatic Music Using Machine Learning Algorithms and Deep Neural Networks	355
Amrutha Jayakumar and Maya L. Pai	
Utilizing the Data Mining Techniques for Obesity Prognosis Based on Eating and Lifestyle Routines of Adolescents and Adults	373
P. Vineetha Sankar and K. Sreekumar	
An Optimization of Feature Selection for Classification Using Modified Bat Algorithm	389
V. Yasaswini and Santhi Baskaran	

A Low-Cost Web Interface for Object Tracking Based on a Wireless Sensor Network	401
Juan P. Carvajal, Arturo Fajardo, and Carlos Paez	
Sentiment Analysis: Choosing the Right Word Embedding for Deep Learning Model	417
Sarita Bansal Garg and V. V. Subrahmanyam	
Detection and Localization of Unmanned Aerial Vehicles Based on Radar Technology	429
Sally M. Idhis, Takwa Dawdi, Qassim Nasir, Manar Abu Talib, and Yara Omran	
Prediction of Parkinson's Disease Using Machine Learning Models—A Classifier Analysis	453
A. T. Rohit Surya, P. Yaswanthram, Prashant R. Nair, S. S. Rajendra Prasath, and Sundeep V. V. S. Akella	
Social Distancing and Crowd Density Distribution System for Public Places and Public Transports Using Computer Vision and NLP	461
Sandeep K. Sharma, Rajiv K. Modanval, Prakhar Tayal, and N. Gayathri	
TV Viewing Behaviour: Analysis Using Machine Learning Algorithms	481
C. Karthika, A. G. Hari Narayanan, and P. P. Vijayalakshmi	
Inverse Kinematics of Robot Manipulator Integrated with Image Processing Algorithms	493
Rajesh Kannan Megalingam, Santosh Tantravahi, Hemanth Sai Surya Kumar Tammana, Nagasai Thokala, Hari Sudarshan Rahul Puram, Naveen Samudrala, and Chennareddy Pavanth Kumar Reddy	
A Comprehensive Study on Workloads in Cloud Computing	505
K. Mallikharjuna Rao and Aravapalli Rama Satish	
Inset Feed Micro-Strip Patch Antenna for Communication Application Using CST	515
Priyanka Shah and Niraj Tevar	
Load Balancing Approach and Diminishing Impact of Malicious Node in Ad Hoc Networks	523
Shrikant V. Sonekar, Rohan Kokate, Manoj Titre, Aniket Bhoyar, Merajul Haque, and Sachin Patil	
ELEMENT: Text Extraction for the Dark Web	537
Ashwini Dalvi, Irfan Siddavatam, Apoorva Jain, Smit Moradiya, Faruk Kazi, and S. G. Bhirud	

Contents	xvii
Anomalies in Punjabi Language WordNet: An IndoWordNet Evaluation	553
Harjit Singh	
Supervised Machine Learning Strategies for Investigation of Weird Pattern Formulation from Large Volume Data Using Quantum Computing	569
Mukta Nivelkar and S. G. Bhirud	
A Comparative Analysis of Intelligent Classifiers for Seizure Detection Using EEG Signals	577
Arshdeep Singh, Debargho Basak, Upamanyu Das, Priya Chugh, and Jyoti Yadav	
Adaptive Fuzzy Logic Models for the Prediction of Compressive Strength of Sustainable Concrete	593
Chakshu Garg, Aman Namdeo, Abhishek Singhal, Priyanka Singh, Rabindra Nath Shaw, and Ankush Ghosh	
Unique Action Identifier by Using Magnetometer, Accelerometer and Gyroscope: KNN Approach	607
Prajyot Palimkar, Varnica Bajaj, Arpan Kumar Mal, Rabindra Nath Shaw, and Ankush Ghosh	
Achieving Maximum Sum Spectral Efficiency with Channel Estimation	633
Ashu Taneja, Ankita Rana, and Nitin Saluja	
Correction to: Anomalies in Punjabi Language WordNet: An IndoWordNet Evaluation	C1
Harjit Singh	
Author Index	645

Editors and Contributors

About the Editors

Monica Bianchini received the Laurea cum laude in Mathematics and the Ph.D. degree in Computer Science from the University of Florence, Italy, in 1989 and 1995, respectively. After receiving the Laurea, for two years, she was involved in a joint project of Bull HN Italia and the Department of Mathematics (University of Florence), aimed at designing parallel software for solving differential equations. From 1992 to 1998, she was a Ph.D. student and a Postdoc Fellow with the Computer Science Department of the University of Florence. Since 1999, she has been with the University of Siena, where she is currently Associate Professor at the Information Engineering and Mathematics Department. Her main research interest is in the field of artificial intelligence & applications, machine learning, with emphasis on neural networks for structured data and deep learning, approximation theory, information retrieval, bioinformatics, and image processing. M. Bianchini has authored more than seventy papers and has been Editor of books and special issues on international journals in her research field. She has been a participant in many research projects focused on machine learning and pattern recognition, founded by both Italian Ministry of Education (MIUR) and University of Siena (PAR scheme), and she has been involved in the organization of several scientific events, including the NATO Advanced Workshop on Limitations and Future Trends in Neural Computation (2001), the 8th AI*IA Conference (2002), GIRPR 2012, the 25th International Symposium on Logic-Based Program Synthesis and Transformation, and the ACM International Conference on Computing Frontiers 2017. Prof. Bianchini served as Associate Editor for IEEE Transactions on Neural Networks (2003-09), Neurocomputing (from 2002), and Int. J. of Computers in Healthcare (from 2010). She is Permanent Member of the Editorial Board of IJCNN, ICANN, ICPR, ICPRAM, ESANN, ANNPR, and KES.

Vincenzo Piuri is Professor at the University of Milan, Italy (since 2000). He was Associate Professor at Politecnico di Milano, Italy, Visiting Professor at the University of Texas at Austin, USA, and Visiting Researcher at George Mason University, USA. He has founded a start-up company in the area of intelligent systems for

industrial applications and is active in industrial research projects. He received his M.S. and Ph.D. in Computer Engineering from Politecnico di Milano, Italy. His main research and industrial application interests are artificial intelligence, intelligent systems, computational intelligence, pattern analysis and recognition, machine learning, signal and image processing, biometrics, intelligent measurement systems, industrial applications, distributed processing systems, Internet of things, cloud computing, fault tolerance, application-specific digital processing architectures, and arithmetic architectures. He published over 400 papers in international journals, international conference proceedings, and books. He is Fellow of the IEEE and Distinguished Scientist of ACM. He has been IEEE Vice President for Technical Activities (2015), Member of the IEEE Board of Directors (2010–2012, 2015), and President of the IEEE Computational Intelligence Society (2006–2007). He is Editor-in-Chief of the IEEE Systems Journal (2013–2019).

Sanjoy Das is currently working as Associate Professor, Department of Computer Science, Indira Gandhi National Tribal University (A Central Government University), Amarkantak, M.P. (Manipur Campus), India. He received his Ph.D. in Computer Science from Jawaharlal Nehru University, New Delhi, India. Before joining IGNTU, he has worked as Associate Professor, School of Computing Science and Engineering, Galgotias University, India, from July 2012 to September 2017. Also, as Assistant Professor G. B. Pant Engineering College, Uttarakhand, and Assam University, Silchar, from 2001 to 2008. His current research interest includes mobile ad hoc networks and vehicular ad hoc networks, distributed systems, and data mining. He has published numerous papers in international journals and conferences including IEEE and Springer.

Rabindra Nath Shaw is a Senior Member of IEEE (USA), currently holding the post of Director, International Relations, Galgotias University India. Dr. Shaw is an alumnus of the applied physics department, University of Calcutta, India. He has more than eleven years of teaching experience in leading institutes like Motilal Nehru National Institute of Technology Allahabad, India, Jadavpur University, and others at UG and PG level. He has successfully organized more than fifteen International conferences as Conference Chair, Publication Chair, and Editor. He has published more than hundred Scopus/ WoS/ ISI indexed research papers in International Journals and Conference Proceedings. He is the editor of several Springer and Elsevier books. His primary area of research is optimization algorithms and machine learning techniques for power systems, IoT applications, Renewable Energy, and Power electronics converters. He also worked as University Examination Coordinator, University MOOC's Coordinator, University Conference Coordinator, and Faculty- In Charge, Centre of Excellence for Power Engineering and Clean Energy Integration.

Contributors

Ritu Agarwal Department of Information Technology, Delhi Technological University, Delhi, India

Sundeep V. V. S. Akella Department of Computer Science and Engineering, Amrita School of Engineering, Coimbatore, India

V. P. Arathi Department of Computer Science and IT, Amrita School of Arts and Sciences, Kochi, Kerala, India

Karanpartap Singh Aulakh Thapar Institute of Engineering and Technology, Patiala, Punjab, India

Varnica Bajaj School of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh, India

Avishek Banerjee Department of Information Technology, Asansol Engineering College, Asansol, India

Kanishk Barhanpurkar Department of Computer Science and Engineering, Sambhram Institute of Technology, Bengaluru, Karnataka, India

Hillol Barman Department of Computer Science and Engg, National Institute of Technology Warangal, Warangal, Telengana, India

Debargho Basak Netaji Subhas University of Technology, Dwarka, Delhi, India

Santhi Baskaran Information Technology Department, Pondicherry Engineering College, Puducherry, India

Pradeep Bedi Department of Computer Science Engineering, Lingaya's Vidyapeeth, Faridabad, Haryana, India

Diwakar Bhardwaj Department of Computer Engineering and Applications, GLA University, Mathura, India

Nishi Bhati Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida, India

S. G. Bhirud Veermata Jijabai Technological Institute, Matunga, Mumbai, India

Aniket Bhoyar JD College of Engineering and Management, Nagpur, Maharashtra, India

Juan P. Carvajal Pontificia Universidad Javeriana, Bogotá, Colombia

Aritra Chakraborty Maulana Abul Kalam Azad University of Technology, Kolkata, India

Santanu Chatterjee Maulana Abul Kalam Azad University of Technology, Kolkata, India

Titash Chowdhury School of Engineering and Applied Sciences, The Neotia University, Kolkata, West Bengal, India

Priya Chugh Netaji Subhas University of Technology, Dwarka, Delhi, India

Ashwini Dalvi Veermata Jijabai Technological Institute, Mumbai, India; K J Somaiya College of Engineering, Mumbai, India

Indrani Das Department of Computer Science, Assam University, Silchar, India

Nataraj Das Computer Science and Engg. Dept, National Institute of Technology Agartala, Jirania, India

Upamanyu Das Netaji Subhas University of Technology, Dwarka, Delhi, India

Victor Das Department of Information Technology, Asansol Engineering College, Asansol, India

Takwa Dawdi Department of Electrical Engineering, University of Sharjah, Sharjah, UAE

Sudip Kumar De Department of Information Technology, Asansol Engineering College, Asansol, India

Suman Deb Computer Science and Engg. Dept, National Institute of Technology Agartala, Jirania, India

Gerard Deepak Department of Computer Science and Engineering, National Institute of Technology, Tiruchirappalli, India

Vijay Egumadiri Department of Electrical Engineering, University of South Florida, Florida, USA

M. Emimal Anna University, Chennai, India

Arturo Fajardo Pontificia Universidad Javeriana, Bogotá, Colombia

K. S. Fathima Shemim Computing and Software Engineering Department, RAK Academic Centre, University of Bolton, Ras al Khaimah, United Arab Emirates

W. M. J. H. Fernando Faculty of Graduate Studies and Research, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka

M. P. A. W. Gamage Faculty of Computing, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka

Kanika Gandhi Thapar Institute of Engineering and Technology, Patiala, Punjab, India

Anupam Garg Computer Science and Engineering Department, Thapar Institute of Engineering and Technology, Patiala, Punjab, India

Chakshu Garg Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University, Noida, Uttar Pradesh, India

Sarita Bansal Garg Maharaja Agrasen Institute of Management Studies, Delhi, India

N. Gayathri School of Computer Science, Galgotias University, Greater Noida, India

Ankush Ghosh School of Engineering and Applied Science, The Neotia University, Kolkata, West Bengal, India

Debasis Giri Department of Information Technology, MAKAUT, Kolkata, India

S. B. Goyal City University, Petaling Jaya, Malaysia

Merajul Haque JD College of Engineering and Management, Nagpur, Maharashtra, India

T. H. Harikrishna Department of Computer Science and IT, Amrita School of Arts and Sciences, Kochi, Kerala, India

Sally M. Idhis Department of Electrical Engineering, University of Sharjah, Sharjah, UAE

T. M. Inbamalar RMK Engineering College, Chennai, India

Apoorva Jain K. J. Somaiya College of Engineering, Mumbai, India

Nipun Jain Computer Engineering, Netaji Subhas Institute of Technology, University of Delhi, New Delhi, India

Amrutha Jayakumar Department of Computer Science & IT, Amrita School of Arts & Sciences, Kochi, Amrita Vishwa Vidyapeetham, India

W. Jino Hans SSN College of Engineering, Chennai, India

C. Karthika Department of Visual Media and Communication, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Kochi, India

Sparsh Katiyar Department of Information Technology, Delhi Technological University, Delhi, India

Amrita Kaur Computer Science and Engineering Department, Thapar Institute of Engineering and Technology, Patiala, Punjab, India

Preeti Kaur Computer Engineering, Netaji Subhas Institute of Technology, University of Delhi, New Delhi, India

Faruk Kazi Veermata Jijabai Technological Institute, Mumbai, India

Ritika Khareta Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida, India

Gurpreet Singh Kharoud Thapar Institute of Engineering and Technology, Patiala, Punjab, India

Netalkar Rohan Kishor Department of Computer Science and Engg, National Institute of Technology Warangal, Warangal, Telengana, India

Rohan Kokate JD College of Engineering and Management, Nagpur, Maharashtra, India

Rahul Krishnan Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India

Nimmala Chaitanya Sai Kumar Department of Electronics and Communication Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India

Rakesh Kumar Department of Computer Engineering and Applications, GLA University, Mathura, India

Loshima Lohi Department of Computer Science & IT, Amrita School of Arts & Sciences, Kochi, India;
Amrita Vishwa Vidyapeetham, Vengal, India

N. Mahiban Lindsay Hindustan University, Chennai, India

Koushik Majumder Department of Computer Science and Engineering, Maulana Abul Kalam Azad University of Technology, Kolkata, India

Arpan Kumar Mal School of Engineering and Applied Science, The Neotia University, Kolkata, West Bengal, India

K. Mallikharjuna Rao Assistant Professor Sr. Grade 1, School of Computer Science and Engineering, VIT-AP University, Amaravati, India

Rajesh Kannan Megalingam Department of Electronics and Communication Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, Kerala, India

Rajiv K. Modanval School of Computer Science, Galgotias University, Greater Noida, India

Smit Moradiya K. J. Somaiya College of Engineering, Mumbai, India

Mohit Motiani Computer Engineering, Netaji Subhas Institute of Technology, University of Delhi, New Delhi, India

Sharad Nailwal Department of Information Technology, Delhi Technological University, Delhi, India

Prashant R. Nair Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India

Aman Namdeo Department of Civil Engineering, Amity School of Engineering and Technology, Amity University, Noida, Uttar Pradesh, India

C S Nandhakishore Department of Computer Science and Engineering, National Institute of Technology, Tiruchirappalli, India

A. G. Hari Narayanan Department of Computer Science and IT, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Kochi, India

Qassim Nasir Department of Electrical Engineering, University of Sharjah, Sharjah, UAE

Debosmit Neogi Computer Science and Engg. Dept, National Institute of Technology Agartala, Jirania, India

M. Nirmala Devi Department of Electronics and Communication Engineering, Amrita School of Engineering, Coimbatore, India

Mukta Nivelkar Veermata Jijabai Technological Institute, Matunga, Mumbai, India

Yara Omran Department of Electrical Engineering, University of Sharjah, Sharjah, UAE

Carlos Paez Pontificia Universidad Javeriana, Bogotá, Colombia

Maya L. Pai Department of Computer Science & IT, Amrita School of Arts & Sciences, Kochi, Amrita Vishwa Vidyapeetham, India

Prajyot Palimkar School of Engineering and Applied Science, The Neotia University, Kolkata, West Bengal, India

Sweta Panigrahi Department of Computer Science and Engg, National Institute of Technology Warangal, Warangal, Telengana, India

Anshu Parashar Computer Science and Engineering Department, Thapar Institute of Engineering and Technology, Patiala, Punjab, India

Abhishek Patel K J Somaiya College of Engineering, Mumbai, India

Saurabh Patel Department of Information Technology, Delhi Technological University, Delhi, India

Debanjan Pathak Department of Computer Science and Engg, National Institute of Technology Warangal, Warangal, Telengana, India

Sachin Patil JD College of Engineering and Management, Nagpur, Maharashtra, India

Darla Vineeth Prithvi Department of Electronics and Communication Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India

Hari Sudarshan Rahul Puram Department of Electronics and Communication Engineering, Amrita Vishwa Vidyapeetham, Kerala, India

Anand Singh Rajawat Department of Computer Science Engineering, Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore, India

S. S. Rajendra Prasath Department of Computer Science and Engineering, Amrita School of Engineering, Coimbatore, India

U. S. N. Raju Department of Computer Science and Engg, National Institute of Technology Warangal, Warangal, Telengana, India

Aravapalli Rama Satish Assistant Professor Sr. Grade 1, School of Computer Science and Engineering, VIT-AP University, Amaravati, India

Ankita Rana Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Chennareddy Pavanth Kumar Reddy Department of Electrical Engineering, University of Dayton, Dayton, OH, USA

A. T. Rohit Surya Department of Computer Science and Engineering, Amrita School of Engineering, Coimbatore, India

Nitin Saluja Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Venkata Chanakya Samsani Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida, India

Naveen Samudrala Department of Electrical Engineering, University of Dayton, Dayton, OH, USA

Vaishnavi Sankar Department of Electronics and Communication Engineering, Amrita School of Engineering, Coimbatore, India

A. Santhanavijayan Department of Computer Science and Engineering, National Institute of Technology, Tiruchirappalli, India

Priyanka Shah Parul University, Vadodara, Gujarat, India

Sandeep K. Sharma School of Computer Science, Galgotias University, Greater Noida, India

Seemu Sharma Thapar Institute of Engineering and Technology, Patiala, Punjab, India

Shanu Sharma Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida, India

Rabindra Nath Shaw Department of Electrical, Electronics and Communication Engineering, Galgotias University, Greater Noida, India

Irfan Siddavatam K. J. Somaiya College of Engineering, Mumbai, India

Abhishek Singhal Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University, Noida, Uttar Pradesh, India

Arshdeep Singh Netaji Subhas University of Technology, Dwarka, Delhi, India

Harjit Singh APS Neighbourhood Campus, Punjabi University, Patiala, India

Priyanka Singh Department of Civil Engineering, Amity School of Engineering and Technology, Amity University, Noida, Uttar Pradesh, India

Trisha Sinha School of Engineering and Applied Sciences, The Neotia University, Kolkata, West Bengal, India

Shrikant V. Sonekar JD College of Engineering and Management, Nagpur, Maharashtra, India

K. Sreekumar Department of Computer Science and IT, Amrita School of Arts and Science, Kochi, Amrita Vishwa Vidyapeetham, India

V. V. Subrahmanyam School of Computer and Information Sciences, IGNOU, New Delhi, India

Sudip Suklabaidya Department of Computer Science and Application, Karimganj College, Karimganj, India

Gayathri Suresh Department of Computer Science and IT, Amrita School of Arts and Sciences, Kochi, Kerala, India

Manar Abu Talib Department of Computer Science, University of Sharjah, Sharjah, UAE

Hemanth Sai Surya Kumar Tammana Department of Electronics and Communication Engineering, Amrita Vishwa Vidyapeetham, Kerala, India

Ashu Taneja Chitkara University School of Engineering and Technology, Chitkara University, Himachal Pradesh, India

Santosh Tantravahi Department of Electronics and Communication Engineering, Amrita Vishwa Vidyapeetham, Kerala, India

Prakhar Tayal School of Computer Science, Galgotias University, Greater Noida, India

Niraj Tevar Parul University, Vadodara, Gujarat, India

Viraj Thakkar K J Somaiya College of Engineering, Mumbai, India

Jobanpreet Singh Thind Thapar Institute of Engineering and Technology, Patiala, Punjab, India

Nagasai Thokala Department of Electronics and Communication Engineering, Amrita Vishwa Vidyapeetham, Kerala, India

Manoj Titre JD College of Engineering and Management, Nagpur, Maharashtra, India

Tushar Vashisth Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida, India

Aditya Vedpathak K J Somaiya College of Engineering, Mumbai, India

P. P. Vijayalakshmi Department of English and Languages, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Kochi, India

P. Vineetha Sankar Department of Computer Science and IT, Amrita School of Arts and Science, Kochi, Amrita Vishwa Vidyapeetham, India

Dr. Ulf Witkowski Electronics and Circuit Technology Research Group, South Westphalia University of Applied Sciences, Iserlohn, Germany

Jyoti Yadav Netaji Subhas University of Technology, Dwarka, Delhi, India

V. Yasaswini Computer Science and Engineering Department, Pondicherry Engineering College, Puducherry, India

P. Yaswanthram Department of Computer Science and Engineering, Amrita School of Engineering, Coimbatore, India

Recent Trends in Electromyography Signal Processing of Neuromuscular Diseases: An Outlook



M. Emimal, W. Jino Hans, T. M. Inbamalar, and N. Mahiban Lindsay

Abstract This paper presents a bibliometric review of several techniques applied to the EMG signals. We reviewed research papers, which were specifically applied for the EMG signals. The EMG signal contains a huge amount of data, thus the EMG signal research grabs the significance of advanced techniques and analysis of data, which are capable of handling ‘Big Data’. Several noise reduction techniques were discussed and it was found that the wavelet-based noise reduction is a promising technique for EMG classification. More prominent feature extraction and classification techniques and their performance were also reviewed. The modern EMG signal analysis mainly emphasizes feature learning, which is specifically ‘deep learning’, which combines feature extraction and classification, also to improve classification accuracy. A performance analysis of Convolutional Neural Network (CNN) was done in the later sections.

Keywords Electromyography (EMG) · Convolutional neural network (CNN) · Big data · Deep learning

1 Introduction

As reported by the World Health Organization (WHO), more than 15 million people in the world suffer from stroke. Among these, five million people do not survive and five million people stay paralyzed life long. Stroke is a result of blood flow interruption to the brain, which affects the brain hemisphere. Hence, the side which is

M. Emimal (✉)
Anna University, Chennai, India

W. Jino Hans
SSN College of Engineering, Chennai, India

T. M. Inbamalar
RMK Engineering College, Chennai, India

N. Mahiban Lindsay
Hindustan University, Chennai, India

supervised by the affected hemisphere results in motor and sensorial dysfunction [1]. There are two main categories of stroke: (i) Ischemic (a blood clot occurs in the blood vessel) and (ii) hemorrhagic (occurs when a blood vessel is weak and bursts, thus, the blood goes into the brain). Brain plasticity plays a vital role in rehabilitation and thus motor recovery. Neural pathway restoration, cortical reorganization, and motor recovery form the basis of long-term plasticity. All the neurological and physical rehabilitation techniques are based on neural plasticity. The rehabilitation helps a patient to regain the normal functionality. A rehabilitation process is burdensome, difficult, and painful; it should be performed in a proper way such that the chances of recovery and regaining are maximized. The rehabilitation process should be initiated soon after a stroke in order to increase motor recovery. The inner workings of the human body differ between stroke patients and that of healthy people. The muscle activation can be measured by measuring the electrical activity associated with it. The study of muscle function by analyzing the electrical signals emerging as a result of muscular contractions is called Electromyography (EMG). The EMG is acquired from the post-stroke patients, and is used for evaluating the motor functionality, and thus is used for accessing the degree of impairment. The EMG signal aids the control devices, thus improving the rehabilitation [2]. The subset of pattern analysis includes pattern recognition, pattern inference, pattern interpretation, pattern reasoning, and pattern learning. It includes the series of processing steps from the input, for identifying patterns for classification, detection, and parameter estimation. A feature in pattern recognition is the structural characteristics, description, transform, or graph, obtained from the pattern. The conversion of a pattern to a feature is termed as feature extraction, and feature selection represents the way of reducing the number of features. Grouping of patterns depends upon the similarity of patterns, or which belongs to the same class is the pattern classification [3]. The surface EMG signals (sEMG) have been used for a variety of applications including rehabilitation, prosthetic hands, diagnosis of neuromuscular disorders, etc. It is known that the sEMG signals used in the robotic rehabilitation devices are used for quantifying the hand function assessment.

This paper provides a detailed elucidation of the necessary steps in EMG signal processing. The steps in EMG signal processing are depicted in Fig. 1. The EMG signal, when passing through various tissues, is affected by different kinds of noise. The amplitude of the EMG signal ranges from 0 to 10 mV. Inherent noise, ambient noise, motion artifact, crosstalk, electrode contact, transducer noise, baseline shifts, etc. are some of the noise present in the EMG signal. By exploiting adaptive noise cancelation techniques, the noise effect is reduced adaptively. The Least Mean Square

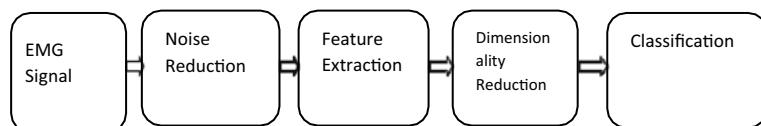


Fig. 1 EMG signal processing steps

(LMS) algorithm is the best cancelation technique [4]. The muscle activation timing of an sEMG signal is detected by obtaining the Motor Unit Action Potential (MUAP). Threshold-based methods can detect the MUAP, or it can be done manually. The Motor Units (MU) timing decides the timing of the surface EMG signal. Literally, the noise which is added into the EMG signal causes an error, or it may lead to the reduction in the signal-to-noise ratio of the EMG signal. The signal from the neighboring muscle may interfere, and this leads to crosstalk and is unavoidable. Inaccurate measurement also leads to inaccurate muscle activation timing. To determine muscle activation, an Artificial Neural Network is used. A well-trained RNN can achieve the state-of-the-art accuracy levels.

2 Noise Reduction in EMG Signals

Yang et al. [5] showed that the Wigner-Ville distribution expresses the motor unit's frequency range. This distribution extremely focuses on the time and instantaneous frequency. This method is not suited for the EMG signal, due to the cross term effect and noise. The wavelet transform is divided into continuous wavelets and discrete wavelets. The processing time for the discrete wavelet transform is less. Moreover, the Continuous Wavelet Transform is more consistent and takes less computational time, and down-sampling is not present. For the successful analysis of EMG signals (non-stationary), DWT is used but the resulting feature vector shows high dimensionality. The expression for CWT is given in Eq. (1).

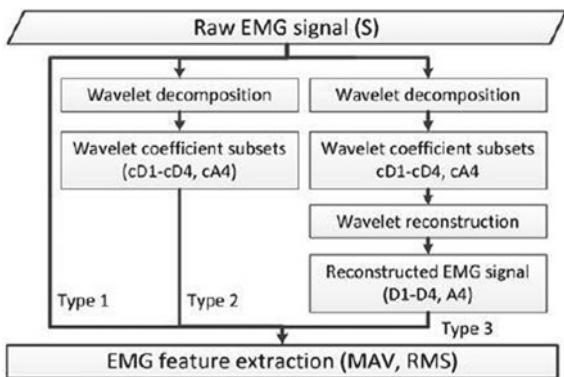
$$\Psi(a+b) = \frac{1}{\sqrt{a}} \Psi\left(\frac{t-b}{a}\right) \quad (1)$$

where $\psi(t)$ is termed as the mother wavelet. The DWT in the discrete-time domain is given in Eq. (2).

$$x(t) = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} d(k, l) 2^{\frac{k}{2}} \psi(2^{-kt} - l) \quad (2)$$

Guglielminotti and Merletti [6]. When the MUAP's shape matches with the wavelet shape, the wavelet transform exhibits good energy localization in the time domain [7]. For multi-component signals, the wavelet transform does not exhibit cross terms [8]. The wavelet transform detects the MUAPs, even when the white noise is present. Morlet and Mexican hat wavelets are the well-known continuous wavelet techniques, but the shape of MUAP is not accurately matched by the Mexican hat wavelet. The Short-term Fourier Transform (SFT) and the Fast Fourier Transform (FFT) deal only with stationary signals. The preprocessing stage of an EMG signal gives better results when wavelet denoising techniques are used. Denoising

Fig. 2 Feature extraction by wavelet coefficients



the random noises from the EMG signal by employing filtering techniques is a difficult task, and so the wavelet-based denoising techniques has been proposed [9]. For an efficient use of the wavelet-based denoising technique, five parameters are to be considered: (i) Wavelet type, (ii) scaling, (iii) selecting the threshold, (iv) threshold rescaling method, and (iv) threshold function. While concerning wavelet denoising, the crucial part is the proper selection of the wavelet. Phinyomark et al. [10] analyzed the functions db2, db5, sym5, sym8, and coif5 as wavelet functions to be used as the denoising techniques. The fifth-order coiflet provides a better reconstruction of the sEMG signal [11]. The wavelet denoising method preserves the information contained in the EMG signal [12]. The threshold selection does not influence the denoising performance. For long-length filters, the Daubechies wavelet gives higher energy. From the comparison of all the wavelet denoising techniques, the Daubechies wavelet gives the better results and the use of db functions, db2, dg4, dg6, db45, and db46 at the decomposition level 4 is recommended. Figure 2 shows how wavelet transform is used for denoising the EMG signal (Fig. 2).

An efficient decomposition technique was proposed, using a continuous wavelet transform which adaptively updates the threshold automatically based on the statistical property of the EMG data, by iteratively estimating the Signal-to-Noise Ratio (SNR) [13]. The stopping criterion is reduced by the stopping criterion for the decomposition levels of the continuous wavelet. Excellent performance was achieved with the bias less than 18 ms, accuracy higher than 97% and timing error lesser than 5%. This method is iterative, operator-independent, and adaptive, works well even for noisy and weak EMG signals. Dawid Gradolewski et al. proposed a noise filtering technique based on Daubechies wavelet function (db4), up to a decomposition level of 10 by using a heuristic algorithm for threshold selection, with a scaling function based on white noise [14]. The wavelet packet decomposition is a wavelet transform technique which is also called optimal sub-band tree structuring, in which the sampled signal is progressed through filters. The filtering performance of the Wavelet Packet Decomposition technique is efficient, without compromising the signal quality (Fig. 3).

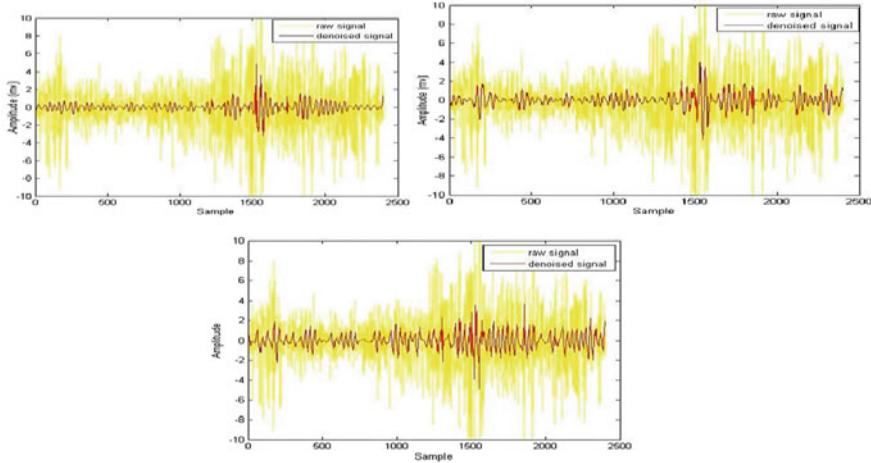


Fig. 3 Denoising the EMG signal by different Daubechies wavelet functions (db2, db4, and db6)

Another denoising technique which is robust is the Universal threshold algorithm [15]. A modified algorithm provided the improved SNR performance. Donoho's wavelet-based denoising algorithm is limited to multifunction control. The pitfall lies in the signal quality; the noise along with the information is lost. In order to alleviate this issue, a thresholding scheme was proposed, namely Global Scale Modified Universal (GSMU). This thresholding is more efficient than that of Donoho's [16]. Most of the denoising techniques were focused on a particular kind of noise. Hence, the noise will not be completely removed from the signal, and the efficacy of the system degrades. A Variational Mode Decomposition (VMD) technique is used to eliminate three types of noise (PLI, BW, and WGN) in the band-limited sub-bands. This method achieved the best performance and the SNR was improved up to 20 dB.

In order to enhance the signal quality, Stachaczky et al. [17] proposed a pre-processing step, which weights the sEMG signal adaptively, to enhance the quality of a signal for a specific time interval. This technique attenuates noise from multi-channel electrodes, the so-called spatial filtering, which yields a high spatial resolution [18]. A more promising technique for minimal distortion was developed by Kun Wang et al., which is applicable to High-Density sEMG, particularly to eliminate Power Line Interference and White Guassian noise. The Canonical Correlation Analysis and the Independent Component Analysis were used as a tool for removing high density noise (Table 1).

3 Higher Order Statistics (HOS)

For the Nerve Conduction Studies (NCS) and the neuromuscular diagnosis, the Higher Order Statistics (HOS) plays a vital role, since much of the information

Table 1 Different noise removal techniques and their results

Noise removal technique used	Results
Morlet and Mexican hat wavelets	MUAP shape does not match
Coiflet wavelet method	Better accuracy and reconstruction
Daubechies wavelet method	The information contained in the signal is preserved
Continuous Wavelet Transform (CWT)	Excellent performance accuracy of 97% was achieved
Wavelet packet decomposition	Efficient signal denoising without compromising the signal quality
Donoho's wavelet	Excellent noise removal, but information is lost
Global scale modified universal thresholding	To overcome the limitation of Donoho's wavelet, this technique was proposed; it preserves the information
Variational mode decomposition	Improved SNR of up to 20 dB
Multi-channel EMG noise reduction	A very high spatial resolution was employed
Canonical correlation analysis and the independent component analysis	This technique was employed to remove the PLI and WGN of HD-sEMG signals

is available in it. The moments and cumulants of third order and more are termed as poly spectra or Higher Order Spectra (HOS). The deviation of statistical properties such as Gaussianity and stationarity can be identified by HOS due to its unique properties. The third-order spectrum is named as bispectrum, the fourth-order as the trispectrum, and the power spectra also belong to the higher order spectral class: it is the second-order spectrum. The joint moments of order r of the random variables x_1, \dots, x_n are given by

$$\begin{aligned} \text{Mom}[x_1^{k_1}, \dots, x_n^{k_n}] &= E\{x_1^{k_1}, \dots, x_n^{k_n}\} \\ &= (-j)^r \partial^r \varphi(\omega_1, \dots, \omega_n) / \partial \omega_1^{k_1} \cdots \partial \omega_n^{k_n} |_{\omega_1 = \cdots = \omega_n = 0} \end{aligned} \quad (3)$$

where $k_1 + \dots + k_n = r$ and $\varphi()$ is their joint characteristic function.

4 EMG Feature Extraction and Classification

Feature extraction transforms a raw EMG signal into an appropriate data structure by noise removal and spotlighting the important data. Plenty of methods are available for extracting features from the EMG signal. Angkoon Phinyomark et al. proposed a feature extraction technique, which calculates the mean and median frequencies. A separate denoising technique is not needed for this algorithm. The EMG histogram, combination of MMNF and Williamson amplitude, were the feature vectors used

for the classification purpose. This method gives better recognition results even for the noisy environments [19]. The features can also be extracted from multi-level wavelet decomposition. The useful resolution components were obtained by selecting different mother wavelets. The noise is also removed in this method. Optimal decomposition of wavelet is obtained from the Daubechies wavelet of the order seven and four [20]. Feature selection plays a crucial role in the success of the EMG analysis. The feature extraction using multiple parameters is a promising effort. Wavelet transform is the best feature extraction method of EMG signals. Autoregressive method, discrete wavelet transform, and wavelet packed energy are used as the techniques for extracting the features from EMG signals. The wavelet packet transform can be done with a two-channel filter bank [21]. Thirty seven time- and frequency-domain features such as Integrated EMG, MAV, modified MAV, simple square integral, zero-crossing, and amplitude of the first burst were calculated and the scatter plot was verified. The results were grouped into four categories. The result shows that the time-domain features reveal better performance compared with the frequency-domain features [22]. The Discrete Wavelet Transform (DWT) suits well for real-time applications.

The lower and higher frequency components of the signal are split into Ca and Cd values, respectively. 1D wavelet decomposes the signal into 3 levels. The wavelet function, db3, from Daubechies wavelet function is used for analyzing the signal. For feature extraction, the Mean Absolute Value (MAV) and RMS value are used. This method provides an improved separability, and it is implemented in real time using machine learning techniques. The MAV feature is given by

$$MAV = \frac{1}{N} \sum_{n=1}^N \sqrt[3]{x_n} \quad (4)$$

where x_n is the nth sample of the EMG signal.

Doulah et al. proposed a technique which extracts and classifies the features. Two techniques were given: (i) High-energy Discrete Wavelet coefficients are extracted on a frame-by-frame basis with the maximum possible values. (ii) The template matching decomposition technique is used to extract the MUAP from the EMG data. From the extracted MUAPs, only the dominant MUAP is selected based on the higher energy, and only the dominant MUAPs are given as input to the classifier. The KNN algorithm is used for classifying the EMG signals. The dimension of the feature is very low, which adds an advantage of computational time saving. Overall, the sensitivity, specificity, and the classification accuracy are more satisfactory [23]. Linear Discriminant Analysis (LDA) is the most efficient technique for feature extraction. The LDA classifier provided a classification accuracy of 94%. In the case of a noisy environment, the performance of the KNN classifier is degraded. The highest classification rate for a KNN classifier was obtained with the value $k = 5$. For the SVM classifier, the difficulty lies in the parameter selection for the kernel function. Three of the kernel functions, namely Radial Basis Function (RBF), Polynomial (PLN), and Linear (LIN) were analyzed and it was found that the polynomial kernel outperformed than the others. The classification accuracy achieved

Table 2 Performance of different classification techniques

Model	Training time (s)	Testing time (s)	Accuracy %
ANN	32.250	1.18	98.20
SVM	1.80	0.20	96.15
LR	0.10	0.05	97.50
LDA	0.09	0.04	97.25
ELM	0.07	0.005	99.75

using SVM was 93%. Another important classifier technique is the General Neural Network (GRNN), in which it does not converge to poor solutions (local minima) and is bounded by maximum and minimum of the recordings. In GRNN, the smoothing parameter decides the performance of the classification. The smoothing parameter lies in the range of 0.05–0.5. The GRNN achieved a classification accuracy of 95%, which is the highest among all classifiers (Table 2).

Principal Component Analysis (PCA) computes the principal components, perhaps using only the first few principal components and leaving the rest. It can also be used in exploratory data analysis and for predictive models. This is used in dimensionality reduction. Dimensionality reduction extrapolates each data point into the first few principal components thus obtaining low-dimensional data, preserving the data's variation as much as possible [24]. PCA and Independent Component Analysis (ICA) can also be used for extracting features from the EMG signal. Experimental findings reveals that DWT provides a superior feature, when compared with PCA and ICA. DWT combined with the SVM classifier provides a better accuracy of 94.28%.

Negi et al. [25]. For rehabilitation devices and prosthetics, Linear Discriminant Classifier (LDC) and the uncorrelated Linear Discriminant Classifier (ULDC) give better results. The waveform length and Williamson amplitude were considered to be the best features for upper-limb motions. Pornchai et al. compared the feature extraction techniques, namely PCA, LDA, ULDA, OFNDA, SRLDA, and SRELM. Also, the classifiers, SVM, LB, NB, KNN, RBF-ELM, AW-ELM, and NN were compared. A classification accuracy of 99% was achieved for the SRELM technique and NN for classification, which is the choicest among all combinations. A novel Ternary Pattern and DWT (TP-DWT) feature extraction technique was put forward by Turker et al. This method includes channel concatenations, feature extraction, feature selection by level 2, and feature classification by using KNN classifier. The classification accuracy obtained was 99.14%.

5 Deep Learning for EMG Signal Analysis

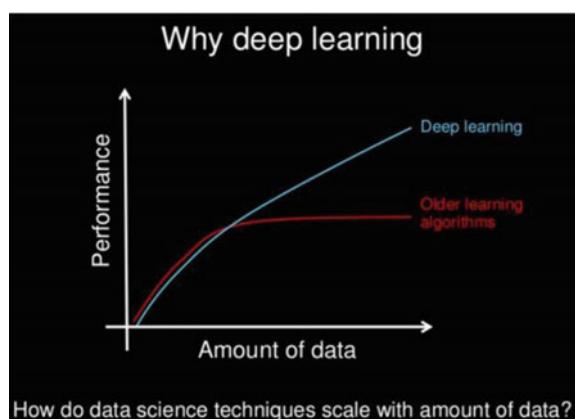
The subset of artificial intelligence (AI) is the deep learning, which emulates the human brain in generating patterns and data processing to be used in decision-making. It is capable of unsupervised learning of unstructured data. Unlike traditional programs, it also utilizes the hierarchy of ANN, which enables the machines to utilize a non-linear approach. The ANNs are built like the human brain, with neurons together forming a network (Fig. 4).

Different techniques for EMG signal processing have been developed and tested. Most of these techniques are prone to handle only a small amount of data or datasets. Traditional EMG signal processing methods are difficult to handle large-scale datasets efficiently. Moreover, classical feature extraction and classification techniques are not feasible for large datasets and so, a proper redesign is needed and the ways of traditional computation should be changed. One feasible approach of analyzing big data is to modify the traditional methods and allow it to work under a parallel computation scenario. For feature selection, renowned population-based approaches, namely Particle Swarm Optimizations (PSO), Genetic Algorithm (GA), and Ant Colony Optimizations (ACO) are the effective techniques. These techniques can also work on the Graphics Processing Unit (GPU).

Convolutional Neural Network (CNN)

The CNN is the popularly used deep learning model. Three hidden layers are there in CNN: convolutional layer, sub-sampling layer or pooling layer, and the fully connected layer. The convolutional layer performs the feature extraction, and dimensionality reduction is done in the pooling layer. The CNN is an efficient technique in pattern recognition and is used as a substitute for SVM, KNN, MLP, LDA, and RF. The classification accuracy obtained by recognizing hand movements using CNN was up to 99.5%. The CNN is also robust to muscle fatigue, inter-subject variability, displacement of electrodes, and long-term use.

Fig. 4 Performance of deep learning versus Older learning algorithms



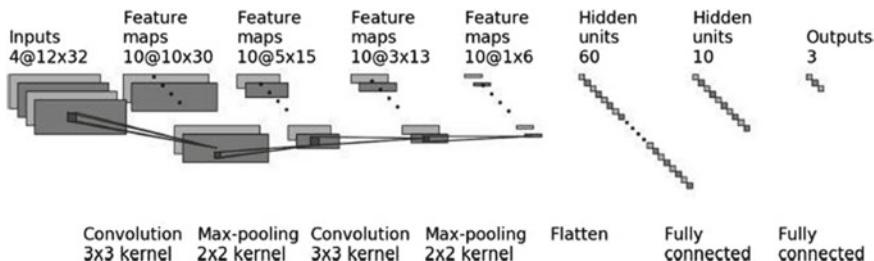


Fig. 5 A convolutional neural network structure

Côté-Allard et al. introduced Transfer learning to enhance the performance and to lower the computational time, using continuous wavelet transform as pre-extraction. Spectrograms are the most commonly used feature extraction methods for deep learning models. Zhai et al. used an input, namely PCA-reduced spectrogram in which the classifier updates automatically. For a simple architecture of CNN, with a short time window for raw EMG signal, the accuracy of the classifier was reduced. Thus, the performance of the algorithm depends on the optimization parameter, network architecture, and model.

Shioji et al. [26] proposed a CNN network of two-class separation, which resulted with an accuracy of 94.9% for accessing wrist kinematics. This method preprocess the EMG data by employing a high-pass filter at the input to CNN. A combined CNN and RNN were proposed by Peng Xia et al. which was trained by gradient descent and back propagation. While comparing the architecture with the Support Vector Regression for the same datasets, the classification accuracy was improved with respect to time [27]. An average of 10.18% increase in the classification accuracy was obtained by employing a self-recalibrating classifier for 50 hand movements, which is automatically updated and does not need user retraining. The results were compared with SVM; this system provided more absolute performance and an efficient training. A multi-input multi-output CNN model which does not contain a pooling layer was proposed by Ryohei Shioji et al. with an average accuracy for recognizing hand movements of 94.6%. Also an accuracy of personal authentication of 95.0% was obtained.

A CNN model with a different scale signal was proposed, which takes frequency spectra as the input and the window length is varied. The classification accuracy is affected by the window length [28]. Compared with the time-domain input model, the frequency-domain input gives a better classification accuracy. Since EMG is taken from the muscles, it has strong temporal dependencies, but the CNN exploits the spatial correlations only. A hybrid model called CNN-LSTM was proposed to utilize the temporal and spatial correlations. This hybrid model is used for analyzing the wrist movements. This hybrid CNN-LSTM model outperforms CNN, RF, and SVR methods. Laezza et al. had a performance analysis of CNN, RNN, CNN + RNN networks, and the results were accordingly 89.01, 91.81, and 90.4%. It is known that the RNN alone gives the best classification performance.

More deep learning models are in practice; some of them are Restricted Boltzmann Machines (RBMs), Deep Belief Networks (DBN), Deep Boltzmann Machines (DBM), Convolutional Boltzmann Machines (CBM), Convolutional Deep Belief Networks (CDBN), Generative Adversarial Networks (GAN), Convolutional Generative Networks (CGN), etc. Comparing the performance of all the networks, the CNN gives better classification results, so CNN is widely used in practice. Further, to enhance the robustness, classification accuracy, and quality of the signal, CNN with a combination of other networks can also be used. The CNN along with LSTM yields a higher accuracy. The LSTM deals with the time-series data; it can capture non-linear and long-term data. Our future research aims at the design of hybrid CNN architectures to increase the classification accuracy and robustness.

6 Conclusion

This paper reviewed several signal processing techniques required for the processing of EMG data. The noise removal is the primary aim, which can be done in many ways; the most commonly used is the wavelet-based denoising technique. Hopefully, this study will be a roadmap for the detection, processing, and classification of the EMG signals, and for diagnosing and rehabilitation of neuromuscular diseases. The plenty of data availability has enabled the researches to utilize the neural networks and also the deep learning techniques. The current research focuses on the hybrid deep neural networks, which improves the overall performance and efficacy of the system.

References

1. Perretta, R.: A preliminary study in EMG based Upper-Limb Stroke Rehabilitation. Masters Thesis
2. Park, S.-H., Lee, S.-P.: EMG pattern recognition based on artificial intelligence techniques. *IEEE Trans. Rehabil. Eng.* **6**(4) (1998)
3. Khokhar, Z.O., Xiao, Z.G., Menon, C.: Surface EMG pattern recognition for real-time control of a wrist exoskeleton. *BioMed. Eng. OnLine* **9**(41) (2010). <http://www.biomedical-engineeringonline.com/content/9/1/41>
4. Khawaled, I.A., Abotabl, A.: Neural muscle activation detection: a deep learning approach using surface electromyography. *J. Biomech.* www.elsevier.com/locate/jbiomech, www.JBiomech.com
5. Yang, W., Yang, D., Liu, Y., Liu, H.: EMG pattern recognition using convolutional neural network with different scale signal/spectra input. *Int. J. Hum. Robot.* **16**(4), 1950013 (2019). <https://doi.org/10.1142/s0219843619500130>. (World Scientific Publishing Company)
6. Guglielminotti, P., Merletti, R.: Effect of electrode location on surface myoelectric signal variables: a simulation study. In *Studies in Health technology and Informatics: Electrophysiological Kinesiology*, vol. 5. IOS Press, Amsterdam, The Netherlands
7. Laterza, F., Olmo, G.: Analysis of EMG signals by means of the matched wavelet transform. *Electron. Lett.* **33**, 357–359 (1997)

8. Ismail, A.R., Asfour, S.S.: Continuous wavelet transform application to EMG signals during human gait. In: Proceedings of the 32nd Asilomar Conference on Signals, Systems & Computers, Pacific Grove, CA, USA, 1–4 November 1998, vol. 1, pp. 325–329
9. Phinyomark, A., Limsakul, C., Phukpattaranont, P.: A comparative study of wavelet denoising for multifunction myoelectric control. In: Proceedings of the International Conference on Computer and Automation Engineering (ICCAE'09), Bangkok, Thailand, 8–10 March 2009, pp. 21–25
10. Phinyomark, A., Limsakul, C., Phukpattaranont, P.: Optimal wavelet functions in wavelet denoising for multifunction myoelectric control. In: Proceedings of the Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, Pattaya, Chonburi, 6–9 May 2009, pp. 1098–1101
11. Zhang, X., Wang, Y., Han, R.P.S.: Wavelet transform theory and its application in EMG signal processing. In: Proceedings of the Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Yantai, China, 10–12 August 2010, vol. 5, pp. 2234–2238
12. Jiang, C.F.; Kuo, S.L.: A comparative study of wavelet denoising of surface electromyographic signals. In: Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS 2007), Lyon, France, 23–26 August 2007, pp. 1868–1871
13. Varrecchia, T., D'Anna, C., Schmid, M., Conforto, S.: Generalization of a wavelet-based algorithm to adaptively detect activation intervals in weak and noisy myoelectric signals. © 2020 Elsevier Ltd, Biomedical Signal Processing and Control, <https://doi.org/10.1016/j.bspc.2019.101838>1746-8094. © 2020 Elsevier Ltd. All rights reserved
14. Bhoi, A.K., Tamang, J., Mishra, P.: Wavelet packet based Denoising of EMG signal. Int. J. Eng. Res. Dev. **4**(2), 78–83 (2012). e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
15. Kalitkar, K.R., Anumula, J.: New wavelet transform Denoising algorithm. Int. J. New Innov. Eng. Technol. **4**(3) (2016). ISSN: 2319-6319
16. Ma, S., Lv, B., Lin, C., Sheng, X., Zhu, X.: EMG signal filtering based on variational mode decomposition and sub-band thresholding. J. Latex Class Files **14**(8) (2015)
17. Stachaczyk, M., Atashzar, S.F., Farina, D.: Adaptive spatial filtering of high-density EMG for reducing the influence of noise and artefacts in myoelectric control, pp. 1534–4320. (c) 2020 IEEE
18. Wang, K., Chen, X., Wu, L., Zhang, X., Chen, X., Wang, Z.J.: High-density surface EMG denoising using independent vector analysis. IEEE Trans. Neural Syst. Rehabil. Eng. <https://doi.org/10.1109/tnse.2020.2987709>
19. Phinyomark1, A., Limsakul, C., Phukpattaranont, P.: Application of wavelet analysis in EMG feature extraction for pattern classification. Meas. Sci. Rev. **11**(2) (2011)
20. Subasi, A.: Classification of EMG signals using combined features and soft computing techniques. Appl. Soft Comput. J. www.elsevier.com/locate/asoc
21. Phinyomark, A., Phukpattaranont, P., Limsakul, C.: Feature reduction and selection for EMG signal classification. Expert Syst. Appl. J. www.elsevier.com/locate/eswa
22. Sharma, Sachin: Wavelet analysis based feature extraction for pattern classification from single channel acquired EMG signal. Elixir Control Engg. **50**, 10320–10324 (2012)
23. Al Omari, F., Hui, J., Mei, C., Liu, G.: Pattern recognition of eight hand motions using feature extraction of forearm EMG signal. Proc. Natl. Acad. Sci., India, Sect. A Phys. Sci. <https://doi.org/10.1007/s40010-014-0148-2>
24. Kehri, V., Ingle, R., Awale, R., Oimbe, S.: Techniques of EMG signal analysis and classification of Neuromuscular diseases. ICCASP/ICMMD-2016. Advances in Intelligent Systems Research, vol. 137, pp. 485–491. © 2017-The authors. Published by Atlantis Press
25. Negi, S., Kumar, Y., Mishra, V.M.: Feature Extraction and classification for EMG signals using linear discriminant analysis. 978-1-5090-3480-2/16. ©2016 IEEE
26. Shioji, R., Ito, S.-I., Ito, M., Fukumi, M.: Personal authentication based on wrist EMG analysis by a convolutional neural network. In: Proceedings of the 5th IIAE International Conference on Intelligent Systems and Image Processing (2017)

27. Zhai, X., Jelfs, B., Chan, R.H.M., Tin, C.: Self-recalibrating surface EMG pattern recognition for neuroprostheses control based on convolutional neural network. *Front. Neurosci.* **11**, Article 379 (2017). www.frontiersin.org
28. Yang, W., Yang, D., Liu, Y., Liu, H.: EMG pattern recognition using convolutional neural network with different scale signal/spectra input. *Int. J. Hum. Robot.* **16**(4), 1950013 (2019). <https://doi.org/10.1142/S0219843619500130>. (World Scientific Publishing Company)
29. Costa, Á., Itkonen, M., Yamasaki, H., Alnajjar, F.S., Shimoda, S.: Importance of Muscle Selection for EMG Signal Analysis during Upper Limb Rehabilitation of Stroke Patients, 978-1-5090-2809-2/17 ©2017. IEEE

Conceptualization, Visualization, and Modeling of Ontologies for Elementary Kinematics



C S Nandakishore, Gerard Deepak, and A. Santhanavijayan

Abstract Ontology, in general, is the study of understanding the way of being. In terms of computer science and engineering, it is the way of explaining a concept in both ways, i.e. human-understandable and machine-understandable formats. This mainly explains the relationship between two entities which are considered. Ontology is conceived where the process has conceptualization and visualization. The modeling of Ontology is done with Protégé—an interface which codes the given Ontology into a .owl file. Flow Mapping has been accomplished for the visualization of the problem taken. In this paper, the relationships between displacement, distance, velocity, speed, acceleration, and retardation are intricated into a dynamic Ontology model, with the condition that all the objects considered obey Newton's laws. Kinematics mainly deals with the motion of objects in a plane or space, which is an elementary concept in physics. The relationship between the objects is established using various attributes such as "Speed is inversely proportional to the time passed;" here the "inversely proportional" describes the relationship between the entities speed and time. An Ontology model with a reuse ratio of 0.0706 is proposed in the paper.

Keywords Ontology · Basic kinematics · Newton's law · Foundations in physics · Elementary kinematics

1 Introduction

Learning a concept with a solid understanding of the relationships among the terms involved gives a clear understanding to the learner. Ontology provides a suitable and efficient method to visualize and maintain any kind of knowledge by arranging the given data in a map indicating their relationships. In general, Ontology may have many links which establish the relationship between parameters. This paper mainly provides a simple and efficient Ontology for basic kinematic concepts taught in high school. Basic Kinematics provides the basement for many concepts in Engineering

C S Nandakishore · G. Deepak (✉) · A. Santhanavijayan
Department of Computer Science and Engineering, National Institute of Technology,
Tiruchirappalli, India

Mechanics, and it is essential for a student to understand any mechanical process which occurs in his vicinity. Learning the foundational concept provides a basic understanding of an object's motion and its rest to learn more advanced concepts which are inculcated in many day-to-day machines. From an Engineering perspective, kinematics is a basement to learn all mechanics related to all machinery.

Interrupting the knowledge in a domain is necessary to learn all the advanced concepts, and a full domain Ontology is proposed in this paper, providing a greater perspective for a student to deep dive into the subject. The main topics are explained with subtopics handling the special cases, which in turn helps in problem-solving. This paper mainly aims to provide ease of teaching to the instructors at the same time it provides the knowledge without any confusion in the form of Ontology which can be interpreted by information and learning systems in the future. Information systems need not be equipped to learn specialized concepts like in kinematics, but rather interpret on the basis of the modeled Ontology.

Motivation: Starting from elementary education, understanding the concepts behind every concept of science boosts the understanding of the learner and improves the overall efficiency of acquiring knowledge from an abstract observation of an event. As an assisting script to the instructor and learner, Ontology models provide great support to nudge students toward learning concepts with a visualization. Moreover, in the era of Web 3.0, information and learning systems must be equipped to infer from linked semantic data rather than learning from scratch. Ontologies that depict specialized concepts play a vital role in achieving this milestone.

Contribution: The methodology used in this model is a hybrid of some existing ontologies used in teaching. The concepts are studied from scratch and arranged in hierarchical order and then conceptualized into a dynamic Ontology. This also includes the conversion of human-readable unclassified data into machine testable data nested in classes. In the implementation phase, the relationships between objects are defined with the help of attributes and properties in the form of annotations in RDF. The entire model is designed using Protégé. As a result, the .owl file is created, and the file is visualized with a graphical interface using WebVOWL. The created Ontology is evaluated using semiotics Ontology which is one of the sustainable and efficient ways to evaluate Ontology models.

Organization: The paper has further divisions after Introduction in which Sect. 2 gives a glimpse of the related works which provided the basement for the paper, Sect. 3 covers Preliminaries—explaining all the terminologies used in the paper, Sect. 4 covers design of Conceptualization, Sect. 5 covers Implementation, Sect. 6 covers the evaluation of Ontology, and Conclusion is depicted in Sect. 7.

2 Related Works

Ganapathi et al. [1] investigated the adoption of Ontology Models in teaching programming languages for computer science newbies. This model represented the structured content for Java programming which led the student to learn program-

ming from scratch using Java programming language. The map by the author in turn explained the structure of the language focusing on many aspects starting from Object-oriented programming to building software. This Ontology mainly depended on the fact that this can be deployed on a system where the student can utilize this in his/her college time. Similar to the approach Sosnovky et al. [2] proposed way an Ontology model to teach and learn C programming for undergraduate students. Wilson et al. [3] proposed ways in which Ontology models can be used in teaching and learning processes. Antoniou et al. [4] describe the evolution of Object Web Language (OWL) language and the details of RDF schema in his book, and the owl format is followed in the entire paper.

Waldron et al. [5]’s book describes the basic concepts of kinematics in detail which is commonly prescribed to teach these concepts. This paper briefly explains the ways in which an Ontology model can be deployed to teach basic kinematics to high school students. Vlieghe and Zamojski’s [6] book—“Towards the Ontology of Teaching” provides a teacher’s perspective of teaching and learning from Ontology related concepts. Grosjean et al. [7]’s the paper on Ontology modeling vividly explains the need for Ontology models and related concepts in learning and practicing medicine.

Beichner et al. [8] studied the ways in which students understand and interpret the concepts in kinematics and their related graphs. Brungardt and Zollman [9] studied the ways in which teaching methods influence the learning curve of high school students by using interactive video modules for teaching kinematics. Rodrigues and Carvalho [10]’s interest made a research paper on which students interrupt the kinematics concepts with popular video games. Lichtenberger et al. [11] discussed the various hindrances occurring in understanding the concepts of kinematics among students by analyzing the test which was given to high school students. Dergham and Gilányi [12] proposed ways in which virtual reality can be used to improve the teaching of kinematics for school students.

3 Preliminaries

The motion of an object in real life is taught with the basic terminologies with discrete cases, and then it is expanded to a continuous phase. This provides a fine basement for students to start with basic mass systems in motion. The above ideology can be explained in analogy with playing separate notes in an instrument and then transforming to play a full-fledged song. Table 1 explains all the terminologies which are commonly used to teach basic kinematics. In the definitions depicted in Table 1, “ t ” refers to time in general, and the frame of reference is kept unaccelerated, i.e. Inertial frame of reference is considered and it obeys Newton’s laws of motion [13].

As mentioned, the laws of motion provide the mathematical background to understand the trajectory of motion which students are taught, and their properties. This plays a major role in reducing the errors by checking with the counter-mathematical equations. By convention, the directions of the objects in motion are represented with a positive sign, and a negative sign is used to represent the direction of the object

Table 1 Preliminary terminologies used in school-level kinematics

Terminology	Definition
Vector quantities	Physical Quantities which have both magnitude and direction. Example: Displacement, Velocity, and Acceleration
Scalar quantities	Physical Quantities which have the only magnitude with no direction mentioned. Examples: Distance and Speed
Frame of reference	A point in a three-dimensional space from which the physical quantities related to an object is measured
Distance (r)	All possible paths between two points in space
Displacement (\vec{s})	The shortest directed distance between two points in space
Speed (v)	The rate of change of distance with respect to time, i.e. $s = \frac{dr}{dt}$
Velocity (\vec{v})	The rate of change of displacement with respect to time, i.e. $\vec{v} = \frac{d\vec{s}}{dt}$
Acceleration (\vec{a})	The rate of change of velocity with respect to time, i.e. $\vec{a} = \frac{d\vec{v}}{dt} = \frac{d^2\vec{s}}{dt^2}$
Retardation	The rate of change of velocity in the direction opposite to the frame of reference. This is also referred to as the negative retardation

in the opposite direction. In addition to this, students are taught the basics of vector algebra to solve the problems in an easy manner.

With reference to the notations from Table 1, we have

$$v = u + at \quad (1)$$

$$s = ut + at^2/2 \quad (2)$$

$$v^2 = u^2 + 2as \quad (3)$$

(where u and v in the equations represent the initial and final velocities or speeds of the objects).

4 Conceptualization

For a better understanding of Ontology by learners, the concepts are mapped in a hierarchical structure with parent (root) classes and the specific cases with subclasses. This Ontology aims to give a clear perspective to the student and the teacher to understand the concepts of kinematics easily, hence accuracy, clarity, and interpretability are increased for better results and understanding. This Ontology has a parent class which expands into several children explaining the situation when a certain parameter changes or if it has been kept constant. The parent class is considered as the

superset of all subclasses. More than 10 equations are used to represent the physical quantities and more than 5 standard definitions are used. The sub-concepts directly explain the situation to the student or the teacher. The domain of Ontology is decided to have all concepts for an object to move in a plane. Figure 1 gives a clear knowledge mapping of the proposed Ontology model.

In general, Ontology is represented in Resource Description Format (RDF), programmed in an eXtensible Markup Language (XML) file. Figure 2 depicts the XML configuration of the given Ontology with the concepts and their division. For explaining the context, time is “*inversely proportional*” to the distance traversed and the distance traversed is “*directly proportional*” to the speed of the object. This provides the background of the model which is coded in the .owl file.

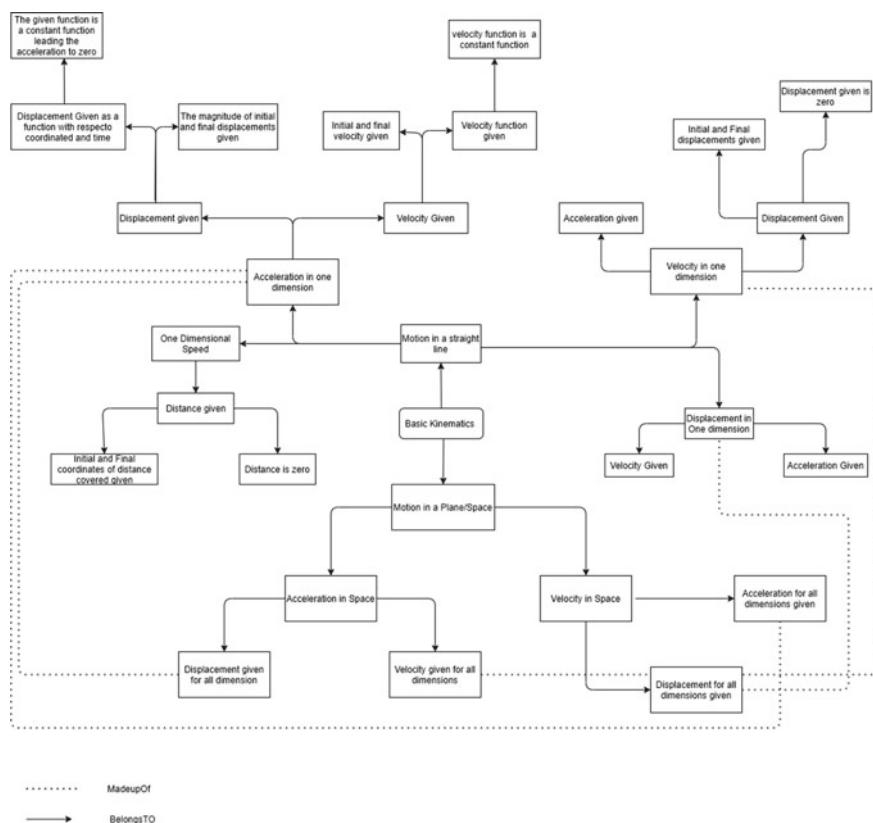


Fig. 1 Knowledge map for the proposed ontology model

```

<?xml version="1.0" encoding="UTF-8"?>
<layout>
  <VSNode splits="0.375 0.625">
    <CNode>
      <Component label="Classes">
        <Property id="pluginId" value="org.protege.editor.owl.OWLAssertedClassHierarchy"/>
      </Component>
      <Component label="Object properties">
        <Property id="pluginId" value="org.protege.editor.owl.OWLObjectPropertyTree"/>
      </Component>
      <Component label="Data properties">
        <Property id="pluginId" value="org.protege.editor.owl.OWLDataPropertyTree"/>
      </Component>
      <Component label="Annotation properties">
        <Property id="pluginId" value="org.protege.editor.owl.OWLAnnotationPropertyTree"/>
      </Component>
      <Component label="Datatypes">
        <Property id="pluginId" value="org.protege.editor.owl.OWLDatatypeList"/>
      </Component>
      <Component label="Individuals">
        <Property id="pluginId" value="org.protege.editor.owl.OWLIndividualsList"/>
      </Component>
    </CNode>
    <CNode>
      <Component label="Selected entity">
        <Property id="pluginId" value="org.protege.editor.owl.SelectedEntityView"/>
      </Component>
    </CNode>
  </VSNode>
</layout>

```

Fig. 2 XML configuration of the ontology created with Protégé

5 Implementation

In general, an Ontology model consists of a parent concept which encapsulates all the child concepts and displays the final relation to the parent by the given relation. This hierarchical system is properly coded in a .owl format model. The whole .owl file is created and formatted using Protégé—a user-interactive GUI to create and edit ontologies. The model prescribed has a parent concept named Basic Kinematics which in turn spans into two main sub-concepts: motion in a straight line and motion in a Space/Plane. The second prescribed sub-concepts are in turn related to the sub-concept one, which discloses the factors which are considered during calculations. Figure 1 explains the schematic representation of all concepts in the proposed Ontology model. The class hierarchy of the model is visualized using Protégé and Fig. 2 provides the class hierarchy of the model.

Taxonomy Employed: Taxonomy is the practice of classifying objects based on their similarities. It follows a hierarchical scheme starting from a common root ending in different leaves covering enormous data which depends on various factors. The child concept of the root concept is said to be derived, and the concept which has a child is said to be the parent concept. Each concept is named with a unique name so that there is no reluctance between different concepts. The hierarchical objects are defined as follows.

Level 4 describes the root concept which acts as the head of the concepts included in ontology, and it is the superset of all sub-concepts and leaf concepts of ontology.

Level 3 provides the outline of the concepts used in Basic Kinematics and their related conditions. Level 3 entities also act as a tool to revise the theory which students have learnt in their classrooms. Level 2 provides the conditions which employ the specific daily-life conditions in which the problems are solved, and the last level, Level 1, of Ontology relates the given data with the mathematical equations.

As described in Fig. 1, for removing redundancy among the sub-concepts, the key words are coded in different caps. The readers can understand the context for the given three quantities, at least one parameter becomes constant when the object changes its motion. For example, when the object is in uniform motion, the velocity of the object becomes constant and its acceleration becomes zero. This applies to the retardation of the object as well. Summarizing the discussion, the sub-concepts represent the special cases which are related to mathematical deductions.

6 Evaluation of Ontology

As the teaching process varies according to its prerequisites and the literary language used, ontology's reusability plays a major role in the deployment of models and its usage in daily life. Semiotics method is used in the paper to evaluate the proposed ontology both qualitatively and quantitatively. For quantitative analysis, the reuse and reference ratio of ontologies are calculated, and the values are compared with the Ontology 101 Model [13] for understanding the schematic and proposing ways to reinforce changes for better usability.

$$\text{Reuse Ratio} = \frac{\text{Number of reused elements}}{\text{Total number of elements in the domain}} \quad (4)$$

$$\text{Reference Ratio} = \frac{\text{Number of referred elements}}{\text{Total number of elements reused}} \quad (5)$$

From Eqs. (4) and (5), the model has a reuse ratio of 0.076 and a null reference ratio. From comparison with the Ontology 101 Model, the proposed model falls under the ideal ontology model category. From Table 2 and Table 3, the parameters from the analysis are taken and analyzed for better enhancement. From Table 2, it is clearly understandable that the proposed model satisfies the ideal conditions for an Ontology model, and the numerical references serve as a great evidence to the effective deployment of models in schools.

The model was tested by 50 candidates who are well versed in the domain and participated in the detailed survey analysis of the model, and the results are presented

Table 2 Qualitative analysis of Ontology

Classes	Subclasses	Attributes	Leaf classes	Reuse ratio	Reference ratio
32	31	6	13	0.076	0

Table 3 Qualitative Parameters

Qualitative parameters	Excellent	Good	Medium	Low
Accuracy	29	15	6	0
Understandability	32	10	6	2
Visualization	25	15	9	1
Relevance	26	14	10	0
Interpretability	30	12	8	0
Clarity	34	9	7	0
Comprehensiveness	28	11	10	1
Concepts	36	10	4	0

in Table 3. The results were found that the accuracy and understandability of the model are high compared to conventional knowledge mapping done in schools. As we can clearly see, parameters such as accuracy and understandability provide a greater perception that the model is well structured and useful for the student to understand concepts properly. The values corresponding to clarity and relevance show that the model is understandable by teachers and instructors, and it provides ease in teaching.

7 Conclusion

A Domain Ontology for teaching basic kinematics concepts has been proposed. On an experimental basis, it is a proven fact that Ontology models help in improving the accuracy and understanding of teaching. To learn a concept like kinematics which applies in day-to-day life, the proposed Ontology provides an effective way to teachers and instructors to teach in a more efficient way. The proposed Ontology model gives a chance to the learners to remember special cases while learning the concepts and applying them in mathematical problems. This model provides a pathway for Information and learning systems to infer and reason specialized concepts in kinematics. The model is visualized properly with XML classifications and a knowledge map is also provided for better interpretation. This model is designed in a manner that it can be reformulated to any other concepts which are taught in kinematics. The reuse ratio of 0.076 and a null reference ratio prove that the proposed model is an ideal model under the purview of Ontology 101 Model, promoting this model for teaching and educational purposes.

References

1. Ganapathi, G., Lourdusamy, R., Rajaram, V.: Towards ontology development for teaching programming language. In: World Congress on Engineering (2011, July)
2. Sosnovsky, S., Gavrilova, T.: Development of educational ontology for C-programming (2006)
3. Wilson, R.: The role of ontologies in teaching and learning. Tech Watch Reports (2004)
4. Antoniou, G., Van Harmelen, F.: Web ontology language: Owl. In: Handbook on Ontologies, pp. 67–92. Springer, Berlin, Heidelberg (2004)
5. Waldron, K.J., Kinzel, G.L., Agrawal, S.K.: Kinematics, dynamics, and design of machinery. Wiley (2016)
6. Vlieghe, J., Zamojski, P.: Towards an Ontology of Teaching, p. 113. Springer International Publishing (2019)
7. Grosjean, J., Merabti, T., Griffon, N., Dahamna, B., Darmoni, S.J.: Teaching medicine with a terminology/ontology portal. Stud. Health Technol. Inf. **180**, 949–53 (2012)
8. Beichner, R.J.: Testing student interpretation of kinematics graphs. Am. J. Phys. **62**(8), 750–762 (1994)
9. Brungardt, J.B., Zollman, D.: Influence of interactive videodisc instruction using simultaneous-time analysis on kinematics graphing skills of high school physics students. J. Res. Sci. Teach. **32**(8), 855–869 (1995)
10. Rodrigues, M., Carvalho, P.S.: Teaching physics with Angry Birds: exploring the kinematics and dynamics of the game. Phys. Educ. **48**(4), 431 (2013)
11. Lichtenberger, A., Wagner, C., Hofer, S.I., Stern, E., Vaterlaus, A.: Validation and structural analysis of the kinematics concept test. Phys. Rev. Phys. Educ. Res. **13**(1), (2017)
12. Dergahm, M., Gilányi, A.: Application of virtual reality in kinematics education. In: 2019 10th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), pp. 107–112. IEEE (2019, October)
13. Newton, Issac, Laws of motion in classical physics 1867

A Deep Neural Approach Toward Staining and Tinting of Monochrome Images



Debosmit Neogi, Nataraj Das, and Suman Deb

Abstract The paper presents a system that colorizes monochrome images driven by deep neural networks without any additional supportive aid. Throughout the paper, we explore feature engineering, transfer learning, color spaces and color channels, and various architectural neural frameworks. Eventually, stained and tinted colorized images or image sequences (video) have been derived from monochrome images that are more aesthetically pleasing, information rich, and more vivid with higher traceability. The experimental results illustrate the viability of our methodology and divulge a favorable route for future work.

Keywords Image colorization · Pixel stain and tint

1 Introduction

Deep Learning is a rapidly growing domain of Machine Learning, which has showcased its prominence in the ever-growing field of image processing and computer vision.

Colorization of black and white or monochrome images has been a subject of numerous research within computer vision and machine learning communities.

With OpenCV 4.4.0.42, we come across pre-erudite networks with popular deep learning frameworks. The actuality that the frameworks are pre-erudite infers that we don't need to devote much time for pedagogical functioning of the framework—rather, we can initiate a forward pass and get a hold of the asset of the yield, to approach to a converging resolution within our neural framework.

OpenCV does not tend to be a contrivance for the pedagogical functioning of networks; by now, great frameworks are at one's fingertips for that purpose. We have seen that networks (such as CNN) have been utilized as classifiers. It builds a judicious perception that OpenCV has a Deep Learning module that can be anchored easily within the OpenCV environment.

D. Neogi · N. Das · S. Deb (✉)

Computer Science and Engg. Dept, National Institute of Technology Agartala, Jirania, India

2 Related Works

Our project, to some extent, was inspired from the works of Zhang et al. [1], where they approached grayscale image colorization by hallucinating potential colors of the original image. We have used their model as our base and worked on top of it. We took inspiration from the works of Jheng-Wei Su et al. [2], where they used an architecture that leverages an off-the-shelf object detector to obtain cropped object images. It also uses an instance colorization network to extract object-level features. They applied a fusion model to full object-level and image-level features to predict the final colors. We also closely studied the work of Jeff Hwang et al. [3] where they proposed a CNN-based architecture for colorizing black and white images. Our framework also reflects some ideas of Alex Krizhevsky and Ilya Sutskever a large deep CNN was trained to categorize 1.2 million high-resolution images that achieved a percentage error of about 37.5% that was at par as compared to previous state of the art [4]. Our work has high potential in manipulating the time series of SAR images of various landscapes and terrains as, for example, SAR images of ocean beds [5]. The proposed neural approach also tends to develop higher traceability of SAR images despite speckle noise [6], which is directly proportional to pixel intensities. This neural architecture can be applied with computational intelligence (Neuro-Fuzzy Logic System) [7] that includes preserving edges and texture information [7].

3 Approach

In order to stain our subject image, we firstly used an extension that caters to our need by providing fast and efficient operations on arrays of homogeneous data. In addition to that, we used another extension of Python. It is particularly taken into use keeping in mind its ability to tackle and solve computer vision problems. The extensions are chosen such that they are highly optimized for numerical operations with a MATLAB-style syntax as well as multidimensional homogeneous arrays.

3.1 Loading the Model

Rather than creating a new model and training it from scratch, we intended to use a pre-erudite Caffe model. Thus, we simply loaded and read a serialized Caffe model from the disk directly. And once the model has been loaded, the subject image has been forwardly propagated inside the Caffe model and the desired classification has been obtained. Inside the Caffe model, we have

- prototxt: The path to the Caffe “deploy” prototxt file. Precisely, it is a model definition where we choose a CNN architecture, and we define its parameter in a configuration file.

- model: The pre-trained Caffe model (i.e. the network weight themselves).
- npy: It is a NumPy file that stores the cluster center points in NumPy format. It consists of 313 cluster kernels, i.e. (0–312).

3.2 1 × 1 Convolutional Cluster Center to the Model

Convolutional cluster center employs deep neural frameworks to take in attribute portrayal which is satisfactory for clustering [8]. Stacked auto encryptor and k-means clustering are one of the favored categorical algorithms for deep clustering. The cluster center serves a purpose that incorporates both clustering loss and reconstruction loss so that the cluster deployment could be grasped simultaneously by the framework along with the representation of attributes [3]. Deep Convolutional Center-Based Clustering (DCCBC) deploys a novel reconstruction loss on the basis of cluster centers rather than usual reconstruction loss. Precise attribute representation gets constructed by the model. The cluster center also traps the primary attributes of instances in different clusters. The cluster center is highly enhanced by gradient descent methodology like mini-batch or stochastic gradient descent and backpropagation. This cluster center accomplished competitive results and outmatched the state-of-the-art clustering algorithms. After extracting the layer Identification from the Caffe model (extracted from the last layer of the network), we are transposing our NumPy file and reshaping the cluster center stored in them as 1×1 matrix. A 1×1 convolutional layer can be used that offers merging operation according to respective channels, often called attribute map amalgamating ensuing a projection layer. This simple methodology is highly optimized for dimensionality depletion, decreasing the number of attribute maps while retaining their arresting attributes. The creation of one-to-one prognostication of the attribute maps to pool attributes across channels or to increase the number of attribute maps, such as after conventional merging layers, can be accomplished by deploying the cluster center. After obtaining the coveted cluster, it has been added to our model.

3.3 BGR to LAB Color Space

3.3.1 BGR Image Representation:

A BGR image can be considered as three distinct images (a red channel image, a green channel image, and a blue channel image) stacked on top of each other, and when fed into the red, green, and blue inputs of a color monitor, it generates a tinted image on the screen [2]. A BGR image constitutes a 3-dimensional array of shape $M \times N \times 3$. Each color pixel in a BGR image is allied with three values which correspond to red, blue, and green channel components of the BGR image at a certain spatiotemporal location [9]. So, when the red, green, and blue intensities within each

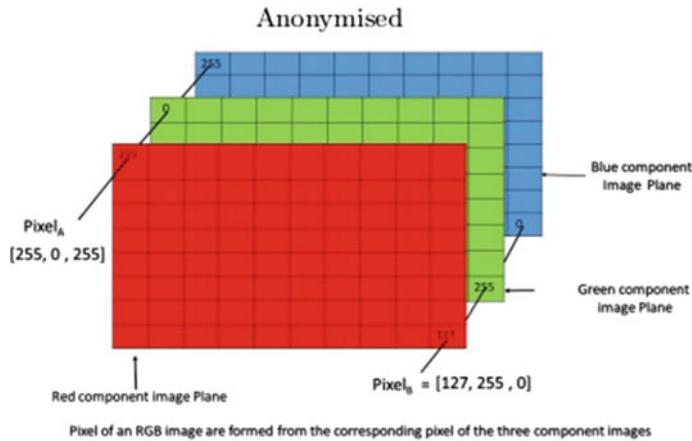


Fig. 1 BGR image pixel representation (<https://www.geeksforgeeks.org/matlab-rgb-image-representation>)

pixel, at specified color plane at the pixel's location, are combined together, the resultant color of that distinct pixel gets determined. Due to the level of accuracy through which a real-life image can be replicated, occasionally a BGR image is cited as a true color image that ultimately fetched the nickname “true color image”.

Now our subject image has been loaded and the scaling operation has been performed. The scaling (8-bit integral representation to floating-point value) is obligatory for normalizing the image pixels in the range [0,1]. After normalizing, its existing channels have been transmuted to Lab channels (Fig. 1).

3.3.2 Lab Image Representation:

Lab channels, i.e. L*a*b Coordinates are more effective in achieving the same color across different media. “Lab” representation illustrates a more error-free and pinpoint color space of an image [10]. This method makes use of a 3-axial system as follows:

- “L”: It stands for lightness, which is a measure of how sharp an image is, i.e. the intensity of the vivid colors present in the image.
- “a”: It expounds color from green to red coordinates.
- “b”: It explicates color from blue to yellow coordinates.

Once the above processes are completed, we moved next toward the extraction of the “L” channel only, through array slicing methods (Fig. 2).

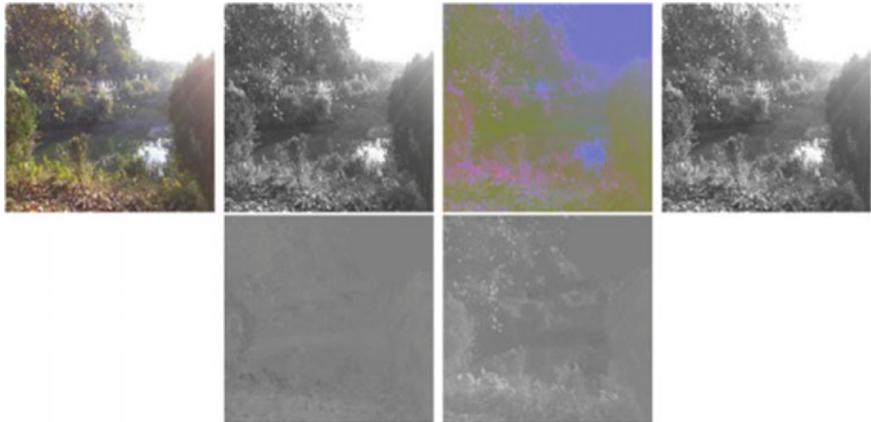


Fig. 2 **a** Original Image with BGR channel, **b** Grayscale version of the original image, **c** BGR image converted to “Lab” channel, **d** Image with extracted “L” channel, **e** Image with extracted “a” channel, **f** Image with extracted, and “b” channel

3.4 Deep Neural Prediction

Caffe is a deep learning framework crafted with expression, speed, and modularity in mind [11]. The architecture of Caffe is exciting as well as expressive. It provides an opportunity for innovation and development. Models and optimization are defined by configuration. So it eliminates the need for hard-coding. It allows quick switching between CPU and GPU which implies smooth running of the models [4].

So, once the “L” channel is extracted, it has been fed to the Caffe model in order to anticipate the “a” and “b” channels [5]. This is the most crucial step behind the whole idea. Once these “a” and “b” channels are predicted, we can convert the predicted “Lab” channels to the “BGR” channel. For this purpose, the anticipated “a” and “b” channels are again resized to the shape of our input image (Fig. 3).

3.5 Concatenation

In order to create the whole image, we need to concatenate the predicted “a” and “b” channels along with the “L” channel. So the “L” channel from the original image (image that was loaded) is again extracted and merged (concatenated) with the predicted “a” and “b” channels using the NumPy array methods [12]. Now we have our total “Lab” image that can now be converted to the BGR channel with ease using the OpenCV library methods (`cv2.cvtColor(Lab, cv2.COLOR_LAB2BGR)`) (Fig. 4).

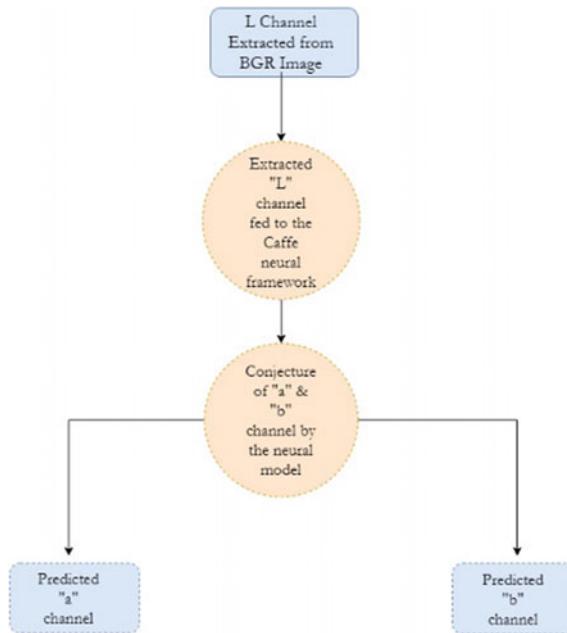


Fig. 3 Process of neural prediction

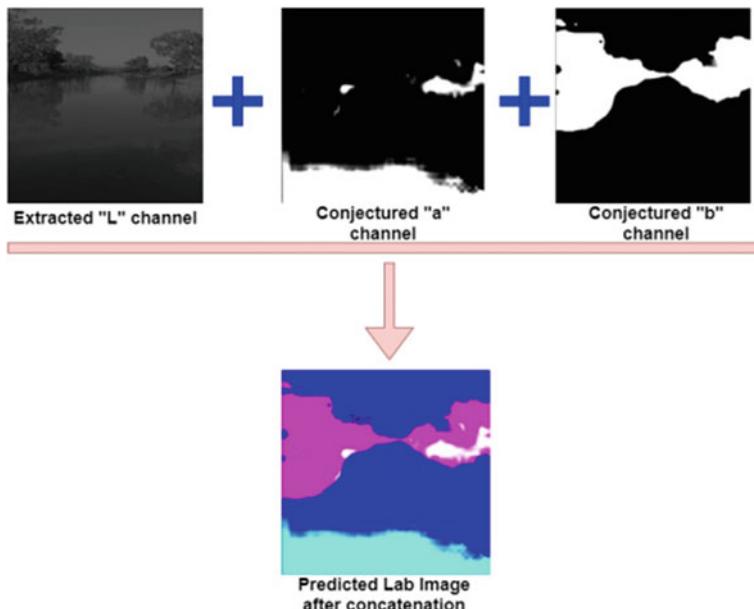


Fig. 4 Concatenation of original "L" channel with conjectured "a" and "b" channels to develop the total conjectured-"Lab" version of the subject image

3.6 8-Bit Integral Representation

The color components of an 8-bit RGB image are integers in the range [0, 255], as, for example, a pixel whose color components are (255,0,0) is displayed as red. The image command displays an RGB image correctly whether its class is double, uint8, or uint16 [13]. The “BGR” image array hence obtained is in the form of floating-point values. So it needs to be converted to an 8-bit integral representation. After the final conversion of the array element’s datatype, we have our stained image ready (Fig. 5).

4 Dataset

We have used the ImageNet dataset for the purpose of training our model. The ImageNet dataset is a large collection of photographs. The dataset was designed for the purpose of training computer vision and deep learning models. According to the ImageNet homepage, the dataset consists of more than 14 million images and more than 21 thousand groups or classes, and more than 1 million images that have bounding box annotations.

5 Further Experimentation

We extended our model implementation toward staining and tinting of real-time subjects [14]. Video is a collection of pictures played within a time interval which is less than human eye persistence time.

Number Of Frames Per Second Frame rate can be expounded as the number of still pictures per unit time of video that usually fluctuate from six or eight frames per second for old mechanical cameras to one hundred and twenty or more frames per second for new competent cameras.

Aspect ratio The commensurate tie-in between the width and height of video screens and video picture elements gets narrated by the term “aspect ratio”. All video formats can be illustrated by a correspondence between height and width, since they are conventionally rectangular. Nevertheless, videos in monitors are constituted through pixels (usually square), yet the pixels have a non-squared aspect ratio.

Color Model And Depth Within a pixel, the cardinality of perceptible colors rests upon the COLOR DEPTH expressed in the number of bits per pixel. The color model is, therefore, the video color constitution and maps ciphered values to perceptible colors transcribed by the system. Thus, a video can be thought of as an image sequence, and each image from the sequence can be tinted to obtain a stained real-time image sequence (video) [15]. So we took our model forward, so that it can stain each of the images in the video image sequence. We obtained every single image using an infinite

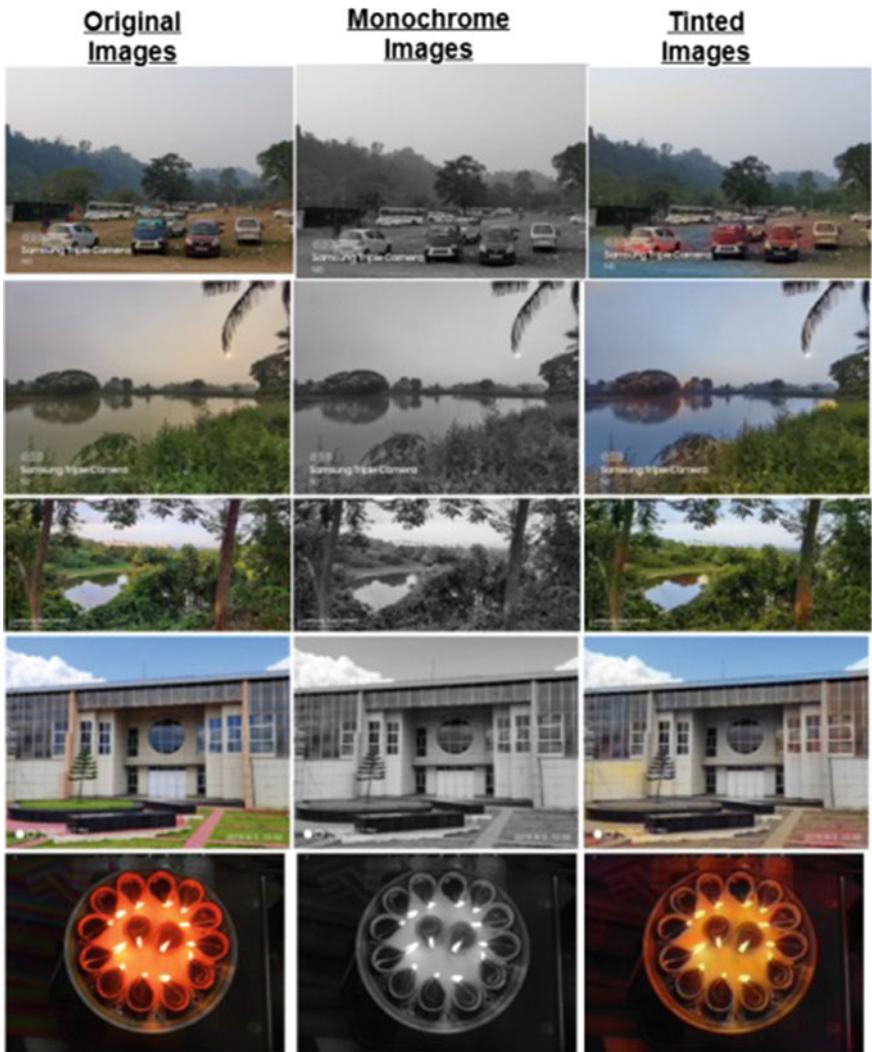


Fig. 5 Deployment of our model to stain some test images

loop (terminates at a particular ASCII value) using OpenCV methods and fed it to the Caffe model to conjecture the “a” and “b” channels from the “L” channel of each image. Final concatenation and conversion from the “Lab” channel to the “BGR” channel results in the total tinting of each image which when played using a loop creates a colored video sequence. However, during real-time object colorization, a problem of lagging video (image sequence) had been encountered (Figs. 6 and 7).

From the above graph, it can be concluded that the frame is non-uniform in a subsequent time interval. Till the time interval of twenty-one seconds, the frame

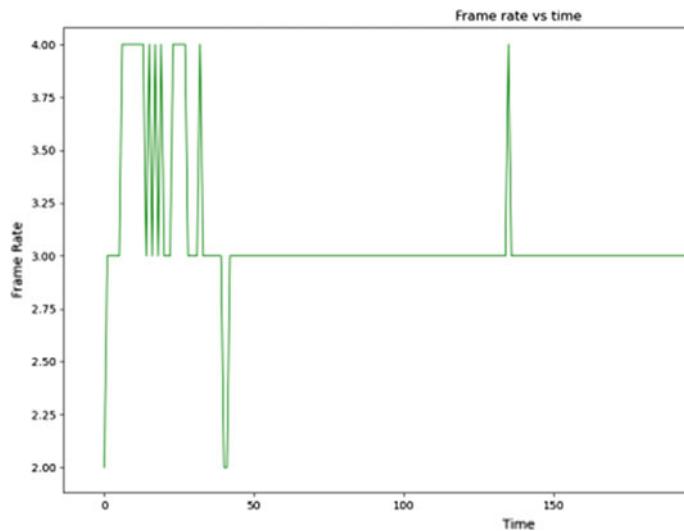


Fig. 6 Frame Rate in real-time image sequence

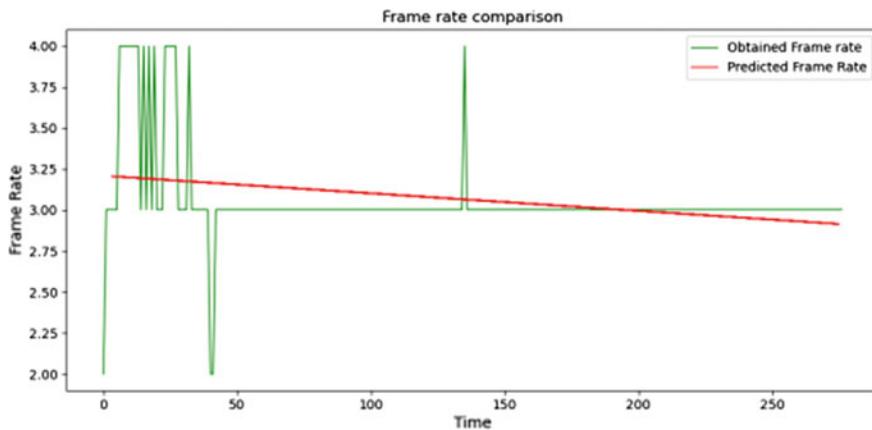


Fig. 7 Best fit line obtained through regression

rate remains constant. So within that time frame, one can have a smooth non-lagging image sequence. But after twenty seconds, there is a significant decrease in the frame rate by two frames per second. After that, again it catches back its height and falls again. Considering the ideal situation where the video would not lag, we would have got a curve parallel to the Time axis. But after plotting the best fit line using the data points used to plot the above "Time versus Frame Rate" graph, we got the following equation of the straight line:

$$Y = -0.0011X + 3.2077$$

Our best fit line equation estimates the actual curve with a precision of 0.86. In other words, over all the data points considered, the total error recorded is 0.14. The negative slope of the graph clearly portrays the fact that the “Frame Rate” decreases with time. This mathematically proves the lagging phenomenon, which otherwise would have given a graph where “Frame Rate” is independent of “Time”, which clearly is not the case.

And this phenomena continues throughout the whole real-time colorization procedure. One reason for this non-uniformity is the processor cache speed and computing potential. Deep neural frameworks are highly hardware-dependent and can literally be optimized through hardware up-gradation. That is the reason for the lagging image sequence.

Viable Usage

We have been able to think of the immense importance of our work in solving real-life problems. Our approach of colorizing a real-time image sequence can be used in surveillance cameras. The surveillance cameras use Infrared(IR) cameras to capture the footage. The IR cameras have sensors that can gauge the amount of light entering the lens aperture and based on that, they switch on to night vision mode. If there is not enough ambient light hitting the sensor, the camera will record in IR or night vision mode and this can lead to missing out on important details. So we can deploy our model for tinting the footage from the surveillance camera, so that the minute details are not missed out, due to a significant increase in traceability (Fig. 8).

6 Conclusion and Future Work

Through our experiments, we have attempted to use deep neural networks for recoloring monochrome and black and white images that are actually reasonable and close to reality. We have further extended our works to real-time video recoloring which potentially can be used to recolorize surveillance camera footage taken under poor lighting for better clarity.

Our works have opened the gates for several future works which can be eventually addressed. One such problem that we touched upon earlier is the lagging of video output. It can be addressed with some better and efficient architecture that synchronizes better. Also, we have seen some discolorization of objects in the real-time video output which can also be addressed in the future with an improved version of our work.

References

1. Zhang, R., Isola, P., Efros, A.A.: Colorful Image Colorization. [arXiv:1603.08511](https://arxiv.org/abs/1603.08511) [cs.CV]
2. Su, J.-W., Chu, H.-K., Huang, J.-B.: Instance-aware Image Colorization. [arXiv:2005.10825v1](https://arxiv.org/abs/2005.10825v1)



Fig. 8 Tinted surveillance footage (<https://www.amazon.in>, <https://www.shutterstock.com>, <https://www.ndtv.com>)

3. Hwang, J., Zhou, Y.: Image Colorization with Deep Convolutional Neural Networks. Stanford University. (cs231n.stanford.edu/reports/2016/pdfs/219_Report.pdf)
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
5. Vijayakumar, S., Swarnalatha, P., Rukmini, S.: A neural network classification approach for oil spill detection on SAR images. *IJOAB J* **7**(5), 225–235 (2016)
6. Singanamalla, V., Vaithyanathan, S.: Neuro-fuzzy approach for speckle noise reduction in SAR images. In: Communications in Computer and Information Science Recent Trends in Image Processing and Pattern Recognition, pp. 251–260 (2017). <https://doi.org/10.1007/978-981-10-4859-3-23>
7. Vijayakumar, S., Santhi, V.: Speckle noise reduction in SAR images using fuzzy inference system. *Int. J. Fuzzy Syst. Appl.* **8**, 60–83 (2019). <https://doi.org/10.4018/IJFSA.2019100104>
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015). [arXiv:1512.03385](https://arxiv.org/abs/1512.03385)
9. <https://www.geeksforgeeks.org/matlab-rgb-image-representation/>
10. <https://www.pyimagesearch.com/2019/02/25/black-and-white-image-colorization-with-opencv-and-deep-learning/>
11. Nguyen, T., Mori, K., Thawonmas, R.: Image Colorization Using a Deep Convolutional Neural Network (2016)
12. <https://www.codespeedy.com/automatic-colorization-of-black-and-white-images-using-ml-in-python/>

13. Larsson, G., Maire, M., Shakhnarovich, G.: Learning Representations for Automatic Colorization (2016)
14. Aranburu, A., Giri, A., Gutierrez, R., Reeves, S.: Image Colorization using Convolutional Neural Networks (2017)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). [arXiv:1409.1556](https://arxiv.org/abs/1409.1556)

Evaluation of IoT Based Automatic Headlight Dimmer Systems



Kanika Gandhi, Karanpartap Singh Aulakh, Jobanpreet Singh Thind, Gurpreet Singh Kharoud, and Seemu Sharma

Abstract IoT has taken over in every aspect of our daily lives. Domains stretch across a vast horizon, starting from home appliances, electronic devices, home systems to industrially manufactured products such as robots, equipment, and even automobiles. Implementation of IoT in automobiles such as self-driving cars is such a domain that is risen in the past few years by a notch but often one area has not been explored to such an extent, and hence poses as a flaw in the manufacturing of vehicles are the headlamps. As per statistics, it has been found that unawareness of the driver and incorrect selection of beam during driving has posed as a major security threat to the traveler himself and the surrounding cars leading to a lot of accidents and in turn raising concern. This paper offers a comparative study of the existing technologies deployed in automatic headlight dimmer systems along with pondering upon the need of automatically controlled systems to switch headlights in automobiles, what are the foundation grounds for headlamp alteration systems, what are the detrimental effects to safety when any driver uses the wrong beam in unsupportable conditions in a vehicle.

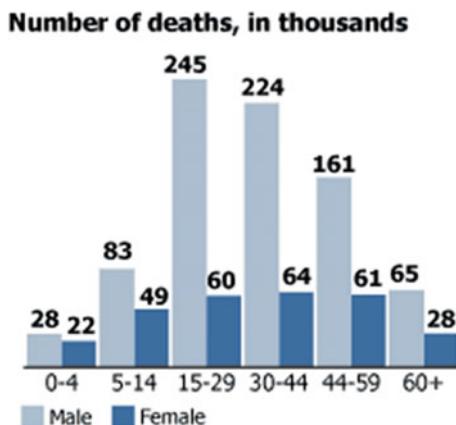
Keywords Headlight dimmer systems · High beam alteration · Troxler effect due to high beam · Harms of high beams · Beam alteration

1 Introduction

IoT has opened multiple doors for research, development, and improvement in our day to day life. As much as man needs technology, the technology and human relation plays a vulnerable yet very vital role in everyday routine. Internet of Things clearly describes how every object can be specified as a complex medley of different technologies. Posing as the catalyst for the digitalization of the future, IoT has already become a big part of our generation. Novel works such as Smart home systems, Smart cities, Smart robotics, Smart vehicles have revolutionized the movement for

K. Gandhi · K. S. Aulakh · J. S. Thind · G. S. Kharoud · S. Sharma (✉)
Thapar Institute of Engineering and Technology, Patiala, Punjab, India
e-mail: seemu.sharma@thapar.edu

Fig. 1 Death Toll Statistics of road accidents worldwide [1]



building a technologically equipped environment. IoT, as one domain of automobiles, has shown various developments by making them more nature and driver friendly. However, the headlamp is usually left untouched in the digitalization of these vehicles. Headlamp plays a vital role in driving and is an area that is often overlooked upon but the repercussions of mismanagement of beam usage can lead to dire disasters and can take driver's or passenger's life (Fig. 1).

Nowadays, a rising concern is the increase in number of road accidents that cause different kinds of injuries, sometimes fatal. The main reason for this problem has been identified as essential to address in the field of automobiles and it has been discovered that more than percent of accidents on the road happen due to wrong use of beam or careless usage and flicking of the beam type emitted from the headlamp of the vehicles. High beam is very dangerous and should only be used in situations where there will be no consequences on the oncoming vehicles or on the driver themselves.

Human error while driving allows driver to use the wrong beam in the wrong situations, hence causing a safety violation to their own and nearby cars. The reason for the same is the confusion caused to one driver due to the sharp rays emitted from the headlight of another vehicle approaching from the opposite side. This is known as dazzling the driver or the fade effect [2]. The scientific explanation for this phenomenon [2] goes by the theory that when there is a vehicle on the road and a strong beam hits the driver, it is not that light at the instant that causes trouble. Once the vehicle on the opposite road moves out of sight the driver experience strong black spots or darkness in his/her vision. This darkness leads to driving on the road without proper view or idea of what's ahead also known as blind driving. As the reaction time extends up to 1.4 s for the driver, the distance of the blind driving duration can extend based on situation to situation, and hence pose a fatal threat to the driver's safety and the passenger's along. This is known as "TROXLER EFFECT" [3].

It is for this reason, there is a system required that will regulate the beam intensity of a car to avoid accidents. This survey focuses mainly on comparative

review of different kinds of beam monitoring systems implemented till date, their implementation technologies, and finding a compatible system in today's world.

2 Background

This section explains the main causes and effect relation for the need of such beam switch systems. There are three main types of scenarios under which the beam directly affects the driver and surroundings which are discussed below.

a. High beams used in crowded city areas or mountain roads with high density of population in the near perimeters.

In crowded cities, where one vehicle is in line with another, it is highly risky to use high beam as it is profoundly easy to cause the dazzling effect or blinding to the driver in the opposite lane as the light of the oncoming vehicle is deflected from its path due to refraction and the intensity is doubled by twice, causing a white like spectrum to appear in the eyes of the person traveling on the opposite side. This momentary blindness gives rise to blanking out while driving as the cognitive senses fail to take an action as per the stimulus received (in this case high intensity light), hence causing an accident.

b. The high beams being used on highways with bright street lights.

Although it is advised to use high beam on highways, similar to the situation that arose when high beam can prove fatal in crowded cities, upon using high beam on already illuminated streets, instead of adding to the visibility, the opposite takes place due to the presence of dazzling effect. The driver traveling on the adjacent road approaching in the contradicting direction, will get dazzled by such intense amount of light produced by the blend of street and headlamps.

c. High beams being used during rainy or moist weather.

On the roads, it is usually advised to drive with high beam to illuminate the distance ahead of the vehicle if the weather is partially clear. During rains, snow, fog or storms small droplets of the water molecules in the air tend to reflect the light emitted from one's own vehicle, hence causing a strong glare and impairing the vision.

High beam does not only have physical effects depending on the surroundings, but also long term effects of wrong beam usage can medically deter the health and livelihood of people.

2.1 What Are the Detrimental Health Effects of Using High Beam?

Medically a few terms have been assigned to the instances with consequences due to high beam usage as per irresponsibility.

1. **Light Pollution:** It is believed to be both manmade and natural phenomenon where the excessive production or overlapping of lights can cause glare to the human eye, hence blurring vision [4]. The high beam from headlamps when used during rain or fog causes splitting of particles, contributing in light pollution along with street lights, hence risking the safety of the cars traveling opposite to the one bearing extreme beam intensity.

2. **In-attentional blindness:** Due to constant disturbance in vision, according to statistics, people tend to develop in attentional blindness or inattentive blindness which refers to missing objects in plain sight [5]. This can occur with simple mistake in paying attention to the road due to momentarily lag between cognitive load and sensory response. In order to serve a move towards the impulse, as reflex, people end up abruptly pressing brakes, flickering their beams, hastily speeding up or even crashing into cars ahead.

3. **Troxler fading:** According to medical sciences, when the light from a high beam is caught by the eye (as shown in Fig. 2), it passes the fovea and fills in the blind spot of our eye which after the lag of 5 s starts getting distorting, leading to the blurriness or haziness of the vehicle with high beam. Due to this malfunction, people tend to stop their cars midway without prior signaling or drive straight into the car in front and get injured because of the collision [6].

Fig. 2 Human retina

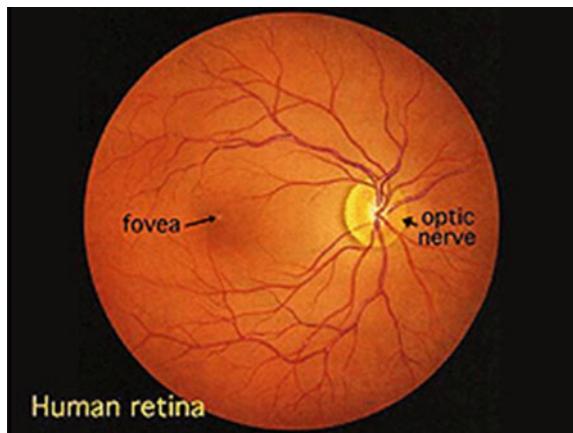


Table 1 Research questions

Question	Motivation
1. How important is it to shed light upon the use of high beam in wrong situations?	With the current lifestyle, almost every household has some mode of personal transport and with newly emerging models, the risk of human error while driving still persists to a great degree. It's a rising concern and must be addressed
2. How harmful is the use of high beam in wrong situations?	The result of high beam can very much so result in minor injuries to severe fatalities
3. Can IoT systems be a good solution to curb this issue of wrong beam at the wrong time?	Yes, technology can help monitor, indicate, alternate and place a threshold on the driver's control when the activity goes out of safety and into danger

3 Methodology for Research

The methodology followed a simple procedure of collection of data (quantitative and qualitative) and formation of the research questions and sections.

3.1 *Review and Selection*

Platforms like Google Scholar, Researchgate, IEEE Xplore, ACM Journal were skimmed through to look for papers containing similar surveys and topic based analysis to shortlist a set of questions for the further movement in process.

3.2 *Question Set*

The questions along with their motivation factors are listed below in Table 1.

3.3 *Selection Procedure*

Keywords such as: Headlight Dimmer Systems, High Beam alteration, Troxler effect due to high beam, Harms of high beams, Beam alteration were used majorly to skim through available papers. Papers were refined according to individual abstract before shortlisting a few.

4 Results and Discussions

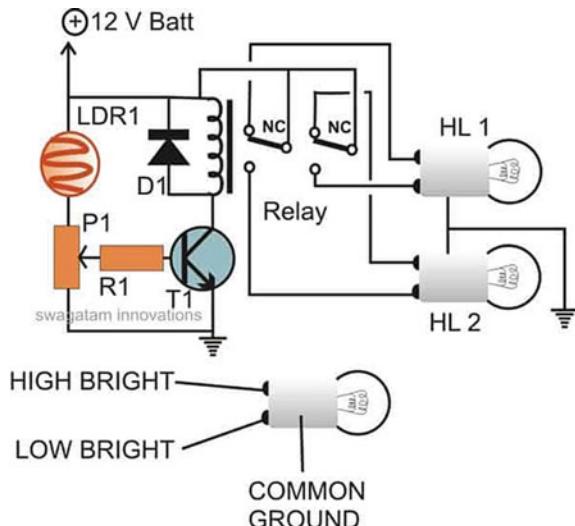
As per different research papers and surveys, the following systems have been proposed for headlight dimming and monitoring during driving.

In the early 1900s, pivotal systems using trunk and attachment with the body of the car [7] was implemented to deploy a lever based automatic system for toggling beam.

In 1917, automatic dimmer systems were developed by arranging switches serially with lamps that allowed the system to sense and pass electric impulses generated upon the shift of magnetization in the switch [8]. In the following years, along with electrical deployed switches, selenium cells paved a new path to link fluid mechanism with dimmer circuits [9] (as shown in Fig. 3) Further, another sophisticated update for the older system had been presented using insulated rod [10] which was connected to the circuit of brake and clutch so that upon activation of either one, the dimmer worked in coordination.

From the late 1950s all the way to present year, optical system are arranged either singularly [11] or in combination with image processing systems [12] or piezoelectric systems [13] to detect illumination further down the road and send signals to the dimmer circuit for response. Towards the start of 2000s, image sensors arranged with CMOS sensors [14], micro cameras [15], and lastly multilevel thresholding [16] provide a platform to perform digital image processing which increased the sensitivity of detection of approaching vehicles to help switch beam by receiving input signals. By the late 2010, blob detection using MTT algorithms [17] took a toll on dimmer systems as it interlinked sensor technology with computer vision for alerting the vehicles and initiating beam toggle mechanism. In the following year, headlight dimmer circuits rose to great popularity as they included [18] intelligent

Fig. 3 Automatic Dimmer System using electric switch, LDR and lamps [9]



vehicle and obstacle detection technologies for headlight switching controls using machine learning. In the late 2015, Zigbee technology blended with V2V and VANET technology proved to be extremely precise way to switch beams due to its ability to detect small targets [19].

Further, in the following years, paired with LDR, the Zigbee technology was a part of light sensing and target detection to produce impulse for beam flip to the vehicle dimmer circuits [20]. Then, Automatic Dimmer Systems using Arduino technology were developed. Arduino initially was linked with LDR [21] to program the input based on light sensing and output based on Arduino signal transferred to relay. To upgrade this technology, paired with gyroscope, the dimmer circuit functions as deflection meter to detect the horizontal angular distance between approaching vehicle to spark the switch case mechanism to take decisions as per surroundings [22]. Later, Automatic Dimmer Systems using transistors were used. Transistors have been an essential part for choosing modes between beams either coupled with operational relay. Three types of transistors are usually preferred, namely photocell transistor [23], junction transistors [24, 25], and photo-resistant transistors. In combination with the dimmer circuits and LDR, transistors aid operational functionality of these systems to reduce glare.

4.1 Industrially Most Favorable Headlight Dimming System Techniques

Table 2 represents a comparative analysis of different automatic dimmer systems discussed in the above section and in the following section.

1. Photo resistor in defogger grill: Nowadays the cars have installed the photo resistors in the defogger grill of the cars that work on the basis of threshold techniques. One such system is proposed in [26]. The photo resistor has set threshold and upon the reflection of ambient light (in this case dawn and dusk) upon the resistor, the threshold value is broken and then the lights switch from on to off situation. These cars have beeping systems that alert the driver. If the lights have not been switched off even after the car has been switched off, they automatically switch off when the car is turned off.

2. AHO: Automatic headlight on is another system or phenomena which has been introduced in motorbikes, especially according to which, since 2017, none of the motorbikes will have the ability to manually switch off the headlights once the engine is revved and going in bikes. This kind of a system is presented in [27].

This was done to prevent accidents as many times the drivers had shut off their headlight in the bikes, thus creating a state of danger for the driver. The drive does not have the control of the headlights once the engine starts, he/she can only switch the beam from high to low but cannot switch it off.

3. DRLs: Daytime Running Lamps have been presented in [28] and are embedded in cars now as a part of construction. These are present at the front of the cars under

the engine and were brought into use in countries where the daylights were dim. These lights do not illuminate the road instead they use the already existing lights and apply lower power settings to give out dim lights that shine ahead but are not at the full potential like the high beams. They are automatic.

4. ACL: The Automatic Light Control system as depicted in [19] is a system ridden with a time delay of 20 s for reaction. This system is affiliated with an automatic switch on of the headlights upon immediate detection of dark conditions by the sensor present in the instrumentation panel. The system allows the car to switch on the lights even in cases where the light is dark in enclosed places with no light hence acting as an assistance system for the driver.

5. Auto-Lamp system: This feature as presented in [29] is mainly witnessed in vehicles of the Ford company where it poses as a different or alternate version of differentiation from the ACL system mentioned above. The system takes in conditions as inputs using a photocell and decides accordingly. Introduced and modeled in year 1982, it has become a patent of Grand Marquis.

6. Twilight Sentinel: This feature dates back to older than Auto Lamp and ACL systems. As per [30], the twilight sentinel system uses the help of a photocell and an amplifier to measure the light intensity before it switches on the lights. It's a little different from the rest as it provides the enlightening of the reverse lamps along with head lamps and the delay can be set by the owner as per their wish

4.2 *Inferences*

Broadly classified, upon the surrounding factors high beam should only be used in 1 of 16 conditions as per this following table (Table 3).

According to this table, the following rules should be followed for usage of high beam:

- Reduction of beam to low upon detection of the intensity light provided by the tail and head lights of the cars nearby our own vehicle and adjustment of the beam of our vehicle.
- High beam cannot be used on the street when a pedestrian, cyclist or jogger is crossing as it causes the blinding effect that blurs the vision of the ones crossing by.
- Automatic detection of location of the vehicle and hence switching to low beam in crowded areas and high beam on clear highways.
- The restriction over use of high beam during moist or wet weather especially during the night.

Based on the above analysis, there has been a notable change in implementation of such beam dimmer mechanism and hence the following inferences can majorly be drawn:

Table 2 Comparative analysis of different automatic dimmer systems

S.no (1)	Year (2)	Objective(s) addressed (3)	Tools & Technology used (4)	Benefit(s)/Outcome(s) (5)	Real time response (6)
1.	1908	Adjustable Headlight	A pivotal system mechanically linking the car body to the 'trunk' which refers to the lever based system [31]	The mechanism works on direction. According to the lane, the headlight intensity is reduced upon movement of steering wheel	Average
2.	1917	Means of adjusting headlight automatically	A parallel combination of electric lamps in line with the engine of the car and a generator to provide power to the prepared apparatus	The apparatus is powered remotely from inside the car but automatically through fluid mechanism. The beam is switched lower by the current passed through the lamps ignited by the generator and reduces manual movement for accident	Poor
3.	1918	Enabling selenium cell for automatic headlight dimmer	A series combination of current wires with a switch composing a magnet in line with a selenium cell	Due to the magnetization of the switch, upon detection of the incident light, the headlight is dimmed when light falls on to the automobile, avoiding blinding effect for the vehicles on the road	Poor
4.	1925	Automatic headlight dimming attachment for automobiles	A rod ridden with fiber or insulating material attached to a spring which deters as per clutch or brake	The headlight intensity is decreased when the traveler applies brake or pulls the clutch to reduce the chances of accident because the driver does not have to release the steering wheel to use the controls on the car	Average

(continued)

Table 2 (continued)

S.no. (1)	Year (2)	Objective(s) addressed (3)	Tools & Technology used (4)	Benefit(s)/Outcome(s) (5)	Real time response (6)
5.	1953	Automatic headlamp dimming circuit	An optical system that is compromised of a lens system and scanning disks. A photoelectric instrument and peak detector	The headlight intensity is alternated when the contrast between the illuminated light from the other cars and general background is measured reducing glare effect	Average
6.	1954	Automatic headlight controlling device for automobiles	A combination system comprising of relay, dual filament headlight and a selector switch [32]	The system provides an alternate way to operate the headlight switch such as through foot to avoid removing the hand from the steering wheel	Good
7.	1956	Automatic headlight control system by multifilament headlight	The illumination system is made of multifilament headlights associated with lower and upper beam circuits, manual beam selection operator and relay [33]	The headlight is reduced or increased based on illumination hence avoiding blinding effect to oncoming drivers	Average
8.	1960	Relay based dimmer circuit for headlight	A dimmer circuit is made of a dimmer relay and a photo sensitive device [34]	Based on energization of the circuit, if light interrupts the vision of the headlight, the intensity of the headlamp is switched to lower beam to avoid accidents	Poor
9.	1980	Automatic headlight dimmer upon use of windshield wipers	A dimmer circuit consisting of photocell, transistor and operational relay for switching between beams	The circuit lowers the beam upon sensing of approaching nearby vehicles and when the windshield wiper is activated, ensuring low beam during rains too	Poor
10.	1984	Automotive accessory control system	Three relays connected in series to five switches with semiconductor switch which connects further to light detector powered by microcomputer [35]	Successfully toggles the beam upon detection of ambient light in designed radar or if the windshield is activated	Average

(continued)

Table 2 (continued)

S.no. (1)	Year (2)	Objective(s) addressed (3)	Tools & Technology used (4)	Benefit(s)/Outcome(s) (5)	Real time response (6)
11.	1992	Dimming apparatus to reduce glare	Light tube operator; Photo-sensitive transistors parallel attached to three junction transistors [36]	Upon detection of electronic beams, this system has successfully reduced glare effect by 63% in oncoming vehicles	Good
12.	1994	Motor vehicle headlight activation apparatus for inclement weather conditions	Connection between the movement of wipers and the activation of headlights upon the start of wipers	The use of photo resistor sets a threshold level which once passed switches the beam intensity per the situation	Poor
13.	1998	Control system to automatically dim headlight lamps	An image processing system connected with an optical system [37]	Accurately detects when a vehicle is approaching on the road and reduces the beam if high	Good
14.	2000	Anti-blind headlights	Light sensor based [38, 39]	Detection of micro beams per location and radiation which tends to reduce dazzling	Poor
15.	2001	Headlamp switch for driving assistance	Twilight Sentinel technology [30]	Aligned photocell and amplifier uses twilight sentinel technology to accurately give real time response without performance delay to switch beams	Average
16.	2004	High-beam headlamp usage on unlit rural roadways	Survey techniques and graphical comparison for observational purposes [40, 41, 42]	Conduction of an observational analysis to make a conclusion regarding the underuse of high beams on rural roadways in comparison to 1960s and studied how the traffic density affects the beam usage	Good

(continued)

Table 2 (continued)

S.no (1)	Year (2)	Objective(s) addressed (3)	Tools & Technology used (4)	Benefit(s)/Outcome(s) (5)	Real time response (6)
17.	2007	Light has to go where it is needed: Future light based driver assistance systems	CMOS image sensors, AFS, ICL and Image processing [43, 44]	Future scope related introduction of few detailed techniques that can will be the future of automatic headlights and halve the accident rate	Average
18.	2008	Nighttime vehicle detection for driving assistance	B&W micro camera mounted in the windshield area and Digital image processing	The system has could develop a connection between the detection of the vehicle lights and beam control although the system proved to be more accurate for the detection of head lights rather than that of tail lights	Average
19.		Temporal coherence analysis for intelligent headlight control	Temporal coherence as a [45] phenomenon for comparison	The author develops a system using analysis with the temporal coherence phenomenon with a 90% accuracy rate to perform analysis of strong luminance	Good
20.	2009	Nighttime vehicle light detection on a moving vehicle using image segmentation and analysis	Image segmentation; Image analysis; Multilevel thresholding	The system uses multilevel thresholding techniques to find out the tail and rear view lights of a vehicle and successfully detect presence of a vehicle at night	Good
21.	2010	Multiple target tracking for intelligent headlights control	MTT algorithm and computer vision techniques	The use of the algorithms for this project has produced a 90% accuracy rate in detecting blobs without hampering and provides real time results by precise decision making algorithm	Good

(continued)

Table 2 (continued)

S.no (1)	Year (2)	Objective(s) addressed (3)	Tools & Technology used (4)	Benefit(s)/Outcome(s) (5)	Real time response (6)
22.	2011	Intelligent headlight control using learning based approaches	SVM technology and Adaboost [46]	Machine learning techniques is used to achieve an automatic vehicle beam controller. The Ada boost method turned to be a little less stable than the SVM technique	Good
23.	2013	Automatic High Beam Controller for Vehicles	LDR, LM358N and IR technology	The system has used IR technology for detection of the high beam of an oncoming vehicle to automatically reduce the light of its own vehicle	Good
24.		Intelligent automatic high beam light controller	LDR Resistance and Relational comparison between Luminance [47]	The system for a car is capable of detecting vehicle lights from 230 m away and is automatically able to switch the beam from high to low upon detection	Average
25.		Fuzzy head light intensity controller using wireless sensor network	Fuzzy controller based on data captured using WSN (Implemented using MATLAB) [48]	WSN based fuzzy controllers and fuzzy union and fuzzy intersection logic to reduce the chances of blind spot upon the crossover of vehicle and the oncoming vehicle	Average
26.	2015	Design and development of an automatic automobile headlight switching system	LDR, NPN transistor, SPDT relay	A prototype using simpler components so as to detect and switch the headlight just like other similar papers	Average

(continued)

Table 2 (continued)

S.no. (1)	Year (2)	Objective(s) addressed (3)	Tools & Technology used (4)	Benefit(s)/Outcome(s) (5)	Real time response (6)
27.	2016	Automatic Rain operated Wiper and Dimmer for vehicles	Two different circuits (Dimmer and wiper) synchronized using LDR, Relay, Transistor and PIC16F877A. Wiper circuit uses capacitive, piezo electric and optical methods of technology	The implemented system reduces the glare faced by driver and reduced the accidents caused by 'TROXLER EFFECT'	Average
28.	2017	A multi featured automatic head light systems prototype for automotive safety	Arduino BASED LED connection to switch on or off based upon detection during night time [49]	Successfully created a device to switch directly to low or high beam light without varying intensity upon the detection of light using LDR	Poor
29.		Automatic headlight dimmer and Noise free horns in automobiles.	V2V technology and smart VANET	The system proposed improves the connectivity speed between the vehicles using Zigbee technology for detection of small targets, increasing precision and giving a better driving experience for the audience as the headlights dim upon detection of objects automatically from high to low and vice versa if the road is clear	Good
30.	2018	Automated headlight intensity control and obstacle alerting system	LDR plus Zigbee based project to deliver switch in beam of one vehicle based upon the other approaching one [50]	Reduction of driving blindness during the night time as the LDRs switches the beam to low upon detection of a vehicle in front of the author's vehicle	Good

(continued)

Table 2 (continued)

S.no (1)	Year (2)	Objective(s) addressed (3)	Tools & Technology used (4)	Benefit(s)/Outcome(s) (5)	Real time response (6)
31.	2019	Light sensing headlamps and taillights in automobiles	Light dependent resistor and voltage dimmer circuit [51]	Up to 147 m, this system detects light and uses switch case mechanism to switch the beam from high to low and vice versa	Good
32.	2020	Arduino based headlight control with multi trait	Arduino Uno in line with a gyroscope for measurement of deflection. LDR and potentiometer along with IR sensor for dimmer circuit [52]	Based on deflection of Arduino and the LDR signal, light is dimmed if a pedestrian, vehicle in same or opposite lane is detected	Good
33.	2021	Adaptive driving beam system using MEMS optical scanning	Microelectromechanical systems optical scanner using piezoelectric scanner. Lissajous pattern for configuration	Based on oncoming vehicle, pedestrian, road and cruising velocity, the pattern reconfigures and conditionally operates the optical system for beam assistance during driving	Good

- Majority of technologies mainly focus upon one factor majorly and that is the presence on an object or vehicle detection only and how the beam reacts upon detection, barely any medium has been constructed to detect any other factors around like climate and stimulate any change in beam upon detection except a few papers.
- The implemented systems mainly focus on light detection throughout like detection of presence of an automobile via absorption of the light emitted by tail lights or the headlights of a vehicle coming in the opposite direction or the presence of an object based on image segmentation analysis.
- The systems already fitted in for automatic or smart headlights end up draining the battery of the vehicle, hence shortening the battery life for usage.
- The sensors for the light detection that are basically or the rain sensors LDRs are always placed near the windshield and weather such as humidity or precipitation usually hamper switch the sensing or the climate affects the quality of the data collected.
- The models developed as of now for instilling an automatic headlight system are ninety percent in only automobiles like cars and barely any other vehicle has been used for testing or implementation, for example, motorbikes.

Table 3 Inference table

Weather conditions		Climate		Physical conditions		Action
Drizzle/Storm/Snow/Rain	Absent	Pedestrian	Absent	Vehicle	Absent	if BEAM = HIGH OR LOW \Rightarrow NO CHANGE
	Absent	Absent	Absent	Present	Present	if BEAM = HIGH \Rightarrow BEAM = LOW; if BEAM = LOW \Rightarrow NO CHANGE
	Absent	Absent	Present	Absent	Absent	if BEAM = HIGH \Rightarrow BEAM = LOW; if BEAM = LOW \Rightarrow NO CHANGE
	Absent	Absent	Present	Present	Present	if BEAM = HIGH \Rightarrow BEAM = LOW; if BEAM = LOW \Rightarrow NO CHANGE
	Absent	Absent	Present	Absent	Absent	if BEAM = HIGH \Rightarrow BEAM = LOW; if BEAM = LOW \Rightarrow NO CHANGE
	Absent	Present	Absent	Present	Present	if BEAM = HIGH \Rightarrow BEAM = LOW; if BEAM = LOW \Rightarrow NO CHANGE
	Absent	Present	Absent	Absent	Absent	if BEAM = HIGH \Rightarrow BEAM = LOW; if BEAM = LOW \Rightarrow NO CHANGE

(continued)

Table 3 (continued)

Weather conditions		Climate		Physical conditions		Action
Drizzle/Storm/Snow/Rain		Overcast/Coudly		Pedestrian		
	Absent	Present			Vehicle	
	Absent	Present			Absent	if BEAM = HIGH ⇒BEAM = LOW; if BEAM = LOW⇒NO CHANGE
	Absent	Present	Present		Present	if BEAM = HIGH ⇒BEAM = LOW; if BEAM = LOW⇒NO CHANGE
Present		Absent		Absent	Absent	if BEAM = HIGH ⇒BEAM = LOW; if BEAM = LOW⇒NO CHANGE
Present		Absent		Absent	Present	if BEAM = HIGH ⇒BEAM = LOW; if BEAM = LOW⇒NO CHANGE
Present		Absent		Absent	Present	if BEAM = HIGH ⇒BEAM = LOW; if BEAM = LOW⇒NO CHANGE

(continued)

Table 3 (continued)

Weather conditions		Climate		Physical conditions		Action
Drizzle/Storm/Snow/Rain		Overcast/Coudly		Pedestrian	Vehicle	
Present	Absent	Present	Absent	Present	Present	
Present	Present	Absent	Present	Absent	Absent	if BEAM = HIGH ⇒BEAM = LOW; if BEAM = LOW⇒NO CHANGE
Present	Present	Present	Absent	Present	Absent	if BEAM = HIGH ⇒BEAM = LOW; if BEAM = LOW⇒NO CHANGE
Present	Present	Present	Present	Absent	Absent	if BEAM = HIGH ⇒BEAM = LOW; if BEAM = LOW⇒NO CHANGE
Present	Present	Present	Present	Present	Absent	if BEAM = HIGH ⇒BEAM = LOW; if BEAM = LOW⇒NO CHANGE
Present	Present	Present	Present	Present	Present	if BEAM = HIGH ⇒BEAM = LOW; if BEAM = LOW⇒NO CHANGE

5 Conclusions and Future Scope

These days, the growing concern is the increase in the number of road accidents that cause various injuries, sometimes fatal. The main cause of this problem has been identified as critical to the automotive sector and it has been found that more than 40% of road accidents are due to improper use of beams or reckless use and blinking of a bag type removed from traffic light. A high beam is very dangerous and should only be used in situations where there will be no impact on oncoming vehicles or on the driver himself. Thus, an IoT based system to monitor and toggle high beam to low upon detection of surroundings is required.

This survey has been written accordingly, studying how does high beam impact daily driving routines, how can it be proved detrimental to health, and how by linking IoT with the automatic headlight dimmer systems, we can propose a solution to create a safe environment for citizens to drive in.

In future, the dimmer systems can be equipped and linked to GPS, to automatically perform as per location detected by the sensors and perform as per fed action response switch case mechanism to give rise to locational equipped IoT based dimmer systems.

References

1. Worly, H.: Road traffic accidents increase dramatically worldwide (2006, March 1)
2. Mace, D., et al.: Countermeasures for reducing the effects of headlight glare. American Automobile Association Foundation for Traffic Safety (2001)
3. Clarke, F.J.J.: A study of Troxler's effect. *Opt. Acta Int. J. Opt.* **7**(3), 219–236 (1960)
4. Longcore, Travis, Rich, Catherine: Ecological light pollution. *Front. Ecol. Environ.* **2**(4), 191–198 (2004)
5. Mack, Arien: Inattentional blindness: looking without seeing. *Curr. Dir. Psychol. Sci.* **12**(5), 180–184 (2003)
6. Bonneh, Y.S., et al.: Motion-induced blindness and Troxler fading: common and different mechanisms. *PLoS One* **9**(3), e92894 (2014)
7. McMannamy, I.D.: Adjustable headlight. U.S. Patent No. 888,641 (26 May 1908)
8. Rhoades, L.T.: Means for automatically regulating automobile-headlights. U.S. Patent No. 1,241,284. (25 Sept 1917)
9. Olson, V.: Automatic headlight-dimmer for automobiles. U.S. Patent No. 1,256,362 (12 Feb 1918)
10. Wiseman, C.M.: Automatic headlight-dimming attachment for automobiles. U.S. Patent No. 1,536,366 (5 May 1925)
11. Jacob, R.: Automatic headlight dimmer. U.S. Patent No. 2,632,040 (17 Mar 1953)
12. Stam, J.S., Bechtel, J.H., Roberts, J.K.: Control system to automatically dim vehicle head lamps. U.S. Patent No. 5,837,994 (17 Nov 1998)
13. Asari, T., et al.: Adaptive driving beam system with MEMS optical scanner for reconfigurable vehicle headlight. *J. Opt. Microsyst.* **1**(1), 014501 (2021)
14. Könning1, T., Amsell, C., Hoffmann, I.: Light has to go where it is needed: future light based driver assistance systems (2007)
15. Alcantarilla, O.F., Bergasa, L.M., Jiménez, P., Sotelo, M.A., Parra, ID., Mayoral, S.S.: Night time vehicle detection for driving assistance. In: IEEE Intelligent Vehicles Symposium (IV) (2008)

16. Chen, Y.L.: Nighttime vehicle light detection on a moving vehicle using image segmentation and analysis techniques. *WSEAS Trans. Comput.* **8**(3), 506–515 (2009)
17. Rubio, J.C., Serrat, J., Lopez, A.M., Ponsa, D.: Multiple target tracking for intelligent headlights control. In: 13th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 903–910 (2010)
18. Li, Y., Haas, N., Pankanti, S.: Intelligent headlight control using learning-based approaches. *IEEE Intell. Veh. Symp. (IV) Baden-Baden* **2011**, 722–727 (2011). <https://doi.org/10.1109/IVS.2011.5940541>
19. Davis, D., George, J., MR, K.: Automatic Head Light Dimmer and Noise free Horns in Automobiles (2017)
20. Arpita K, Akhila, M.J., Avi Kumar, R.: Automated headlight intensity control and obstacle alerting system. *Int. J. Eng. Res. Technol. (IJERT) NCESC–2018* **6**(13) (2018)
21. Manjula, S.C., Sushmitha, M.R., Parameshachari B.D.: Automated Headlight Intensity Control and Obstacle Alerting System (2017)
22. Pal, S., Bhaskaran, S.: Arduino based conventional headlight with multi trait. In: 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE (2020)
23. Rosenblum, I.: Automatic headlight system. U.S. Patent No. 4,236,099 (25 Nov 1980)
24. Akinsanmi, O., Ganjang, A.D., Ezea, H.U.: Design and development of an automatic automobile headlight switching system. *Int. J. Eng. Appl. Sci.* **2**(8) (2015)
25. Chapakanade, P., Gangurde, P., Peje, S., Shende, D.R.: Automatic rain operated wiper and dimmer for vehicle. *Int. Res. J. Eng. Technol. (IRJET)* **3**(04), 2376–2378 (2016)
26. Gillespie, L.W.: Automatic headlight control. U.S. Patent No. 2,240,843 (6 May 1941)
27. Parkes, W.B.: Motor vehicle headlight activation apparatus for inclement weather conditions. U.S. Patent 5,374,85 (1994)
28. Elvik, Rune: A meta-analysis of studies concerning the safety effects of daytime running lights on cars. *Accid. Anal. Prev.* **28**(6), 685–694 (1996)
29. Asaduzzaman, A., Islam, Mohammad., Paul, Shuva., Alam, Farhat, Rahman, Md: Automatic High Beam Controller for Vehicles. *Int. J. Sci. Eng. Res.* **4**, 5 (2013)
30. Wolfe, G.B.: Head lamp switch with twilight sentinel control. U.S. Patent No. 6,288,492 (11 Sept 2001)
31. Cawley, M.J.: Automobile-headlight. U.S. Patent No. 1,168,058 (11 Jan 1916)
32. Bordewieck, R.W., Roebuck, J.O.: Automatic headlight control device. U.S. Patent No. 2,686,277. (10 Aug 1954)
33. Gandelot, H.K.: Automatic headlight control system. U.S. Patent No. 2,749,478 (5 Jun 1956)
34. Vanaman, F.P., Odom, C.H., Gordon, H.L.: Automatic headlight dimming system. U.S. Patent No. 2,959,709 (8 Nov 1960)
35. Goode III, J.W.: Automotive accessory control system. U.S. Patent No. 4,435,648 (6 Mar 1984)
36. Lawler, L.N.: Automatic headlight dimmer apparatus. U.S. Patent No. 5,086,253 (4 Feb 1992)
37. Tofflemire, Troy C., Whitehead, Paul C.: An evaluation of the impact of daytime running lights on traffic safety in Canada. *J. Saf. Res.* **28**(4), 257–272 (1997)
38. Beam, N.E.: Adaptive/anti-blinding headlights. U.S. Patent No. 6,144,158 (7 Nov 2000)
39. Schofield, K., Larson, M.L., Vadas, K.J.: Vehicle headlight control using imaging sensor. U.S. Patent No. 6,097,023 (1 Aug 2000)
40. Sullivan, J., Adachi, G., Mefford, M., Flannagan, M.: High-beam headlamp usage on unlighted rural roadways. *Light. Res. Technol.* **36**(1), 59–65 (2004). <https://doi.org/10.1191/1477153504li104oa>
41. Elvik, R., Christensen, S.F., Olsen, S.F.: Daytime running lights. A systematic review of effects on road safety. No. TOI-688/2003 (2003)
42. Schofield, K., Larson, M.L., Vadas, K.J.: Vehicle headlight control using imaging sensor identifying objects by geometric configuration. U.S. Patent No. 6,559,435 (6 May 2003)
43. Simons, D., et al.: Induced visual fading of complex images. *J. Vis.* **6**(10), 9–9 (2006)
44. Hsieh, P.-J., Tse, P.U.: Illusory color mixing upon perceptual fading and filling-in does not result in ‘forbidden colors’. *Vis. Res.* **46**(14), 2251–2258 (2006)

45. López, A., Hilgenstock, J., Busse, A., Baldrich, R., Lumbrieras, F., Serrat, J.: Temporal coherence analysis for intelligent headlight control. In: 2nd Workshop on Planning, Perception and Navigation for Intelligent Vehicles, Nice, France (2008)
46. Bazzan, A.L.C., Azzi, G.G.: An investigation on the use of navigation devices in smart transportation systems. *Anais do VIII Simpósio Brasileiro de Sistemas de Informação* (2012)
47. Alsumady, M., Alboon, S.: Intelligent Automatic High Beam Light Controller. ©2013 Old City Publishing, Inc. Published by license under the OCP Science imprint, a member of the Old City Publishing Group. 00, pp. 1–8 (2013)
48. Nutt, V., Kher, S., Raval, M.: Fuzzy headlight intensity controller using wireless sensor network. In: 2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Hyderabad, pp. 1–6 (2013). <https://doi.org/10.1109/fuzz-ieee.2013.6622538>
49. Jadhav, M.S.S., Mulla, A.A.: A multi featured automatic head light systems prototype for automotive safety
50. Hussain, Rasheed, Zeadally, Sheralli: Autonomous cars: research results, issues, and future challenges. *IEEE Commun. Surv. Tutor.* **21**(2), 1275–1313 (2018)
51. Swaroop, A., et al.: Design of light-sensing automatic headlamps and taillamps for automobiles. In: Proceedings of the Third International Conference on Microelectronics, Computing and Communication Systems. Springer, Singapore (2019)
52. Gayatri, S.S., et al.: Design and implementation of automatic vehicle headlight dimmer. *IJRAR Int. J. Res. Anal. Rev. (IJRAR)* **7**(1), 267–271 (2020)
53. Shreyas, S., et al.: Adaptive headlight system for accident prevention. In: 2014 International Conference on Recent Trends in Information Technology. IEEE (2014)

Enhanced Energy-Efficient Fuzzy Logic Clustering and Network Coding Strategy for Wireless Sensor Networks (EEE-FL-NC)



K. S. Fathima Shemim and Dr. Ulf Witkowski

Abstract Battery energy is limited in Wireless Sensor Networks (WSNs); thus energy is a key element in designing WSNs. In particular, the effective utilization of energy becomes the major challenge during the design of routing protocols for WSNs, and the ultimate aim of the routing protocols is to extend the network lifespan of Wireless Sensor Networks by efficiently utilizing node energy. Clustering is one kind of mechanism in Wireless Sensor Networks to prolong the network lifetime and to reduce overall energy consumption. This paper used fuzzy logic for electing cluster heads based on 7 different descriptors—(delay, distance from the base station, RSSI, density, residual energy, location suitability, and Compacting) in each round. Dedicated network coder nodes in the bottleneck zone used network coding algorithms to improve data transmission rate. This sequentially improves the overall network lifetime. NS2 simulation tool and C++ programming language have been used to simulate and implement the proposed routing protocol. The simulation results proved that EEE-FL-NC protocol outperforms based on throughput, energy consumption, and the average lifetime of the network while comparing with the state-of-the art protocols like FL-NC-EEC, LEACH, K-means-LEACH, FL-EEC, and F-LEACH-[1]/D.

Keywords Clustering · Fuzzy logic · Network coding · Routing

K. S. Fathima Shemim (✉)

Computing and Software Engineering Department, RAK Academic Centre, University of Bolton, Ras al Khaimah, United Arab Emirates

e-mail: F.KS@bolton.ac.uk

Dr. U. Witkowski

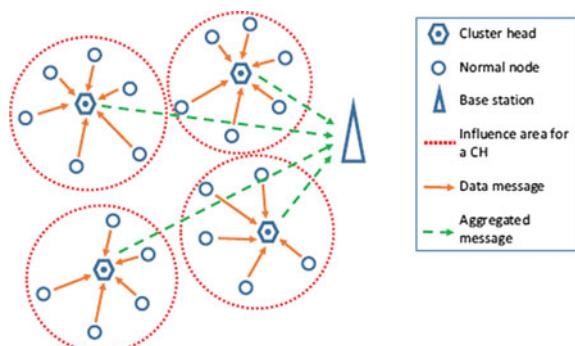
Electronics and Circuit Technology Research Group, South Westphalia University of Applied Sciences, Iserlohn, Germany

e-mail: witkowski@fh-swf.de

1 Introduction

Wireless Sensor Networks (WSNs) gained universal attention in recent times because of the improvement of Micro-Electro-Mechanical Systems (MEMS) technology. WSNs have received remarkable attention due to their extensive usage in disaster management, environmental monitoring, military applications, smart homes, agricultural applications, security, transportation, etc. [1–3]. WSN consists of lots of nodes with limited resources in terms of battery power, processing, and storage capability. These sensor nodes that are arbitrarily positioned in the field are usually used to measure and collect surrounding environmental conditions like light, sound, pressure, temperature, moisture, etc. The collected information will be passed to the sink or destination node with several hopes based on the routing algorithm selected for the network. As the nodes in WSNs are battery powered with restricted energy, routing protocols have an important role in reducing energy consumption during data transmission. Different researchers proposed numerous routing protocols for WSNs to enhance the energy efficiency of the network. Many researchers used the concept of clustering of sensor networks and efficient routing to preserve the energy of nodes, and these techniques will significantly improve the lifetime of WSN [3, 4]. Clustering-based protocols use dedicated nodes to transmit data to the sink or destination node. Here, the entire network will be dividing into different clusters and the cluster head will be acting as the coordinator for the next level of communication. As in LEACH [3] protocol, EEE-FL-NC protocol also comprises two stages, i.e., setup stage and steady stage [5]. Figure 1 shows an illustration of clustering operation in wireless sensor networks. In the setup stage, the clusters of the network were created and Cluster Head (CH) is elected among sensor nodes present in the cluster. The proposed algorithm used fuzzy logic with 7 different parameters (delay, distance to Base Station (BS), RSSI, density, residual energy, location suitability, and Compacting) to select the most efficient cluster head for each cluster in every round. Also used network coding mechanism before sending data packets to the destination. These two techniques together can significantly reduce the energy usage and expand the lifespan of the WSNs [3, 5, 6].

Fig. 1 Clustering in WSNs



2 Related Works

In WSNs, nodes are arbitrarily deployed in most of the applications. After deployment, all these sensor nodes collectively perform the specific job by communicating with each other based on the routing algorithm used for the network. All these sensor nodes are battery powered and it is challenging to replace or recharge nodes battery once deployed, so the major design challenge of WSN is by what means the network can efficiently utilize the energy of nodes as well as increase network lifespan. Based on different constraints like throughput optimization, data aggregation, data redundancy, network coverage, compaction, and data communication techniques, different routing algorithms are proposed to improve energy efficiency and network lifetime [4, 5, 7, 8]. From those hierarchical energy-efficient routing algorithms are more popular by efficiently utilizing nodes by dividing the entire network into different clusters and assigning different roles to nodes. By using data gathering and fusion, hierarchical routing protocols diminish the number of transactions to the destination or BS. LEACH is one of the popular self-organizing, adaptive clustering protocols. LEACH protocol has two stages, i.e., setup stage and steady stage. LEACH randomly picks CH for each cluster in each round and tries to distribute energy consumption between nodes in the networks. Weighted LEACH (W-LEACH) is a modified LEACH, which can handle non-uniform networks by modifying the intra-cluster data communication phase. In Centralized LEACH (LEACH-C), depending on the remaining energy and the location of each node, Base Station (BS) selects the best CH in each round. By doing this, LEACH-C guarantees the optimal number of CHs [6, 8]. Energy-Efficient Weight-Clustering (EWC) considered different metrics to choose CHs in each round. EWC used node degree, node residual energy, distance, etc. to elect CH in the assumption that nodes are homogeneous, static, and distributed randomly. KMMDA (K-Means Minimum Mean Distance Algorithm) improves network life by applying the K-Means algorithm to compute nodes' mean distance while selecting cluster head. Energy-Efficient CH Selection (NECHS) protocol used fuzzy logic approach to select the CH in each round based on two descriptors called remaining energy and number of neighboring nodes, and this improves the energy efficiency of the entire network by considering the best node for cluster head. While LEACH-FL (Fuzzy Logic-LEACH) used node density, the distance between BS and the CH, energy level, etc. as the descriptors to elect cluster head, which is more accurate when comparing with NECHS [8–10]. CHEF [6, 9] used distributed Cluster Head election method to enhance network expansibility. CHEF used two descriptors (energy and local distance) for fuzzy logic and selected a pool of candidate nodes for CH based on a threshold value. As discussed, many clustering protocols have been emerged to solve energy efficiency problems; many of them are based on the LEACH protocol. However, more research work is still needed to find a more energy-efficient routing protocol that is scalable and robust and to extend network lifetime by enhancing energy consumption [6, 8–11].

3 Proposed System Model for EEE-FL-NC

Enhanced Energy-Efficient Fuzzy Logic Clustering and Network Coding (EEE-FL-NC) strategy is proposed in this paper. As in LEACH, EEE-FL-NC, the proposed protocol, also has a setup stage and steady-state stage except in the setup stage it uses fuzzy logic with 7 different descriptors (delay, distance from the base station, RSSI, density, residual energy, location suitability, and Compacting) to select CH in every round. EEE-FL-NC protocol used Network encoding scheme in the bottleneck zone to encode received packets before transmitting to the base station. Here, more capacity of data can be forwarded to the destination node with an equivalent number of deliveries. EEE-FL-NC sequentially increases the overall lifespan of the network by joining fuzzy logic algorithms together with network coding techniques. EEE-FL-NC protocol makes few assumptions during performance evaluation. Assume that all nodes in the network have the same initial energy and are distributed randomly across the specified area where recharging of the batteries is impossible. All nodes having uniform communication and processing capabilities. Base Station will be able to estimate the sensor node location by adopting the weighted centroid localization technique [4, 6, 7, 10, 12].

3.1 Radio Energy Dissipation Model

In this proposed system, the first-order energy model is applied. Equation (1) shows the energy dissipated for the transmission of l-bit over the distance d. To receive the l-bit message at the receiver side, radio disburses energy as in Eq. (2), where E_{elec} the electronics energy is the energy dissipation that consumes per bit for the receiver or transmitter circuitry, ϵ_{fs} and ϵ_{mp} are the amplifier models, the distance between the receiver and the transmitter is denoted as d, and d_0 is the reference distance value [8, 11, 13, 14].

$$E_{Tx}(l, d) = \begin{cases} l * E_{elec} + l * \epsilon_{fs} * d^2 & \text{if } d < d_0 \\ l * E_{elec} + l * \epsilon_{mp} * d^4 & \text{if } d \geq d_0 \end{cases} \quad (1)$$

$$E_{Rx}(l) = l * E_{elec} \quad (2)$$

$$d_0 = \sqrt{\frac{\epsilon_{fs}}{\epsilon_{mp}}} \quad (3)$$

3.2 Fuzzy Logic System for EEE-FL-NC Protocol

In the cluster formation stage of the EEE-FL-NC protocol, to elect the CH, a Fuzzy Logic System has been implemented. We used the Mamdani model. Fuzzy Logic System as in Fig. 2 consists of 4 stages, a fuzzifier, the fuzzy inference engine, a rule-base for Fuzzy Inference Engine (FIE), and a defuzzifier. Fuzzifier creates fuzzy sets based on the 7 input variables given with crisp values. During Rule evaluation, apply IF-THEN logic rules to fuzzy descriptors to find a new fuzzy set. IF-THEN rules have several input conditions and logical operators. The fuzzy inference mechanism mainly does all of the calculations for the output, which combines input values and IF-THEN rules. The fuzzy set $\mu(x_i)$, represented as in Eq. (4), for N nodes is $N = \{x_1, x_2, x_3, \dots, x_n\}$. The fuzzy Logic model deals with the 7 different elements, each element converted into a fuzzy linguistic variable by using the selected membership functions. The Defuzzifier transforms the fuzzy output into a real output value or the crisp value which intern decides the chance of a node to become a CH [8, 15, 16].

$$\mu(x_i) = \{\mu(x_1), \mu(x_2), \dots, \mu(x_n)\}, \quad (0 \leq i \leq 1) \quad (4)$$

EEE-FL-NC protocol used 7 different descriptors as input variables to the Fuzzy Logic System. 1. Energy: If we choose a node with advanced remaining energy as CH, it improves the lifetime of the network. 2. Surrounding node density: Selection of a node with more neighbors in the vicinity to become a CH improves energy consumption. 3. Compaction: Total energy consumption will be minimized if we select a node with a higher degree of compaction. 4. BS Distance: Consumed energy will be very less if the gap between the CH and BS is less. 5. End–End Delay: Average time duration of a packet to reach its destination is considered as an end-to-end delay. 6. Received Signal Strength: By selecting CH with a high RSSI value, Cluster Members will be able to send packets to CH with the lowest transmission power. 7. Average energy of the network: The best position for a CH is always a position with lower total interaction energy. With these 7 descriptors, FLS decides

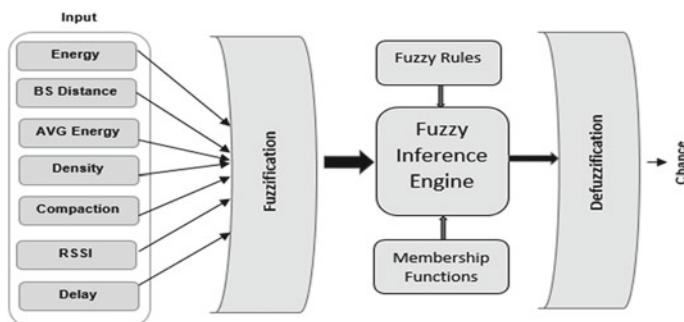


Fig. 2 Fuzzy logic system

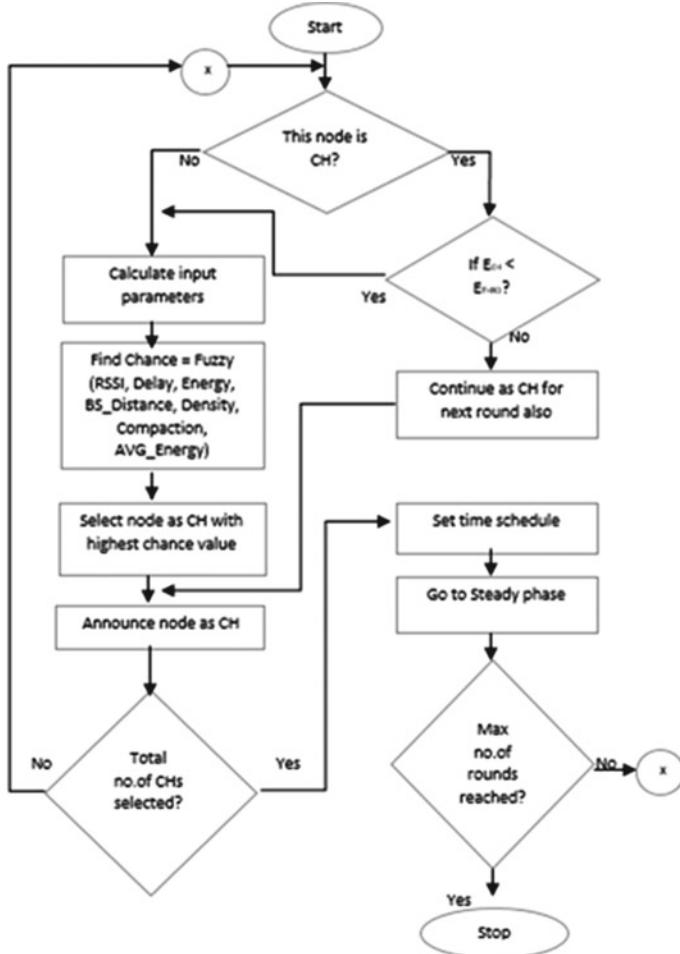


Fig. 3 Cluster head selection

the best node as the CH in each round [5, 8, 10, 17]. The detailed algorithm for selecting Cluster Head in each round for each cluster is depicted in Fig. 3.

3.3 Network Coding Technique

In the proposed work, we used network coding technique in the bottleneck zone, which in turn increases network energy efficiency by increasing the volume of data transmission with the same number of transmissions. In the bottleneck zone, sensor nodes will be having different roles, i.e., relay node or the encoder node.

The relay node directly transfers the data to its destination whereas the encoder node compresses the data with XOR encoding technique with the previously received data in the encoder queue [6, 12]. At the receiver side, this data packet will be decoded to retrieve actual data. Figure 4 explains the flow of the network coder node. NC algorithm helps to improve WSNs' overall lifespan by the healthier utilization of bandwidth and by reducing the number of failure nodes in the bottleneck region [6, 12]. Network coding is a method that enables the inner node to encode data packets collected from their adjacent nodes in a network. In this work, XOR encoding for the linear network coding technique has been used. The encoded packets that are transmitted in the network are elements in $GF(2) = \{0, 1\}$, and bitwise XOR in $GF(2)$ is adopted as an operation. At the time of encoding, the node which is acting as encoder node chooses an encoding vector with a sequence of coefficients $q = (q_1, q_2, q_3 \dots q_n)$ from $GF(2^s)$. G_i ($i = 1, 2, 3, 4, 5, \dots, n$) is a set of n packets received at the encoder node which will be linearly encoded into a single packet. This output

Fig. 4 Network coding

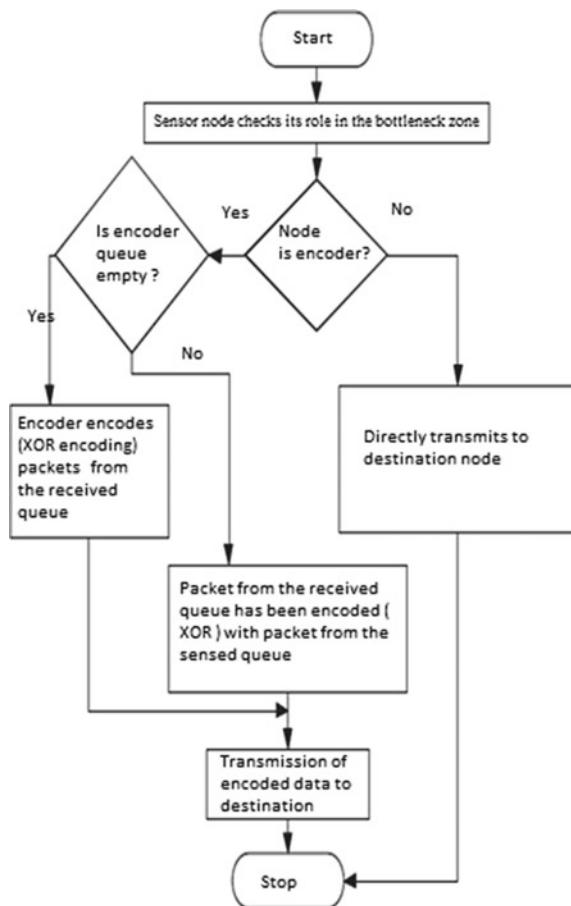


Table 1 NS2 simulation parameters

Parameter	Value
Version	NS2.34
Initial energy of a sensor node	2 J
Total number of nodes	50, 100, 150, and 200
Queue size	50
Routing protocol	EEE-FL-NC
Traffic type	CBR, UDP
Size of packet	512 bytes
Deployed region	X = 150, y = 300

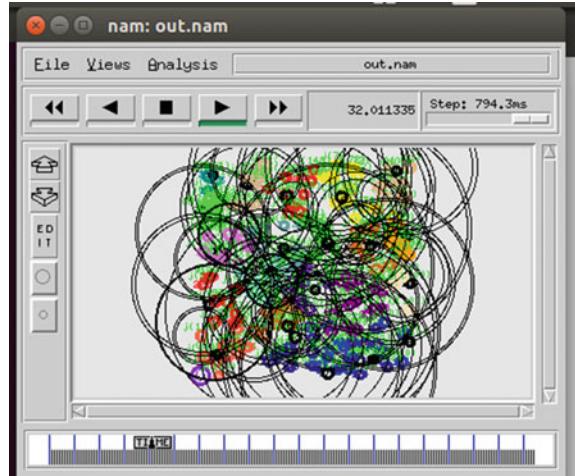
packet is transmitted with the n coefficient [6, 7, 17]. This encoded data is called an information vector. The same encoding vector will be used in the destination node to decode the received packet. To retrieve the original packet from the received encoded packet, the receiver node solves a set of linear equations. The receiver node will be using the encoding vector q which is already received with the encoded data packet. A set $(q_1, Y_1), (q_2, Y_2), \dots, (q_m, Y_m)$ that has been collected at the recipient helps to decode data, where Y_i represents the information and q_i represents the coding vector [6, 12, 17].

4 Simulation Setup

EE-FL-NC protocol is implemented in C++ and simulated using a network simulator called NS2.34. The performance of EE-FL-NC is compared in NS2 and plotted the results using XGraph. We used different parameters for the testing scenario, also varying the total number of sensor nodes. The main parameters used for the scenario are displayed in Table 1. Assume the nodes are randomly deployed in the field with dimensions 150×300 , with the initial energy of the sensor node as 2 J. Packet size is set as 512 bytes and Queue set as 50. Network Animator (NAM) is used to animate the selected scenario; NAM is a Tcl/TK-based animation tool for observing network simulation and real-world packet delivery traces. Figure 5 shows NAM output for 200 nodes. Color coding has been given for sensor nodes in the network to differentiate between live nodes, dead nodes, cluster heads, etc. in each round.

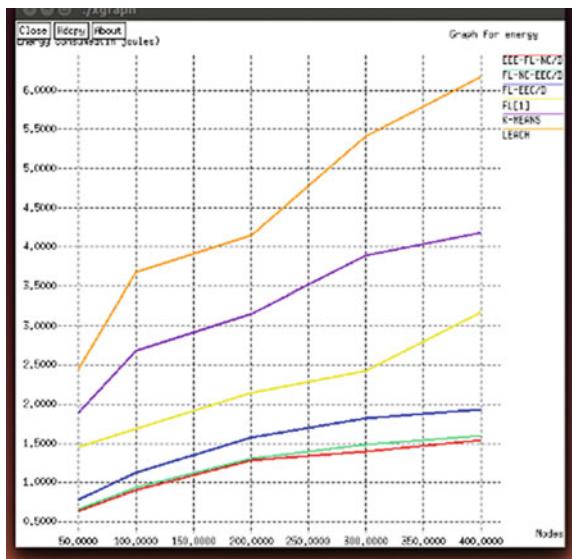
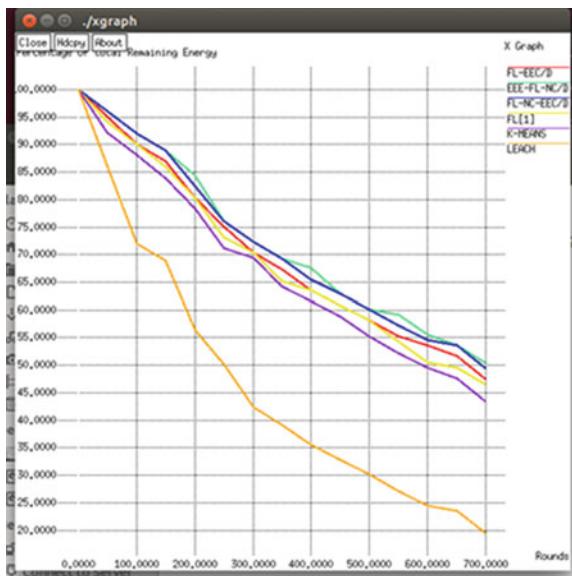
5 Results and Discussion

Performance analysis of the proposed protocol has been done over 200 nodes. AWK scripting language has been used to filter and process data from the generated trace log file. By using extracted data, graphs have been plotted with XGraph as given in

Fig. 5 NAM output

Figs. 5, 6, 7, 8, 9, 10, and 11. Figure 5 shows that the average end-to-end throughput is higher when compared the proposed EEE-FL-NC protocol with LEACH, FL-LEACH, K-MEANS, and FC-NC-EEC/D protocols, which is measured by the average number of packets collected at the BS per second. The throughput of the EEE-FL-NC protocol improves by selecting the best Cluster Head with the highest RSSI value, with low transmission power, and with the highest remaining energy for each round. Figure 6 shows the total energy consumed in each round [4, 5, 7, 10, 15]. When

Fig. 6 Average throughput

Fig. 7 Energy consumed**Fig. 8** Percentage of total remaining energy

comparing with LEACH, the EEE-FL-NC protocol uses very minimal energy for the transaction and thereby improves overall network lifetime. From Figs. 8 and 9, it's proved that the number of alive nodes is more and the number of dead nodes and the half node dead is less in the proposed algorithm by efficiently utilizing node energy while selecting CMs and its CH using fuzzy logic. By using a network coding

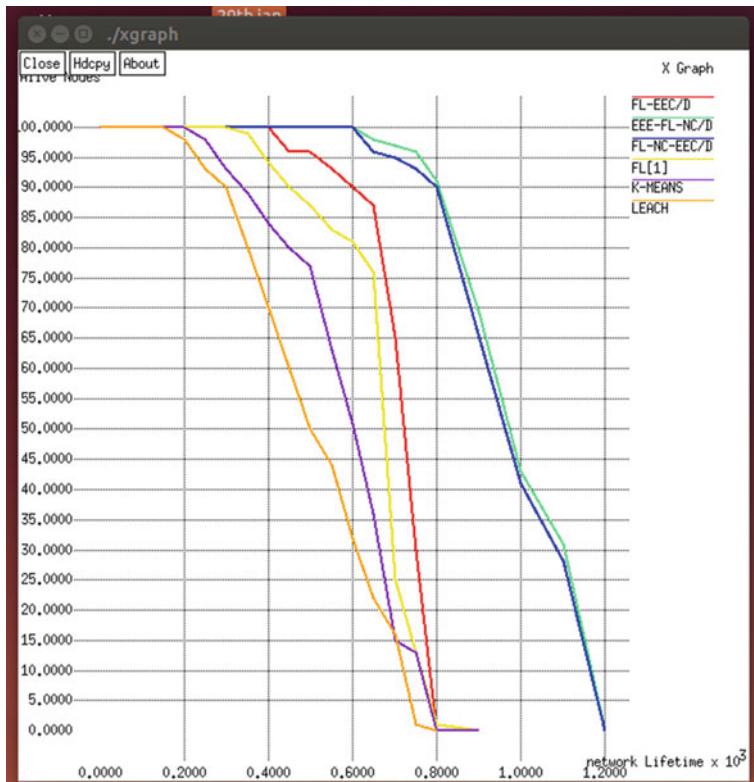


Fig. 9 Alive nodes

algorithm, data traffic has been reduced and utilized the bandwidth effectively, so the proposed protocol outperforms in terms of network throughput. Figures 10 and 11 show the improvements in the stability period, which indicates the time where the first node dead in the network for each protocol. From the simulation results, it's proved that the proposed EEE-FL-NC protocol significantly diminishes the energy dissipation of the entire network thereby increasing the lifespan of the network.

6 Conclusion

In the proposed work, we used a fuzzy system with 7 descriptors to select the best Cluster Head for its cluster members to accomplish optimal clustering. Also Used XOR encoding decoding mechanism to utilize the bandwidth effectively to transmit more packs with a limited time duration. The proposed algorithm is an enhanced version of the FL-NC-EEC/D protocol by adding RSSI and Delay descriptors as input to the Fuzzy System and combining network coding algorithm. In WSNs, the

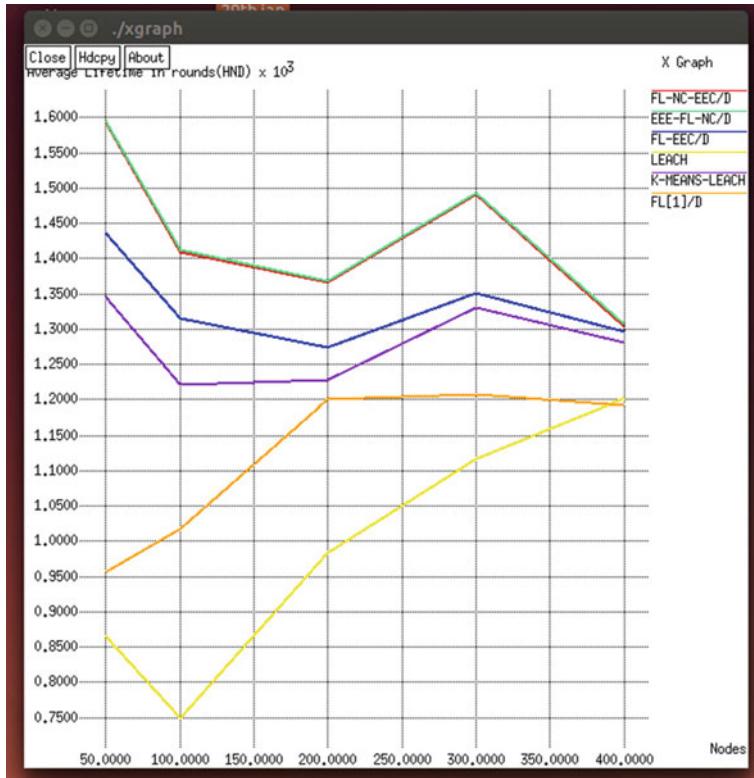


Fig. 10 Half Node dead

rate of the depletion energy is directly related to the nodes failure pattern statistics of the sensor nodes. So, to avoid energy holes in the bottleneck zone, we used a network coding algorithm with the best cluster head selection mechanism for the clusters. The simulation results authenticate that the proposed protocol outperforms the compared state-of-the-art protocols in terms of energy consumption, throughput, and the network lifetime.

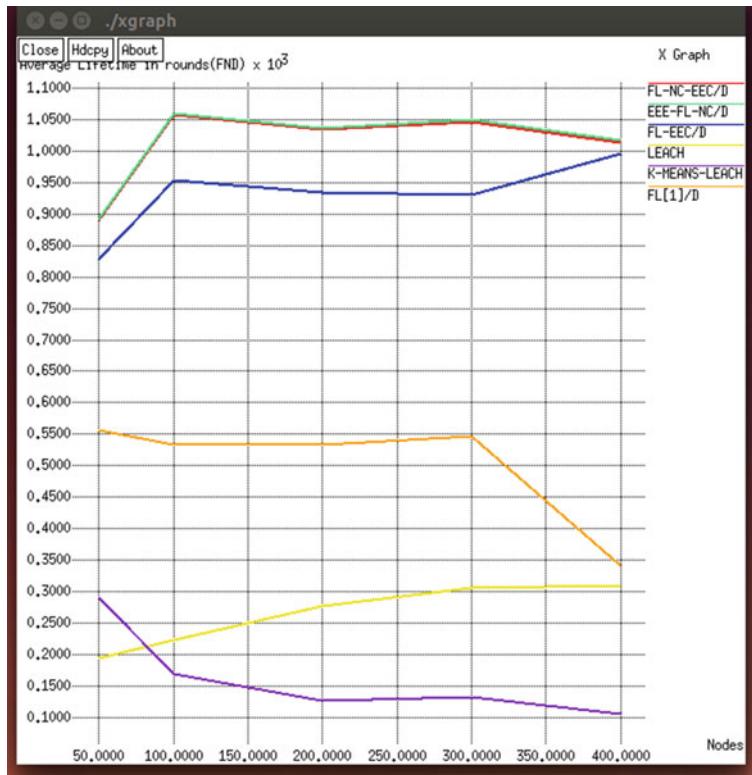


Fig. 11 First Node dead

References

1. Jasim, A.A., Idris, M.Y.I., Razalli Bin Azzuhri, S., Issa, N.R., Rahman, M.T.: Energy-efficient wireless sensor network with an unequal clustering protocol based on a balanced energy method (EEUCB). *Sensors* **21**(3), 784 (2021)
2. Abuhmida, M., Radhakrishnan, K., Wells, I.: Performance evaluation of mobile Ad Hoc routing protocols on wireless sensor networks for environmental monitoring. In: 2015 17th UKSim-AMSS International Conference on Modelling and Simulation (UKSim), pp. 544–548. IEEE (2015, March)
3. Shemim, K.F., Witkowski, U.: Energy efficient clustering protocols in WSNs: performance analysis and comparison of EEAHP protocol with LEACH and EAMMH using MATLAB. In: 2020 Advances in Science and Engineering Technology International Conferences (ASET), pp. 1–5. IEEE (2020)
4. Amri, S., Khelifi, F., Bradai, A., Rachedi, A., Kaddachi, M.L., Atri, M.: A new fuzzy logic based node localization mechanism for wireless sensor networks. *Futur. Gener. Comput. Syst.* **93**, 799–813 (2019)
5. Shemim, F., Shajahan, S.: Enhanced energy aware multi-hop hierarchical routing algorithm for wireless sensor networks. In: 2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA), pp. 1–4. IEEE (2017, November)

6. Fathima Shemim. K.S, Dr. Witkowski, U.: Improving lifespan of wireless sensor networks using fuzzy logic clustering and network coding (FL-NC) techniques. In: Int. J. Adv. Sci. Technol. **29**(10s), 8172–8182 (2020). <http://sersc.org/journals/index.php/IJAST/article/view/24267>. Accessed 10 Jan 2021
7. Sun, Q.T., Yin, X., Li, Z., Long, K.: Multicast network coding and field sizes. IEEE Trans. Inf. Theory **61**(11), 6182–6191 (2015)
8. Zhang, Y., Wang, J., Han, D., Wu, H., Zhou, R.: Fuzzy-logic based distributed energy-efficient clustering algorithm for wireless sensor networks. Sensors **17**(7), 1554 (2017)
9. Park, G.Y., Kim, H., Jeong, H.W., Youn, H.Y.: A novel cluster head selection method based on K-means algorithm for energy efficient wireless sensor network. In: 2013 27th International Conference on Advanced Information Networking and Applications Workshops, pp. 910–915. IEEE (2013, March)
10. Mohamed, R.E., Saleh, A.I., Abdelrazzak, M., Samra, A.S.: Survey on wireless sensor network applications and energy efficient routing protocols. Wirel. Pers. Commun. **101**(2), 1019–1055 (2018)
11. El-Sayed, H.H.: Performance Evaluation of Clustering EAMMH, LEACH SEP, TEEN Protocols in WSN (2018)
12. Kafaie, S., Chen, Y.P., Dobre, O.A., Ahmed, M.H.: Network coding implementation details: A guidance document (2018). [arXiv:1801.02120](https://arxiv.org/abs/1801.02120)
13. Kandris, D., Tsioumas, P., Tzes, A., Pantazis, N., Vergados, D.D.: Hierarchical energy efficient routing in wireless sensor networks. In: 2008 16th Mediterranean Conference on Control and Automation, pp. 1856–1861. IEEE (2018, June)
14. Radhika, S., Rangarajan, P.: On improving the lifespan of wireless sensor networks with fuzzy based clustering and machine learning based data reduction. Appl. Soft. Comput. **83**, (2019)
15. Tamandani, Y.K., Bokhari, M.U.: SEPFL routing protocol based on fuzzy logic control to extend the lifetime and throughput of the wireless sensor network. Wirel. Netw. **22**(2), 647–653 (2016)
16. Elshrkawey, M., Elsherif, S.M., Wahed, M.E.: An enhancement approach for reducing the energy consumption in wireless sensor networks. J. King Saud Univ. Comput. Inf. Sci., **30**(2), 259–267 (2018)
17. Liu, X.: Atypical hierarchical routing protocols for wireless sensor networks: a review. IEEE Sens. J. **15**(10), 5372–5383 (2015)

Malware Classification Using Automated Transmutation and CNN



Ritu Agarwal, Saurabh Patel, Sparsh Katiyar, and Sharad Nailwal

Abstract The ostentatiously steep increase in the number of malwares a year has given rise to a growing demand for a decision support system which would help against the same. The reason for such a steep increase in malware is because of slight mutation added to them to avoid detection. Thus, the malware from the same family with slight mutation is proliferated across the internet causing harm. We have come up with a Convolutional Neural Network (CNN) (Ciresan DC, Meier U, Masci J, Maria Gambardella L, Schmidhuber J, IJCAI Proceedings-international joint conference on artificial intelligence, vol 22, p 1237 (2011)) approach to battle a quarter of a million malwares a day, which heavily relies on the classification of the malwares based on the visual resemblance of malware samples of the same family. Binaries of malwares are represented visually as it has the ability to retain changes while keeping the global structure intact to help in detecting variations. The feasibility and performance of the proposed approach have been tested on the Malimg dataset which has shown avant-garde performance results. 98.97% precision was obtained by the proposed process.

Keywords Malware classification · CNN · Image processing · Neural networks

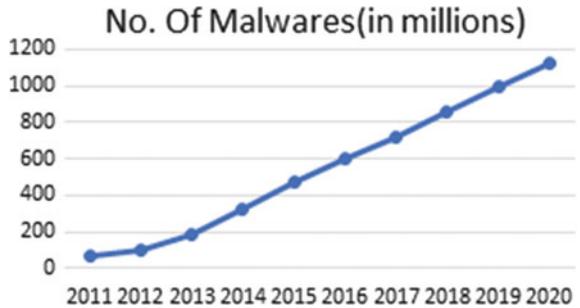
1 Introduction

Malware refers to that set of software programs that are designed to damage or perform any type of harmful actions to the computer system. The main purpose of this kind of malicious software is to steal important sensitive information and for spying on the system. Thereby posing threat to the integrity, confidentiality, and functioning of the digital system, which are the building blocks for a successful advancement toward digitalization. To make matters worse, there has been a steep increase in the number of malwares each year. According to the report of the Macfee Lab 2020 [2], more than 1.2 Billion total malwares have been detected in the second

R. Agarwal · S. Patel (✉) · S. Katiyar · S. Nailwal

Department of Information Technology, Delhi Technological University, Delhi 110042, India

Fig. 1 Last 10 years malware statistics showing a significant increase in malware with each year [28]



quarter of 2020 which is relatively much larger than the previous quarter's empirical data. There also has been an increase in the number of new malwares found each day and this increase is due to obfuscation and slight alteration to malwares to prevent detection. Each new alteration to any signature/fingerprint of malware gives rise to a new malware that can't be detected (Fig. 1).

One of the traditional approaches toward studying/analyzing malware is to extract the signature of the malware which contains their fingerprints. A signature is an algorithm that identifies a specific malware uniquely.

This fingerprint can be matched with other malware signatures. This is how the detection was done. However, due to the sudden increase in the number of new signatures of malwares (According to Macfee Lab report 2020, 50 million new in the second quarter of 2020), this traditional method turns out to be trivial, non-convenient, and doesn't show satisfactory results.

Previous researches on malware classification conclusively suggest that malware samples are deployed as alternatives to previously known samples and hence share common behaviors. This in turn can act as a prospect for developing a decision system for detecting and classifying malwares easily by relating to the malware's family irrespective of the variance.

In this research, we take up the above-mentioned idea and turn it into an applicable method using the convolutional neural network. The CNN has had seen a tremendous amount of success in various fields of artificial intelligence [3] (including Computer vision, speech recognition, and Natural Language Processing) [4] and especially has been successful on images. With the recent success of CNN in various fields (mentioned above) and images, we believe it is perfect for us to use the technique for the classification of malwares as well. So, for applying the CNN, we needed to convert malware into images, which we achieved by representing the binaries of each malware into a grayscale image and then trained CNN for classification.

2 Related Work

As the threat of malicious software has increased in the past decade, there has been a sudden increase in the usage of machine learning [5] algorithm for detection and the classification of malicious software. The increase in these approaches' success is due to a reduction in computing power cost and significant research in machine learning. Limited research has been done in visualizing the malware even after using several tools and techniques. Yoo [16] has used the self-organizing map to detect and classify malicious software. In [17], Leibrock et al. have developed a framework for visualizing the malware by reverse engineering. Functional areas are identified and then de-mystified by a visualization of the node-link whereby nodes represent the address and links represent transitions of state between domains. In [18], Goodall et al. have developed an environment for visual analysis of the malware. In [19], Trinius et al. have used Treemaps [6] to display the distribution of the operations and used thread graph for the sequence of operations. Using Static and Dynamic Analysis, several methods have been applied based on the feature extraction. Gandotra et al. [20] and Ranvee and Hiray [21] have given an overview that extracting the features of the malware is better rather than directly dealing with the raw malware as an abstracted view of the software program is being provided by it and then use these features to train a machine learning model.

The feature can be bifurcated into two different types:

- Static features
- Dynamic features.

Static features are considered to be features that are extracted from the malware binary without being executed. Tesauro et al. [22] have classified malware using an artificial neural network by extracting a list of byte sequence trigram. In [23], Tain et al. have used a feature to classify trojan [7] of seven different types, and the feature used is the length of the program and was able to obtain an average accuracy of 88%.

Dynamic features are those features that can be drawn out by executing and observing the behavior of the malwares. In this model, the behavior of the malwares and malware API calls [8] have been used. In [24], Rieck et al. have classified malwares according to their family in which feature is based on the behavior analysis of the malware. They used a 10,072 malicious sample labeled dataset in which an anti-virus program is labeled and the dataset has been split into 14 malware families. Then, they supervised and evaluated the activities of all the malware in a sandbox environment [9] that produced a behavioral report. They generated a function vector based on the frequency of each malware by using this report. For training and testing, a support vector machine [10] was used on the families of 14 malwares and 88% average classification accuracy is being reported.

Besides these, many visualization techniques have been proposed to understand the behavior of the malwares. Like Nataraj et al., [25] has classified the malicious software using image processing technique according to its visualization as grayscale images. In their work, by extracting the GIST characteristics from the malware

images, they used a qualified k-nearest neighbor [11] for classification. There is also a method that is proposed that is similar to ours which is good but due to changes proposed in our model (using VGG-16[12] with few layers removed), we have achieved higher accuracy.

3 Background

Convolutional Neural Network (CovNet) is a feed forward based on a deep learning algorithm and which is based on an animal's audio-visual cortex. CNN can be represented as a type of neural network model in which it is possible to extract higher representations of image information. CNN takes the raw pixel data of the image, trains the model, then extracts the features automatically for better classification, unlike classical image recognition, where one is required to define image features. Importance (biases and weights(learnable)) is being allocated to the plethora of aspects/objects in the image upon taking the image as an input which in turn helps in differentiating.

3.1 Convolutional Layers

The convolutional layer incorporates a progression of channels that should be instructed about the parameters. The layers in the image sequence are utilized to force a specific measure of activities related to convolution. For each spatial contrast, a dot product is taken between the channel just as the input image by moving the filter over the height and width of the source. The main point of these layers is to wipe out highlights, for example, borders, shading, angle direction, and so on. The first layer excludes low-level capacities. Low-level characteristics incorporate inclination, shading, borders, and direction. The model could adjust to a significant level job with an unpredictable layer framework by adding layers, which thus would give us a superior comprehension of pictures and accordingly more prominent precision with more subtleties.

3.2 Pooling Layers [13]

The pooling Layer is responsible for truncating the spatial scale of the Convolved Function (downsampling). This is done to reduce the amount of computing power needed to manage the data's size. Moreover, it is useful for extricating useful features which are positional and rotational invariant, thus the effective training of the model takes place. The maximum value is returned from the portion of the image where the Max Pooling layer is present. Whereas, an average of all values is returned by the Kernel over the image when Average Pooling is done.

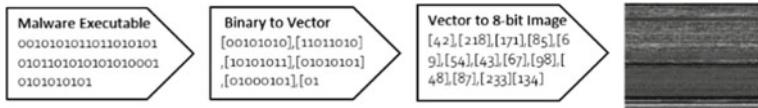


Fig. 2 Malware visualization process summary

3.3 Fully Connected Layers

The function of this layer is to take the output of convolutional layers and pooling layers as an input and then classification is done on it. Every neuron is connected to each and every other neuron of another layer. For classifying, the flattened matrix passes through a fully connected layer.

4 Proposed Approach

4.1 Visualizing Malware from Binary to Image

Malware is mostly present in executable formats that are slightly modified to change the signature of the code. These changes cannot be detected by using traditional approaches. But looking at the binary file, it can be observed that most of the code remains the same, resulting in a similar architecture of the malware but a different signature.

In the proposed approach, malware binary (0, 010, 101, 101, 010) is first divided into chunks of 8 bits. Then the 8-bit vector is converted to the decimal equivalent of unsigned integers [00000000] as 0 (black) and [11111111] as 255 (white). Figure [2] shows the representation of the same. Further, in the 2D-Matrix, the decimal vector is being reshaped which is represented as a grayscale image. Width is of fixed size whereas the size of the height is variable corresponding to the file size.

4.2 Model Outline

To identify malware, we use the Convolutional Neural Network (CNN). Here, we have used a custom CNN model (VGG-16 with few layers removed) which takes the grayscale image (malware image) as an input and classifies the image from 25 different classes as part of one of those 25. The score from the highest is taken as a forecast by getting the scores from the various groups.

4.3 Learning

The Malimg dataset has 25 classes of malware with 9,339 Images in total. In the proposed model, we used cross-entropy loss in order to train the CNN Network. The categorical cross-entropy loss function is calculated by the following function:

$$\text{Loss} = - \sum_{i=1}^{\text{output size}} y_i \log i \quad (1)$$

where y^i is the i -th scalar value in the model output and y^i is the target value to be achieved. Scalar values are represented by output size in the model output. The loss is calculated to measure the probability distribution of distinguishing two classes from each other. The Activation Function used with Convolutional Neural Network is ReLU. Rectified Linear Activation Function (ReLU) [14] is a linear activation function that, if the input value is negative, gives the output as zero but if the input is given as positive then the output is the same as the input. ReLU is found to improve the performance of deep learning models and is widely used by the community because of the ease of use and less overload of training. It has often shown better performance than many other activation functions. The activation function that we used for classifying is SoftMax. SoftMax [15] is usually used in the last layers of CNN. SoftMax transforms logits, i.e., The numerical output is taken from the linear layer of the last layer of the classification of the multi-class neural network and translated into probabilities by taking the exponents of each output and then using the sum of those exponents to normalize each number, so that the total output vector adds up to 1.

5 Experiments

5.1 Datasets and Experimental Setting

All the experiments were performed on the Malimg dataset:-

A total of 9339 malware samples in the Malimg dataset [31] are represented as grayscale images. In the dataset, there are 25 families/groups present and each one of the malware in the dataset belongs to one of several families, and the number of samples dispersed among the classes equally. The malware in the dataset belongs to the VB.AT family, Malex.gen! J, Yuner.A, Autorun.K, VB.AT, Yuner.A, Rbot! Gen, packaged, packed (UPX). For preliminary analysis, unpackaging of them was performed. We have randomly chosen 90% of the samples from the dataset in our experiment to train the model and the rest of the 10% are used for model testing. In the

end, we have (8405) training samples and (934) testing samples for the performance evaluation of the proposed model (Fig. 3).

For experimental purpose, we have used Google Colaboratory [25] with GPU enabled. Configuration of the virtual machine was as follows: Dual Core (TM) Intel Xenon CPU (2.30 GHz) with 12 GB RAM and Nvidia P40 GPU 12 GB RAM. Operating System used was Ubuntu 20.04 64bit. Colab is widely used by researchers all over the world.

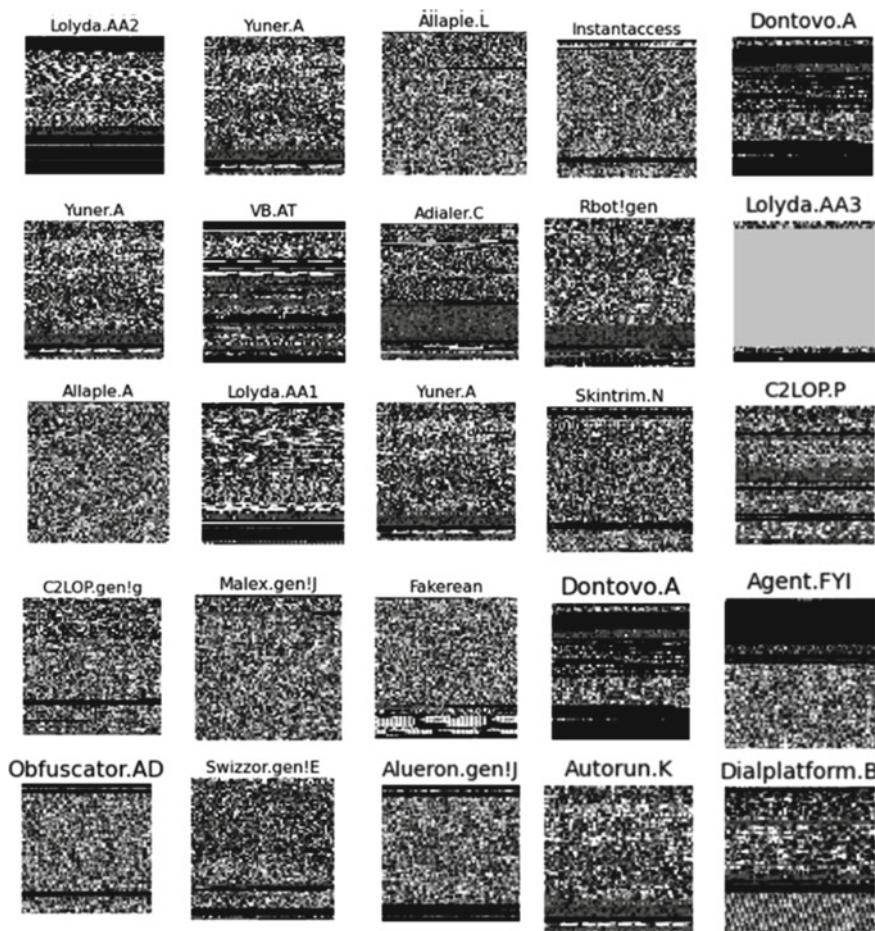


Fig. 3 Malware images belonging to various malware families

Table 1 Quantitative results on the Malimg dataset

Method	Accuracy (%)
Nataraj et al	97.18
GIST + SVM	93.23
M-CNN[27]	98.52
(Proposed Model)	98.97

5.2 Experiments and Evaluation

Previous work has been done on this particular dataset. We have used the percentage model to evaluate the accuracy of the proposed model and predict the class of malware.

The proposed model is trained for 20 epochs with batch sizes of 10,000 on the Malimg dataset. The performance of various methods with their claimed accuracy by different authors is shown in Table 1. An accuracy of 93.23% was achieved by GIST + SVM, whereas our proposed model has achieved 98.97% accuracy. According to Nataraj et al. method, the best accuracy achieved was only 97.18%.

6 Conclusion

Analyzing the Macfee Lab report, we can see that the number of malware and new malware has been increasing year by year, so this increase in malicious software poses a serious and significant threat to the systems. As the world is moving toward digitalization, this threat increases and has become a catalyst for a world with higher cyber theft and felonies. Hence, it became essential to study/analyze the malware's behavior and categorize its samples so that a program can be developed to counter this sudden ostentatious increase in malware. Thus, we have proposed a functional Convolutional Neural Network for the de-obfuscation of malware, which in turn help us in countering the rise. The result of the experiment shows the effectiveness of the method that we have proposed.

References

1. Ciresan, D.C., Meier, U., Masci, J., Maria Gambardella, L., Schmidhuber, J.: Flexible, high performance convolutional neural networks for image classification. In: IJCAI Proceedings-International Joint Conference on Artificial Intelligence, vol. 22, p. 1237 (2011)
2. <https://www.mcafee.com/enterprise/en-us/assets/reports/rp-quarterly-threats-nov-2020.pdf>
3. Galeon, D., Gphd, C.: Dubai just appointed a “State Minister for Artificial Intelligence”, 20 Oct 2017. <https://futurism.com/dubai-just-appointed-a-state-minister-for-artificial-intelligence>. Accessed 22 Nov 2017

4. Yi, J., Nasukawa, T., Bunescu, R., Niblack, W.: Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. In: Third IEEE international conference on data mining, 2003. ICDM 2003, pp. 427–434. IEEE, Nov 2003
5. https://www.internetsociety.org/resources/doc/2017/artificial-intelligence-and-machine-learning-policy-paper/?gclid=Cj0KCQiA2af-BRDzARIsAVQUOcDI_pemfCfUVVY8-EOAKFObhZlgJpNUy0F35p_ZdwKRmvsrSO8msaAoroEALw_wcB
6. <https://towardsdatascience.com/treemaps-why-and-how-cfb1e1c863e8>
7. Shugang, T.: The detection of Trojan Horse based on the data mining. In: Sixth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD '09, pp. 311–314 (2009)
8. https://medium.com/@jamesbeck_12148/how-does-an-api-call-work-e771adc08a71
9. <https://searchsecurity.techtarget.com/definition/sandbox>
10. Bartlett, P., Shawe-Taylor, J.: Generalization performance of support vector machine and other pattern classifiers. In: Burges, C., Scholkopf, B. (eds.), Advances in Kernel Methods—Support Vector Learning. MIT Press (1998)
11. Wang, H.: Nearest neighbours without k: a classification formalism based on probability, technical report, Faculty of Informatics, University of Ulster, Northern Ireland, UK (2002)
12. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
13. <https://medium.com/ai-in-plain-english/pooling-layer-beginner-to-intermediate-fa0dbdce80eb>
14. Abien, F.A.: A neural network architecture combining Gated Recurrent Unit (GRU) and support vector machine (SVM) for intrusion detection in network traffic data (2017). [arXiv:1709.03082](https://arxiv.org/abs/1709.03082)
15. Facchinei, F., Pang, J.-S.: Finite-dimensional Variational Inequalities and Complementarity Problems, vol. I. Springer Series in Operations Research, Springer, New York (2003)
16. Yoo, I.: Visualizing windows executable viruses using self-organizing maps. International Workshop on Visualization for Cyber Security (VizSec) (2004)
17. Quist, D.A., Liebrock, L.M.: Visualizing compiled executables for malware analysis. In: International Workshop on Visualization for Cyber Security (VizSec), pp. 27–32 (2009)
18. Goodall, J.H., Randwan, H., Halseth, L.: Visual analysis of code Security. In: International Workshop on Visualization for Cyber Security (VizSec) (2010)
19. Trinius, P.H.T., Gobel, J., Freiling, F.C.: Visual analysis of malware behavior using treemaps and thread graphs. In: International Workshop on Visualization for Cyber Security (VizSec), pp. 33–38
20. Gandotra, E., Bansal, D., Sofat, S.: Malware analysis and classification: a survey. *J. Inf. Secur.* **5**, 56–64 (2014)
21. Ranvee, S., Hiray, S.: Comparative analysis of feature extraction methods of malware detection. *Int. J. Comput. Appl.* **120**, 1–7 (2015)
22. Tesauro, G., Kephart, J., Sorkin, G.B.: Neural networks for computer virus recognition. In: IEEE International Conference on Intelligence and Security Informatics, vol. 11 (1996)
23. Tian, R., Batten, L.M., Versteeg, S.C.: Function length as a tool for malware classification. In: 3rd International Conference on Malicious and Unwanted Software (MALWARE) (2008)
24. Rieck, K., Holz, T., Willems, C., Dussel, P., Laskov, P.: Learning and classification of malware behavior. In: Fifth Conference on Detection of Intrusions and Malware & Vulnerability Assessment (DIMVA'08), pp. 108–125 (2008)
25. Nataraj, L., Karthikeyan, S., Jacob, G., Manjunath, B.S.: Malware images: visualization and automatic classification. In: Proceedings of the 8th International Symposium on Visualization for Cyber Security, VizSec '11, pp. 4:1–4:7. ACM, New York, NY, USA (2011)
26. <https://colab.research.google.com/notebooks/intro.ipynb>
27. Mahmoud, K., Mrigank, R., Noman, M., Neil, D.B.B., Wang, Y., Iqbal, F.: Malware Classification with Deep Convolutional Neural Network (2018)
28. <http://dl.acm.org/citation.cfm?id=2016908>

Content-Based Image Retrieval Using Energy-Based Frequency Domain Features



Hillol Barman, Netalkar Rohan Kishor, U. S. N. Raju ,
Debanjan Pathak , and Sweta Panigrahi 

Abstract Content-Based Image Retrieval (CBIR) has become one of the trending areas of research in computer vision. In traditional CBIR, the features in a spatial domain, such as color, texture, shape, and point features, are extracted. It is often considered that apart from the spatial features, the features extracted from the frequency domain of the images can give further information on the features of an image. In this paper, we are proposing two novel methods for the purpose of feature extraction from the 2-dimensional Discrete Cosine Transform (DCT) of an image: DCT_256_Zigzag and DCT_256_2 × 2. These methods take into consideration the lower frequencies in order to determine the features in the frequency domain. The advantage of using the zigzag scanning is to have the maximum low frequency values having Higher Energies comparatively. These two features are combined with two of the existing spatial domain features: Local Binary Patterns (LBP) and Interchannel voting features to generate a global feature vector for an image. For a query image, its feature vector is compared with feature vectors of every other image in the database using d1-distance, and the images with least distance are considered most similar images to the query image. To evaluate the efficiency of these two methods, five standard performance measures such as Average Precision Rate (APR), Average Recall Rate (ARR), F-Measure, Average Normalized Modified Retrieval Rank (ANMRR), and Total Minimum Retrieval Epoch (TMRE) are used. Six benchmark image datasets Corel-1000, Corel-5000, Corel-10000, VisTex, STex, and Color-Brodatz are used to corroborate the performance of these methods.

Keywords CBIR · DCT · Frequency domain · Interchannel voting · LBP · TMRE

H. Barman · N. R. Kishor · U. S. N. Raju () · D. Pathak · S. Panigrahi

Department of Computer Science and Engg, National Institute of Technology Warangal, Warangal 506004, Telengana, India

e-mail: usnraju@nitw.ac.in

1 Introduction

An image is a kind of important information carrier. With the development of information and networking, the varieties and quantity of an image are increasing fast. The number of images captured in 2006 exceeded 250 billion worldwide. In 2019, the number of photos taken exceeded 1.4249 trillion and by 2020 the number is predicted to rise by 0.8%–1.4363 trillion images. How to retrieve the desired image rapidly and accurately among massive image data storage is becoming an urgent problem.

Traditional image retrieval systems are mostly based on metadata such as keywords, tags, and/or descriptions associated with the image. These kinds of systems may produce lots of garbage in search results if the metadata of images is not well filled. Also having humans manually enter keywords for images in a large database can be inefficient, expensive, and may not capture every keyword that describes the image.

Content-based image retrieval research has produced a number of search engines that can retrieve images based on local or global features derived from color, texture, and simple shape information. The term “content” signifies that images are retrieved based on some features which can be calculated from the actual content of images. The retrieval process depends on the similarity between the query image and all other images of the given dataset. Feature vector comparison is one of the possible ways to find similarities between the corresponding images. In traditional CBIR, features of an image can be Local and Global texture features, Point features, and Shape features or as mentioned above, Color features.

To produce color features of an image, feature extraction procedures like color histogram [1], color correlogram [1], color autocorrelogram [2, 3], and interchannel voting between hue and saturation [4] can be applied on the image. For texture feature extraction, LBP [5], ULBP [5], CS_LBP [6], LEP [7], LDP [8], and LTrP [9] can be used. One more texture feature descriptor is GLCM, which reveals knowledge about pixel pair co-occurrence of the image [10, 11]. For extracting shape information, HOG [12], Wavelet Fourier descriptor [13], Fourier descriptor [13], angular pattern and binary angular pattern [14], Convex Hull [15], etc. can be used. In [16–18], different image retrieval methods have been covered and discussed. Suresh et al. [19] proposed a new color feature named interchannel voting among the three components of a HSI image. This method explores the interrelationship among three components: Hue (H), Saturation (S), and Intensity (I) of a color image.

The frequency domain representation of an image represents the rate at which the values of the pixels change in the spatial domain. The transformation of an image from the spatial domain to the frequency domain can be done with the help of Discrete Fourier Transform (DFT) or Discrete Cosine Transform (DCT). The 2-dimensional Discrete Fourier Transform (DFT) of an image can be achieved with the help of (1), and the Inverse Discrete Fourier Transform (IDFT) can be done with the help of (2) for an image of size $M \times N$ [20].

$$F(u, v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) e^{-j2\pi(ux/M + vy/N)} \quad (1)$$

$$f(x, y) = \frac{1}{MN} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} F(u, v) e^{j2\pi(ux/M + vy/N)} \quad (2)$$

where $0 \leq u < M$ and $0 \leq v < N$, $0 \leq x < M$ and $0 \leq y < N$, $f(x, y)$ represents an image, and $F(u, v)$ represents the DFT transform of the image. DFT consists of both real and imaginary parts. We have another way of representing the image in the frequency domain with the help of DCT which uses only the real part for representation. The Property of DCT that makes it more suitable for feature extraction is that it has better energy compaction than the DFT, with which it can constitute the majority of energy coefficients in the sequence with a few number of transform coefficients. This makes DCT more suitable for feature extraction than DFT. The equations for performing the DCT and Inverse Discrete Cosine Transform (IDCT) of an image are given in (3) and (4).

$$F(u, v) = \frac{2}{\sqrt{MN}} C(u) C(v) \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) \cos \frac{(2x+1)u\pi}{2M} \cos \frac{(2x+1)v\pi}{2N} \quad (3)$$

$$f(x, y) = \frac{2}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} C(u) C(v) F(u, v) \cos \frac{(2x+1)u\pi}{2M} \cos \frac{(2x+1)v\pi}{2N} \quad (4)$$

where $0 \leq u < M$ and $0 \leq v < N$, $0 \leq x < M$ and $0 \leq y < N$, $f(x, y)$ represents an image, and $F(u, v)$ represents the DCT transform of the image. The features extracted from the frequency domain can be used for image enhancement, restoration, compression, watermarking, representation and description, and for CBIR. Kobayashi et al. [21] have used the frequency domain feature for CBIR. Stuchi [22] et al. have used frequency domain layers of CNN architecture for feature extraction of CBIR. Here in this paper, we proposed and used two features based on frequency domain and used them to combine with spatial domain features.

2 Methodology

The general structure of any CBIR is given in Algorithm-1. The process of obtaining features by using the proposed methods is given in Algorithm-2 and Algorithm-3. In this paper, we proposed the following two methods for feature extraction. The features are obtained from the information from the frequency domain. The two proposed algorithms are given in Algorithm-2 and Algorithm-3.

Algorithm-1: General CBIR process

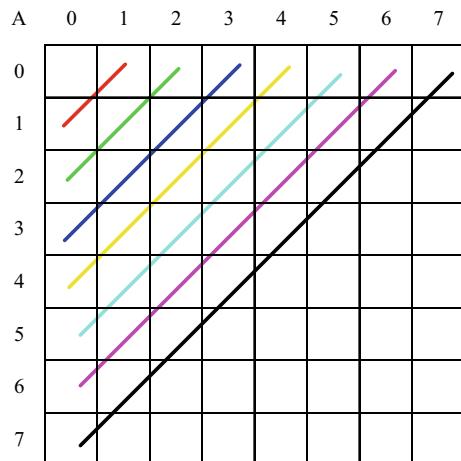
1. Take a query image
 2. Extract the features from the image using the same method used to extract features of image dataset
 3. Compare the feature vector of the query image with the feature vectors of the images stored in the image dataset to determine the closest image w.r.t. distance.
 4. Retrieve the image(s) having lowest distance which thereby means the image(s) are having most similarity to the query image
-

Algorithm-2: Feature Extraction Algorithm: DCT_128

1. Take the image from dataset
 2. Divide the image into blocks of 8×8 size. If the Block cannot be divided into 8×8 block then padding (0-padding the right and bottom of the image) is done to get an 8×8 block
 3. Working from left to right, top to bottom, compute the DCT by using Eq. (3) for each of these 8×8 blocks
 4. Now, for each block calculate the 7-bit code using the values in the DCT as
 - a. For Most Significant Bit, calculate the average of the values of the first diagonal, i.e. A01 and A10 as shown in Fig. 1. After finding the average of the numbers (using Eq. (5)), based on the value obtained from the average, the bit is set if the value is greater than or equal to zero else it is not set
 - b. Similarly, for 2nd MSB, calculate the average of all the values of the second diagonal, i.e. A02, A11, and A20 as shown in Fig. 1. After finding the average using Eq. (6), based on the value obtained the bit is set, and so on. In this way, we will get a 7-bit code. For the remaining 5-bits, Eqs. (7) to (11) are used
 5. After getting 7-bit code for one block, step 4 is repeated for each block in the image
 6. After calculating the 7-bit codes for the entire image, we store the histogram of the 7-bit codes as the feature vector of the image
-

$$a[0] = (A01 + A10)/2 \quad (5)$$

Fig. 1 Values used for computing the pattern for proposed DCT_128



$$a[1] = (A02 + A11 + A20)/3 \quad (6)$$

$$a[2] = (A03 + A12 + A21 + A30)/4 \quad (7)$$

$$a[3] = (A04 + A13 + A22 + A31 + A40)/5 \quad (8)$$

$$a[4] = (A05 + A14 + A23 + A32 + A41 + A50)/6 \quad (9)$$

$$a[5] = (A06 + A15 + A24 + A33 + A42 + A51 + A60)/7 \quad (10)$$

$$a[6] = (A07 + A16 + A25 + A34 + A43 + A52 + A61 + A70)/8 \quad (11)$$

Algorithm-3: Feature Extraction Algorithm: DCT_256

1. Take the image from dataset
2. Divide the image into blocks of 8×8 size. If the image cannot be divided into 8×8 blocks, then padding (0-padding the right and bottom of the image) is done to get an image that can be divided into perfect 8×8 blocks
3. Working from left to right, top to bottom, compute the DCT by using Eq. (3) for each of these 8×8 blocks
4. Now, for each block, calculate the 8-bit code using the values in the DCT as given in Eqs. (12) to (19) and making into binary patterns as explained in Algorithm-2. The energy values used for calculating this are shown in Fig. 2
5. After getting 8-bit code for one block, step 4 is repeated for each block in the image
6. After calculating the 8-bit codes for the entire image, we store the histogram of the 8-bit codes which is the feature vector of the image.

$$a[0] = A01 \quad (12)$$

$$a[1] = A11 \quad (13)$$

$$a[2] = A10 \quad (14)$$

$$a[3] = (A02 + A03 + A12 + A13)/4 \quad (15)$$

$$a[4] = (A22 + A23 + A32 + A33)/4 \quad (16)$$

$$a[5] = (A20 + A21 + A30 + A31)/4 \quad (17)$$

Fig. 2 Values used for computing the pattern for proposed DCT_256

A	0	1	2	3	4	5	6	7
0								
1								
2								
3								
4								
5								
6								
7								

$$a[6] = \left(\begin{array}{l} A04 + A05 + A06 + A07 + \\ A14 + A15 + A16 + A17 + \\ A24 + A25 + A26 + A27 + \\ A34 + A35 + A36 + A37 \end{array} \right) / 16 \quad (18)$$

$$a[7] = \left(\begin{array}{l} A40 + A50 + A60 + A70 + \\ A41 + A51 + A61 + A71 + \\ A42 + A52 + A62 + A72 + \\ A43 + A53 + A63 + A73 \end{array} \right) / 16 \quad (19)$$

3 Results and Discussions

3.1 Analysis

3.1.1 Average Precision Rate (APR) and Average Recall Rate (ARR)

Precision is defined as a ratio between the number of total relevant images retrieved and the number of total images retrieved for a given query, as given in (20).

$$P(i) = \frac{1}{n} \sum_{k=1}^n Rank(k, i) \quad (20)$$

where $Rank(k, i) = 1$ if image(k) belongs to the same class as image(i), else $Rank(k, i) = 0$ and n is the number of images retrieved.

3.1.2 Recall

Recall is defined as the ratio between the number of total relevant images retrieved and the number of total images having the same class as a query image, given in (21).

$$R(i) = \frac{1}{N_{ic}} \sum_{k=1}^n Rank(k, i) \quad (21)$$

where N_{ic} is the number of images in the database belonging to the same class as image(i).

Average precision for different step sizes m_1, m_2, \dots, m_k is known as APR. Similarly, average recall for different step sizes is known as ARR.

3.1.3 F-Measure

It is represented by a single value to reflect the relationship between precision and recall. It is obtained by assigning equal weight to both precision and recall in the harmonic mean calculation as given in Eq. (22).

$$F - Measure(n) = \frac{(2 \times APR \times ARR)}{(APR + ARR)} \quad (22)$$

3.1.4 Average Normalized Modified Retrieval Rank (ANMRR)

It is used to measure the retrieval accuracy. To calculate ANMRR for each image, we consider only those images whose rank is less than $2 \times (\text{number of images in the class})$. If an image's rank is less than $2 \times (\text{number of images in the class})$, then the score of that image is the rank of the image, else it is a predefined fixed number. Now the average score is calculated and then the normalized score.

3.1.5 Total Minimum Retrieval Epoch (TMRE)

Minimum Retrieval Epoch (MRE) is calculated as the ratio between the average total number of traversed images required to retrieve all images having the same class of each query image and the number of images in that class, given in (23).

$$MRE(i) = \frac{\max(k) \forall k \exists Rank(k, i) \in C_i}{N} \quad (23)$$

Thus, Total Minimum Retrieval Epoch (TMRE) is the average MRE for all the images in the dataset, given in (24).

$$TMRE = \frac{1}{N} \sum_{k=1}^n MRE(k) \quad (24)$$

3.2 Results

For performance evaluation of the two proposed methods, six benchmark color datasets are used. Three of these image datasets are natural image datasets and three are color texture datasets. Each of the query images gives a feature vector by following the steps of the proposed method during evaluation. By using d1-distance, the comparison between the query image feature vector and dataset images' feature vector is carried out. A rank matrix is then obtained from the distances calculated. By using this rank matrix, the five performance measures Precision, Recall, ANMRR, TMRE, and F-measure are calculated and compared with different existing methods. The results of all these are given in this section.

Dataset-1(Corel-1 K): This dataset [12] consists of a total of 1000 images. There are 10 categories with each category consisting of 100 images. The different categories are Africans, Beaches... Food. The size of each image in this image dataset is 384×256 or 256×384 . Three images from each group, a total of 30 images of this dataset, are shown in Fig. 3. The five performance measure values for this image dataset are shown in Table 1.

Dataset-2(Corel-5 K): Corel-5 K image dataset [13] consists of 5000 images, with 50 categories where each category is of 100 images. Figure 4 shows a total of 50 images, one image from each of the 50 categories. As in the Corel-1 K image dataset, here too all the five performance measures are evaluated and shown in Table 2.

Dataset-3(Corel-10 K): The third natural image dataset considered is the Corel-10 K image dataset [13]. This dataset consists of 100 categories with 100 images in each category resulting in a total of 10000 images. This dataset contains a total of 53 different-sized images: 128×192 , 192×128 , etc. In Fig. 5, a total of 100 images are given, one from each of the 100 categories of this image dataset. Table 3 shows all the five performance measure results of the Corel-10 K dataset.

Dataset-4(VisTex): This image dataset is the first texture image dataset considered called the VisTex texture dataset [14] with 484 images. Out of these 484 texture images, 40 are considered for experimentation. The actual image dimension is 512×512 . Each image of these 40 is made into 16 nonoverlapping sub-images where each sub-image is of dimension 128×128 , which results in a total of 640 texture



Fig. 3 Corel-1 K samples (three images per category)

Table 1 Performance measures for Corel-1 K

	APR	ARR	F-Msr	ANMRR	TMRE
LBP	69.18	38.69	31.228	0.5208	8.03
ULBP	69.26	44.42	33.929	0.4574	7.49
Interchannel	80.15	50.94	39.502	0.3880	7.35
DCT_64	47.87	30.38	22.739	0.6081	8.49
DCT_128	48.73	33.09	24.485	0.5719	8.16
DCT_256	45.98	30.07	22.457	0.6066	8.21
LBP + DCT_64	70.34	39.72	31.896	0.5105	7.97
LBP + DCT_128	70.02	40.23	32.081	0.5048	7.92
LBP + DCT_256	67.60	39.83	31.491	0.5085	7.86
Interchannel + DCT_64	80.40	51.24	39.728	0.3850	7.30
Interchannel + DCT_128	80.59	51.39	39.840	0.3833	7.25
Interchannel + DCT_256	80.24	51.28	39.738	0.3847	7.19

image datasets. From these 640 images Fig. 1, 17, 33, 49 ... 625 which are the 1st sub-image of each of 40 actual texture images, are shown in Fig. 6. All the performances are obtained on this 640 texture image dataset. Table 4 shows the five performance measure values.

Dataset-5 (STex): The other color texture dataset considered is the Salzburg Texture Image Dataset (STex) [15] which contains a total of 476 texture images. All these 476 images are considered to test the performance of different methods.

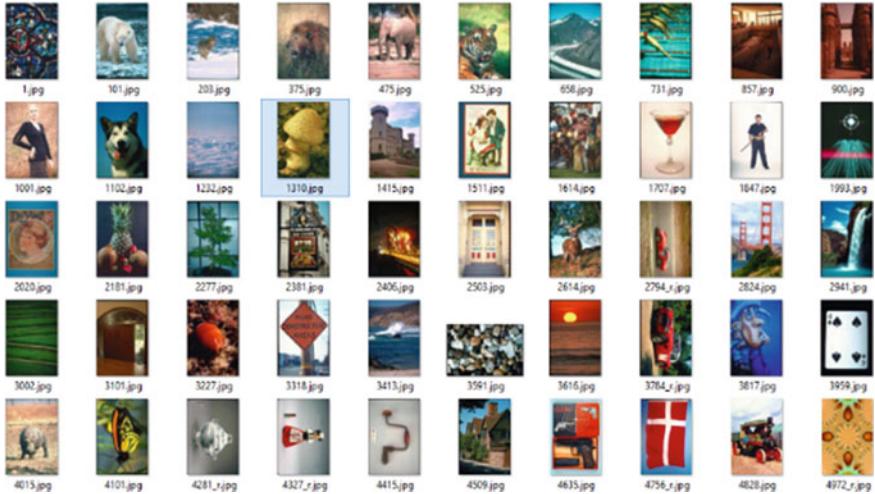
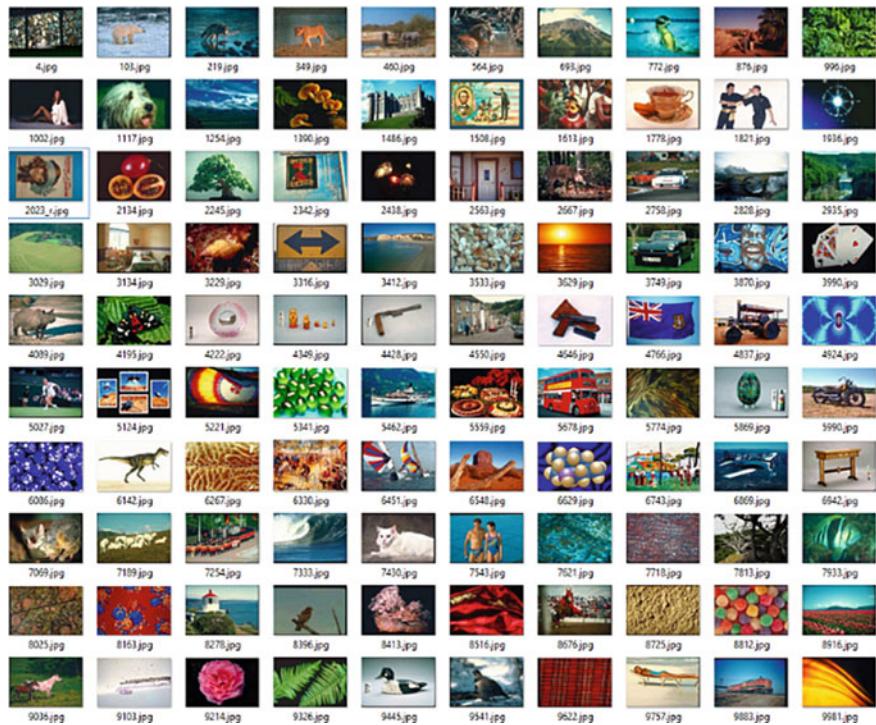


Fig. 4 Corel-5 K samples (one image per category)

Table 2 Performance measures for Core-5 K

	APR	ARR	F-Msr	ANMRR	TMRE
LBP	46.19	19.85	16.667	0.7446	41.62
ULBP	46.94	21.10	17.584	0.7276	38.87
Interchannel	62.60	30.23	25.250	0.6269	36.55
DCT_64	18.59	7.17	5.858	0.9018	46.30
DCT_128	22.39	9.64	7.768	0.8693	44.85
DCT_256	20.36	8.45	6.841	0.8852	46.42
LBP + DCT_64	45.99	20.31	16.976	0.7393	41.39
LBP + DCT_128	44.60	20.23	16.799	0.7397	41.00
LBP + DCT_256	40.10	18.53	15.177	0.7589	41.04
Interchannel + DCT_64	63.06	30.57	25.499	0.6233	36.43
Interchannel + DCT_128	63.32	30.93	25.770	0.6190	36.17
Interchannel + DCT_256	62.80	30.72	25.583	0.6211	36.22

Here also, each texture image is made into 16 non-overlapping sub-images which results in a total of 7616, where each sub-image is of dimension 128×128 . Figure 7 shows 40 of these 7616, where each of these 40 is considered from 40 different actual texture images from 476 texture images. All the five performance measures are given in Table 5.

**Fig. 5** Corel-10 K Samples**Table 3** Performance measures for Corel-10 K

	APR	ARR	F-Msr	ANMRR	TMRE
LBP	38.17	15.20	12.969	0.8057	83.20
ULBP	38.61	15.65	13.318	0.8001	79.67
Interchannel	53.57	23.10	19.728	0.7121	73.57
DCT_64	15.25	4.68	4.045	0.9368	93.48
DCT_128	17.68	6.23	5.243	0.9163	90.92
DCT_256	16.68	5.61	4.766	0.9245	93.44
LBP + DCT_64	38.07	15.43	13.099	0.8029	82.95
LBP + DCT_128	36.49	15.23	12.816	0.8048	82.26
LBP + DCT_256	32.39	13.69	11.398	0.8229	82.36
Interchannel + DCT_64	54.01	23.35	19.936	0.7093	73.49
Interchannel + DCT_128	54.43	23.62	20.147	0.7060	73.06
Interchannel + DCT_256	54.09	23.52	20.016	0.7074	73.27



Fig. 6 Forty VisTex texture images considered

Table 4 Performance measures for VisTex

	APR	ARR	F-Msr	ANMRR	TMRE
LBP	97.07	81.00	64.357	0.1216	4.33
ULBP	98.13	82.75	65.522	0.1081	3.96
Interchannel	98.67	81.56	65.564	0.1280	5.21
DCT_64	35.43	15.68	15.250	0.7940	30.29
DCT_128	37.38	17.38	16.622	0.7695	28.98
DCT_256	38.32	17.99	17.181	0.7662	30.74
LBP + DCT_64	96.72	80.10	63.752	0.1256	4.47
LBP + DCT_128	95.74	76.41	61.798	0.1449	4.55
LBP + DCT_256	93.52	72.41	59.065	0.1775	5.16
Interchannel + DCT_64	98.83	81.70	65.655	0.1260	5.17
Interchannel + DCT_128	98.63	81.91	65.665	0.1243	5.06
Interchannel + DCT_256	98.63	81.55	65.495	0.1261	5.09

Dataset-6 (Color Brodatz): This is the last image dataset considered, color Brodatz texture image dataset [16]. We made each of the images into 25 non-overlapping sub-images, which results in a total of 2800 images. The first sub-images from each of these 112 are shown in Fig. 8 and the results are shown in Table 6.



Fig. 7 Forty of the STex texture images from 7616 images

Table 5 Performance measures for STex

	APR	ARR	F-Msr	ANMRR	TMRE
LBP	76.92	47.22	41.849	0.4526	104.11
ULBP	82.54	52.39	45.915	0.4027	86.54
Interchannel	93.44	68.53	57.245	0.2422	42.43
DCT_64	30.55	10.71	11.542	0.8668	322.74
DCT_128	31.29	11.41	12.082	0.8583	315.64
DCT_256	32.91	12.76	13.197	0.8425	341.15
LBP + DCT_64	75.86	47.13	41.488	0.4539	102.27
LBP + DCT_128	72.00	44.57	39.232	0.4801	101.53
LBP + DCT_256	67.71	41.36	36.512	0.5134	104.71
Interchannel + DCT_64	93.51	68.89	57.451	0.2391	42.11
Interchannel + DCT_128	93.62	69.03	57.523	0.2375	40.79
Interchannel + DCT_256	93.35	69.00	57.481	0.2371	41.13

4 Conclusion and Future Extension

This paper proposes two new DCT-based features for CBIR. The two proposed methods individually show better performance when compared with the existing DCT_64 on the three natural image datasets and also on three color image texture

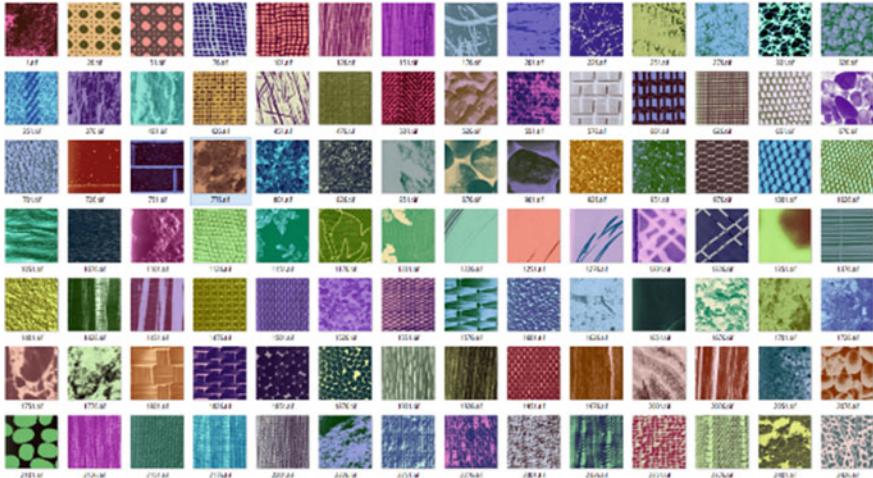


Fig. 8 112 Textures each from Color Brodatz texture

Table 6 Performance measures for Color Brodatz

	APR	ARR	F-Msr	ANMRR	TMRE
LBP	89.29	70.22	54.997	0.2247	13.92
ULBP	91.97	74.39	57.630	0.1940	12.36
Interchannel	99.64	90.41	67.304	0.0648	6.28
DCT_64	36.33	17.25	15.597	0.7847	68.35
DCT_128	35.73	17.39	15.455	0.7829	66.38
DCT_256	38.46	19.67	17.274	0.7599	72.90
LBP + DCT_64	89.32	70.07	54.857	0.2233	13.51
LBP + DCT_128	86.66	67.55	53.028	0.2433	13.54
LBP + DCT_256	84.44	64.37	50.809	0.2733	15.06
Interchannel + DCT_64	99.62	90.57	67.368	0.0637	6.09
Interchannel + DCT_128	99.62	90.54	67.337	0.0634	6.05
Interchannel + DCT_256	99.59	90.24	67.242	0.0645	6.21

datasets. The results show that when the proposed features are used combined with Interchannel voting, the performance is increased when compared with LBP + DCT_64. For Interchannel + DCT_128, the natural datasets give us an improvement of 29.81% and improvement of 26.77% for the color texture datasets, i.e. for the six datasets, it is 28.29%.

Future Extension: To improve the accuracy of CBIR, frequency extractor layers in Deep CNN models can be used. To reduce the time for processing, the Map Reduce paradigm can be used along with SPARK.

References

1. Singha, M., Hemachandran, K.: Content based image retrieval using color and texture. *Signal Image Process* **3**(1), 39–57 (2012). DOIurl: <https://10.5121/sipij.2012.3104>
2. Huang, J., Kumar, S.R., Mitra, M., Zhu, W.J., Zabih, R.: Image indexing using color correlograms. In: Proceedings of IEEE computer society conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, USA, USA, 17–19 June, pp. 762–768 (1997). <https://doi.org/10.1109/CVPR.1997.609412>
3. Chun, Y.D., Kim, N.C., Jang, I.H.: Content-based image retrieval using multiresolution color and texture features. *IEEE Trans. Multimedia* **10**(6), 1073–1084 (2008). <https://doi.org/10.1109/TMM.2008.2001357>
4. Bhunia, A.K., Bhattacharyya, A., Banerjee, P., Roy, P.P., Murala, S.: A novel feature descriptor for image retrieval by combining modified color histogram and diagonally symmetric co-occurrence texture pattern (2018). [arXiv:1801.00879](https://arxiv.org/abs/1801.00879)
5. Ojala, T., Pietikäinen, M., Maenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7), 971–987 (2002). <https://doi.org/10.1109/TPAMI.2002.1017623>
6. Heikkilä, M., Pietikäinen, M., Schmid, C.: Description of interest regions with center-symmetric local binary patterns. *Comput. Vis. Graph. Image Process.* **43**(8), 58–69 (2006). https://doi.org/10.1007/11949619_6
7. Verma, M., Raman, B., Murala, S.: Local extrema co-occurrence pattern for color and texture image retrieval. *Neurocomputing* **165**, 255–269 (2015). <https://doi.org/10.1016/j.neucom.2015.03.015>
8. Zhang, B., Gao, Y., Zhao, S., Liu, J.: Local derivative pattern versus local binary pattern: face recognition with high-order local pattern descriptor. *IEEE Trans. Image Process.* **19**(2), 533–544 (2009). <https://doi.org/10.1109/TIP.2009.203588>
9. Murala, S., Maheshwari, R.P., Balasubramanian, R.: Local tetra patterns: a new feature descriptor for content-based image retrieval. *IEEE Trans. Image Process.* **21**(5), 2874–2886 (2012). <https://doi.org/10.1109/TIP.2012.218809>
10. Haralick, R.M., Shanmugam, K., Dinstein, I.H.: Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **6**, 610–621 (1973). <https://doi.org/10.1109/TSMC.1973.4309314>
11. Clausi, D.A.: An analysis of co-occurrence texture statistics as a function of grey level quantization. *Can. J. Remote. Sens.* **28**(1), 45–62 (2002). <https://doi.org/10.5589/m02-004>
12. Hu, R., Barnard, M., Collomosse, J.: Gradient field descriptor for sketch based retrieval and localization. In: 2010 IEEE International Conference on Image Processing, Hong Kong, China, 26–29 Sept, pp. 1025–1028 (2010). <https://doi.org/10.1109/ICIP.2010.5649331>
13. Hu, R.X., Jia, W., Ling, H., Zhao, Y., Gui, J.: Angular pattern and binary angular pattern for shape retrieval. *IEEE Trans. Image Process.* **23**(3), 1118–1127 (2013). <https://doi.org/10.1109/TIP.2013.2286330>
14. Osowski, S.: Fourier and wavelet descriptors for shape recognition using neural networks—a comparative study. *Pattern Recognit.* **35**(9), 1949–1957 (2002). [https://doi.org/10.1016/S0031-3203\(01\)00153-4](https://doi.org/10.1016/S0031-3203(01)00153-4)
15. Mathew, S.P., Balas, V.E., Zachariah, K.P.: A content-based image retrieval system based on convex hull geometry. *Acta Polytechnica Hungarica* **12**(1), 103–116 (2015). <https://doi.org/10.12700/aph.12.1.2015.1.7>
16. Rui, Y., Huang, T.S., Chang, S.F.: Image retrieval: Current techniques, promising directions, and open issues. *J. Vis. Commun. Image Represent.* **10**(1), 39–62 (1999). <https://doi.org/10.1006/jvci.1999.0413>
17. Smeulders, A.W., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(12), 1349–1380 (2000). <https://doi.org/10.1109/34.895972>

18. Kokare, M., Chatterji, B.N., Biswas, P.K.: A survey on current content based image retrieval methods. *IETE J. Res.* **48**(3–4), 261–271 (2002). <https://doi.org/10.1080/03772063.2002.11416285>
19. Kanaparthi, S.K., Raju, U.S.N., Shanmukhi, P., Aneesha, G.K., Rahman, M.E.U.: Image Retrieval by Integrating Global Correlation of Color and Intensity Histograms with Local Texture Features. *Multimedia Tools Appl.* 1–37 (2019). <https://doi.org/10.1007/s11042-019-08029-7>
20. Rafel, C.G., Richard, E.W.: *Digital Image Processing*. 3rd (edn.), Person
21. Kazuhiro, K., Qiu, C.: Content-Based Image Retrieval Using Features in Spatial and Frequency Domains, ICSIIT 2015, CCIS 516, pp. 269–277 (2015). https://doi.org/10.1007/978-3-662-46742-8_25
22. Improving Image Classification with Frequency Domain Layers for Feature Extraction, IEEE International workshop on Machine Learning for Signal Processing, Sept. 25–28, Tokyo, Japan (2017)

Decision Tree-Based Event Detection Framework for UWSN Routing to Optimize Energy Consumption During Transmission



Rakesh Kumar and Diwakar Bhardwaj

Abstract In this article, a novel framework for event detection is proposed to detect event on the basis of data priority. The priority of critical time-bounded events is considered as critical event, and other events consider as non-critical events. The framework uses the decision tree-based classifier to classify the application-specific event in underwater environment. Despite using delay-sensitive routing as suggested by various researchers, the proposal proposed a reserve route-based routing scheme for fast and reliable transmission of critical data. The routing approach proposed in this article provides energy-efficient transmission through suppressing the duplicate packet transmission as delay-sensitive routing, as well as through reducing the transmission of control packets in frequent route discovery. The viability of the proposal verifies through simulation in terms of delay, network throughput, and lifetime of the network.

Keywords Energy-efficient routing · Delay sensitive routing · Underwater wireless sensor network · Decision tree classifier

1 Introduction

Recent developments in underwater wireless communications have empowered the growth of low-cost, power-efficient, multifunctional sensor nodes that are minor in size and communicate untied in short distances [1]. The various application areas demanded multifunctional sensor nodes for sensing real-time data in underwater environment. This sensed data further route to surface stations for processing through the sink nodes usually fitted at the water surface to understand underwater environment for various applications. These applications are categorized as long-term non-critical

R. Kumar (✉) · D. Bhardwaj

Department of Computer Engineering and Applications, GLA University, Mathura, India
e-mail: rakesh.kumar@glau.ac.in

D. Bhardwaj

e-mail: diwakar.bhardwaj@glau.ac.in

applications and short-term critical applications. In a long-term non-time-critical applications, the time of response do not play a vital role. The sensor nodes deployed underwater to monitor a particular area collect the information and relay it to intermediate sensor nodes and finally transmit the information to surface sinks that are equipped with dual modems, i.e., acoustic and radio modem. The surface nodes transmit information to the on-shore command center through radio communication. These applications include environmental monitoring, undersea exploration, Mine reconnaissance, etc. On the other hand, the applications of short term, demanding less time for information transmission is categorized as short-term time-critical application. The applications wherein time of response plays a vital role include disaster prevention, assisted navigation, distributed tactical surveillance, etc.

The fast and reliable time-bound routing for time-bounded critical applications is a challenging task in UWSNs, due to the characteristics of underwater environment. Some researchers suggest delay sensitive routing mechanism for such application areas. These delay-sensitive routing provide reliable transmission, but converse high battery power to route data packets. The cost in terms of battery power is very high in case of delay-sensitive routing, which also leads to imbalanced energy consumption and increases the chance of early die of sensor nodes in the network.

This article proposed a framework to identify the data priority based on the applications and triggered an event, which is classified as critical and non-critical event. This classification is done through the use of decision tree classifier, which classifies the event based on the environmental and non-environmental data of underwater scenario. To implement the proposal, a 3D network architecture is proposed as shown in Fig. 1, in which various sensor nodes are deployed randomly in underwater sensing area uniformly. Multiple sink nodes are fitted on water surface to communicate with underwater sensor nodes as well as surface station. For the same, all sink nodes are equipped with acoustic as well as radio transceiver. The proposed routing is based on the depth-based routing mechanism. All sensor nodes know their depth from the surface with the help of low-cost depth sensor, and routing is done in vertical direction only.

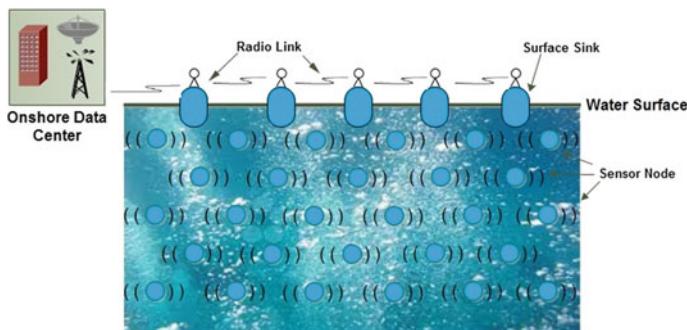


Fig. 1 Underwater network architecture

The preceding segment delivers the detailed framework of the proposal and evaluates the objective of the proposal through simulation.

2 Related Work

This section discussed some literature regarding the work done by various researchers in context to energy-efficient routing as well as delay-sensitive routing.

In Depth-Based Routing Protocol [2], author proposed routing mechanism based on the depth of the sensor nodes. A holding time is calculated on the basis of depth of the node to select next relay node. Another depth-oriented routing named EEDBR [3] proposal also considers residual energy of node to calculate holding time with depth of the node. In [4, 5], authors provided the enhancement of the DBR by adding residual energy in calculating holding time and control the hop count through network partitioning.

Besides the energy-efficient and energy-balanced approach, many researchers suggest delay-sensitive routing mechanism specifically designed for time-critical events. The objective of these approaches is to deliver data quickly. These approaches focus on fast and reliable data communication.

An approach named Delay-sensitive opportunistic routing for underwater sensor networks [6] applies the idea of opportunistic-based routing for maximizing goodput while meeting end-to-end latency requirements. In the article [7], authors extend the Delay-sensitive opportunistic routing for underwater sensor networks by adding energy efficiency throughout the routing process by considering nodes as relay on the basis of remaining energy.

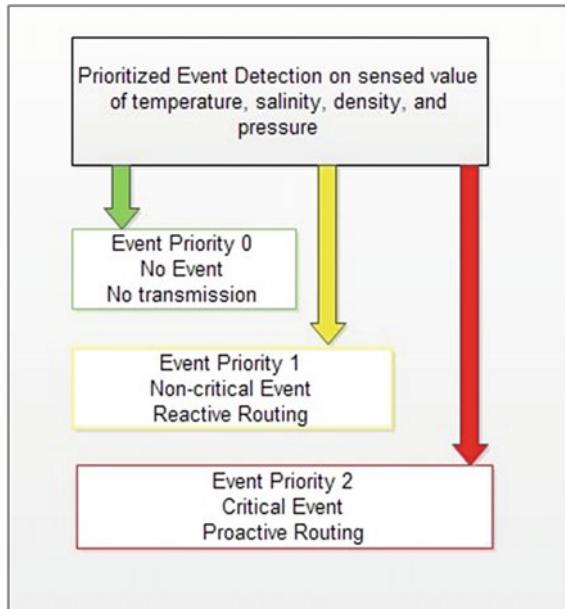
In another work named DRADS [8], authors enhanced the opportunistic-based routing through modifying the metric called the predicted end-to-end latency. The same author proposed the extension of DRADS in article [9] by introducing interference aware and cooperative routing.

The literature discussed above provide the insight that no strategy provides services to all types of applications in underwater environment. Therefore, we try to propose an event-driven framework to classify events and proposed a routing approach for critical time-bounded applications to reduce the delay in transmission.

3 Proposed Framework

The proposed framework for event-driven data communication is shown in Fig. 2. The proposed framework classifies the occurrence of event in three categories based on data priority of environmental and non-environmental data. The proposed framework detect application-specific event based on four underwater parameters; temperature, salinity, density, and pressure.

Fig. 2 Proposed framework for event-driven data communication



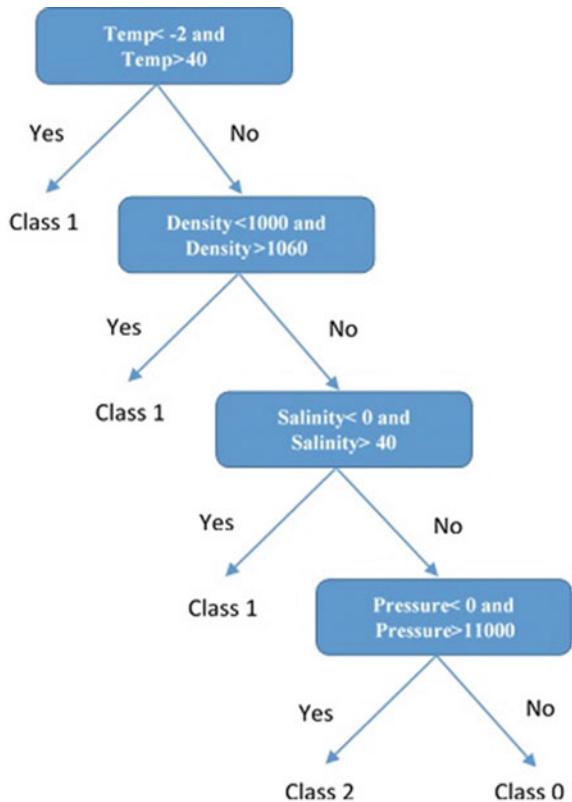
For application-specific event detection, the thresholding of parameter values is considered as per dataset provided by [10] for Atlantic Ocean. The dataset analysis is used for thresholding environmental parameter values and discussed below (Table 1).

Another study [11] carried by of US geological survey says that the effect of inner surface activities or object movement changes the pressure level of water greater than the depth level threshold of pressure by more than 2%. This helps in detecting object movement and other inner surface activity such as earthquake by tracking the change in pressure level on the basis of depth profile of the ocean.

The decision tree classifier is used to classify events in the proposal. It is a straightforward technique to classify data. In our proposal, decision tree is built on the basis of attributes provided by dataset of underwater parameters. Figure 3 shows the decision tree classification model.

Table 1 Ocean parameter range

Parameter	Range	Average value
Temperature	-20 c to 40o c	3.50 c
Salinity	0 g/kg to 42 g/kg	34.9 g/kg
Pressure	0 dbar to 11,000 dbar	1850 dbar
Density	1000 to 1060 kg/m ³	1036 kg/m ³

Fig. 3 Decision Tree Model

The algorithm for reserve route discovery is receiver-initiated approach, in there sink node decides the best reserve route on the basis of residual energy of the participating nodes. Which route has the maximum energy grades is selected as the reserve route for transmission of critical data packets.

Initially, at network setup phase, each sensor node initiates the reserve route setup phase for critical data transmission. The sensor node which initiates the process broadcasts a HELLO packet with source id, slice id, depth, and residual energy of the sensor node. All intermediate sensor nodes that fall between the source node and sink node participate in the reserve route setup phase. Node receiving the request packet from the previous slice wait for a δt time and then process the following steps and broadcast the request to its neighbors:

- Replace the slice id of the packet with node slice id.
- Replace the depth field of the request packet with its own depth.
- Replace the residual energy field with adding own residual energy with max of received residual energy of multiple request packet with same source id in δt time.
- Append the own source id in route information field of packet with maximum residual energy as per Eq. 2.

P _{type}	S _{id}	D _{sen}	R _{ENs}	R _s
-------------------	-----------------	------------------	------------------	----------------

Fig. 4 Hello Packet Format

$$RE_L = \text{Max}\{\text{RE}_1, \text{RE}_2, \text{RE}_3, \dots, \text{RE}_N\} \quad (1)$$

δt time is considered to suppress the duplicate request from same source node. δt is the waiting time equal to the one propagation delay between neighbors and calculated as follows:

$$t = \frac{\text{CommunicationrangeR}}{\text{SignalSpeedS}} \quad (2)$$

After reaching the request packet at sink node, it also waits for δt time to receive multiple request from same source and decide the route founded on the maximum remaining energy of the route. The sink node then broadcasts the reply with maximum residual energy route information to all immediate neighbors.

The format of the *HELLO* packet is shown in Fig. 4.

In the above hello packet format:

- Ptype is one-bit field that represents the type of packet- 0 represents query packet, and 1 represents reply packet.
- Sid is one-byte field that represent the id of source node.
- Dsen one-byte field represents the depth of the sensor node.
- RENs one-byte field represents the summation residual energy of participating sensor nodes.
- Rs ten-byte field represents the route information at route discovery procedure.

The details of the algorithm for reserve route setup is described below:

Algorithm1: Reserve Route Setup.

At Intermediate Node.

1: Upon receiving hello Packet.

2: if $P_{type} = 0$ AND $SL_{id}(\text{Source}) = SL_{id}(\text{Receiver}) + 1$ AND $D_{rec} < D_{sen}$ then.

3: Wait for δt time.

4: Set $REN_{max} = REN_s$.

5: Loop Until Timer > 0.

6: if $REN_s > REN_{max}$ then.

7: $REN_{max} = REN_s$.

8: end if.

9: end loop.

10: Set $SL_{id} \leftarrow \text{Slice Id}(\text{Receiver})$.

11: $D_{sen} \leftarrow \text{Depth of receiver node}$.

12: $R_s \leftarrow R_s + \text{Receiver Id}.$
 13: $\text{REN}_s \leftarrow \text{REN}_{\max} + \text{Residual Energy of Receiver Node}.$
 14: *Goto step 6.*
 15: **else if** $P_{\text{type}} = 1$ **then.**
 16: Extract byte from R_s in $\text{Relay}_{\text{next}}.$
 17: Extract second last byte from R_s in $\text{Relay}.$
 18: Extract depth in $D_{\text{sen}}.$
 19: **if** $D_{\text{sen}} = \text{Depth of receiver node AND } \text{Relay}_{\text{next}} = \text{Sink ID}$ **then.**
 20: $\text{ReserveRoute} \leftarrow \text{Relay}_{\text{next}}.$
 21: Truncate byte from $R_s.$
 22: *Goto step 6.*
 23: **else if** $\text{Relay} = \text{Node ID}$ **then.**
 24: $\text{ReserveRoute} \leftarrow \text{Relay}_{\text{next}}.$
 25: Truncate byte from $R_s.$
 26: *Goto step 6.*
 27: **else.**
 28: drop the hello packet.
 29: **end if.**
 30: **else.**
 31: drop the hello packet.
 32: **end if.**
At Sink Node.
 33: Upon receiving hello Packet.
 34: **if** $P_{\text{type}} = 0$ **then.**
 35: Start Timer for δt time.
 36: Set $\text{REN}_{\max} = \text{REN}_s.$
 37: **Loop Until** $\text{Timer} > 0.$
 38: **if** $\text{REN}_s > \text{REN}_{\max}$ **then.**
 39: $\text{REN}_{\max} = \text{REN}_s.$
 40: **end if.**
 41: **end loop.**
 42: Set $P_{\text{type}} = 1.$
 43: Set $R_s = R_s + \text{Sink Id}.$
 44: Broadcast hello packet.
 45: **end if.**

After executing the proposed algorithm, each sensor node on the network reserves the next relay node entry in their reserve route entry field, and utilizes the route when encountered critical event sensed directly or received data packet with event priority 2 from the previous node. Data forwarding takes place simply by reading the event class from data packet itself. If the event class is 2, then check the reserve route entry from the sensor node and forward the data packets to the node address stored in reserve route entry field.

4 Experimental Results

In this section, the performance of the proposal termed as Prioritized Event-Based Routing (PEBR) is evaluated in terms of delay, network throughput, and lifetime of the network against three well-known energy-efficient routing protocols named DBR, EEDBR, and EEBET. In the simulation, environment deployment of the sensor nodes is in uniform manner, and sink nodes are deployed on water surface randomly. Simulation may take place with varying network radius ranging from 1 to 5 km. Other simulation parameters are presented in the following Table 2.

The evaluation of network lifetime of the proposed PEBR against DBR, EEDBR, EEBET is depicted in Fig. 5. The Proposed PEBR has longer lifetime of network compared to other techniques due to balanced energy consumption. The proposal suppresses the control packet flow through reserve route entry, which leads to energy efficiency and prolonging the network lifetime. On the other hand, EEBET uses two hop communication frequently, which leads to more energy consumption reduces the network lifetime. In the other two techniques DBR and EEDBR, routing loops initiate packet duplication, which overall consumes more energy. Hence, network lifetime is shorter.

Table 2 Simulation Parameters

Parameters	Values
Network Size	Ranging from 1 to 5 km
Nodes	500
Primary Energy	150 J
Message Generation Rate	1 msg per 10 rounds

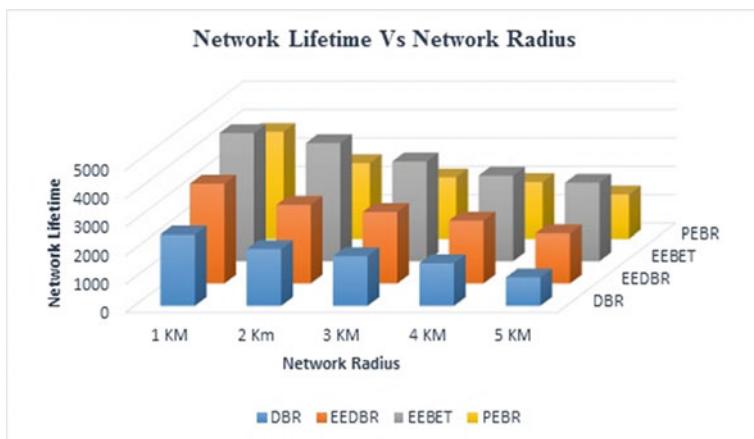


Fig. 5 Network lifetime versus varying Radii

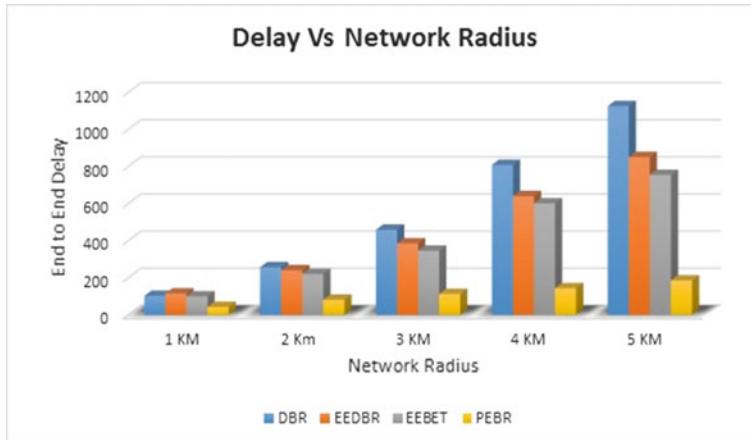


Fig. 6 Delay versus varying radii

The end-to-end delay is shown in Fig. 6 between PEBR, DBR, EEDBR, and EEBET. The proposal encounters a very shorter delay in comparison to other techniques. This is due to availability of reserve route, which reduces the latency to find the next relay node and provide fast and reliable transmission. The frequency of time-bound critical event occurrence is very less, therefore reserve route discovery process initiates less frequently, and increases the chance to find the route in reserve route entry data structure. Hence, overall reduces the delay of the network. On the other hand, the next relay node discovery is done at each node in other techniques, which leads latency to find the relay node and increase the delay of the network.

The throughput analysis for DBR, EEDBR, EEBET, and PEBR is represented in Fig. 7. The throughput is analyzed in terms of packet delivery ratio. High ratio depicted as high throughput. The delivery ratio is very high in case of PEBR for diversified network radius. The PEBR reduces the energy depletion at sensor node by reducing the flow of control packet and frequent route discovery.

The PEBR also transmits data packets in a single hop, optimizes the energy intake, reduces the chance of node failure due to lack of energy, also reduces the chance of dropping data packets, and increases the throughput. EEBET also performs better than the other two techniques, due to use of energy balancing and network partitioning. DBR and EEDBR have less throughput due to packet duplication and routing loops.

The simulation result discussed above verifies that the PEBR is performed well in terms of all parameters discussed. PEBR performs well in diversified networks.

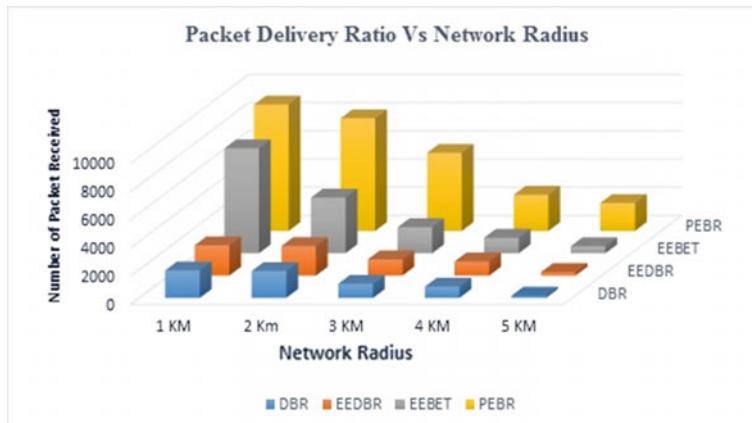


Fig. 7 Network throughput versus varying Radii

5 Conclusions

In this article, we have proposed a novel framework to detect event as short-term critical time-bounded, and long-term non-critical events. The decision tree classifier is used to classify the events. Based on the event classification, a routing approach is developed and analyzed to route data packets for critical time-bounded events in fast and reliable manner. The proposed routing scheme is utilizing the reserve route entry to route the data packets toward the destination sink in place of delay-sensitive routing. The revaluation and analysis of simulation results verify that the proposal meets all the objectives of the research and provides reliable and fast data delivery for critical time-bounded applications in energy-efficient manner.

References

1. Akyildiz, I.F., Pompili, D., Melodia, T.: Underwater acoustic sensor networks: research challenges. *Ad Hoc Netw.* **3**(3), 257–279 (2005)
2. Yan, H., Shi, Z.J., Cui, J.H.: DBR: depth-based routing for underwater sensor networks. In: International Conference on Research in Networking, pp. 72–86. Springer, Berlin, Heidelberg (2008)
3. Wahid, A., Lee, S., Jeong, H. J., & Kim, D. (2011). Eedbr: energy-efficient depth-based routing protocol for underwater wireless sensor networks. In: International Conference on Advanced Computer Science and Information Technology, pp. 223–234. Springer, Berlin, Heidelberg
4. Kumar, R., Mishra, M.K.: Improved the network lifetime through energy balancing in depth based routing protocol for underwater sensor network. In: Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT), pp. 26–27 (2018)
5. Kumar, R., Bhardwaj, D., Mishra, M.K.: Enhance the lifespan of underwater sensor network through energy efficient hybrid data communication scheme. In: 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC),

- pp. 355–359. IEEE (2020)
- 6. Kumar, R., Bhardwaj, D., Mishra, M.K.: EBH-DBR: energy-balanced hybrid depth-based routing protocol for underwater wireless sensor networks. *Mod. Phys. Lett. B* **21**50061 (2020)
 - 7. Hsu, C.C., Liu, H.H., Gómez, J.L.G., Chou, C.F.: Delay-sensitive opportunistic routing for underwater sensor networks. *IEEE Sens. J.* **15**(11), 6584–6591 (2015)
 - 8. Javaid, N., Jafri, M. R., Ahmed, S., Jamil, M., Khan, Z. A., Qasim, U., Al-Saleh, S. S.: Delay-sensitive routing schemes for underwater acoustic sensor networks. *Int. J. Distrib. Sens. Netw.* **11**(3), 532676 (2015)
 - 9. Shakeel, U., Jan, N., Qasim, U., Khan, Z. A., Javaid, N.: DRADS: Depth and reliability aware delay sensitive routing protocol for underwater WSNs. In: 2016 10th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), pp. 78–83. IEEE (2016)
 - 10. Javaid, N., Shakeel, U., Ahmad, A., Alrajeh, N., Khan, Z.A., Guizani, N.: DRADS: depth and reliability aware delay sensitive cooperative routing for underwater wireless sensor networks. *Wirel. Netw.* **25**(2), 777–789 (2019)
 - 11. Montgomery, E.T., Martini, M.A., Lightsom, F.L., Butman, B.: Documentation of the US geological survey oceanographic time-series measurement database (No. 2007–1194). *Geol. Surv. (US)* (2008)
 - 12. Jin, Z., Ma, Y., Su, Y., Li, S., Fu, X.: A Q-learning-based delay-aware routing algorithm to extend the lifetime of underwater sensor networks. *Sensors* **17**(7), 1660 (2017)
 - 13. Puyate, Y.T., Rim-Rukeh, A.: Variability with depth of some physico-chemical and biological parameters of Atlantic Ocean water in part of the coastal area of Nigeria. *J. Appl. Sci. Environ. Manag.* **12**(1) (2008)

Recognizing Child Unsafe Apps Through User Reviews on the Google Play Store



Ashwini Dalvi^{ID}, Irfan Siddavatam, Viraj Thakkar, Aditya Vedpathak, and Abhishek Patel

Abstract Google Play Store serves as a platform to host, download, and review android applications. Many researchers have explored the user review section and worked on approaches and solutions that would prove a more effective pipeline to enable developer feedback on application issues and praised features proving the section's abundance of information. This work uses this same data to attempt a novel use case of determining child unsafe apps on Google Play Store. User reviews are collected using a crawler and categorized for selected keywords relating to child, media, and India. Since Google Play Store does not provide a definitive number of downloads, this work attempts to mitigate this challenge by instead calculating the user density for an application. The user density helps establish the engagement users have with an application and is calculated by the difference in the timestamps of the most and least recent reviews divided by the sum of total reviews and its upvotes for an application. 60,620 reviews from 1,600 applications were extracted to validate the proposed concept. This concept has proved effective in recognizing applications that present child unsafe content while also offering a novel concept of calculating user density.

Keywords User review · Google play store · Content classification

1 Introduction

The proliferation of smartphone technology has made it easy to access different categories of mobile applications ranging from entertainment to education. Further, the emergence of application stores like Google Play Store and Apple AppStore facilitate the spread of a variety of mobile applications. Developers upload their mobile applications on the store, and interested users download these applications, with some writing reviews for the same. Recently, researchers have started exploring the facet of app store mining for a variety of purposes. The user reviews are often

A. Dalvi (✉) · I. Siddavatam · V. Thakkar · A. Vedpathak · A. Patel
K J Somaiya College of Engineering, Vidyavihar, Mumbai 400077, India
e-mail: ashwinidalvi@somaiya.edu

studied to offer constructive feedback to application developers about what the main points of applications are and which features are most appreciated among users [1].

From its establishment in 2008, Google Play Store is the most popular platform for android developers to host their applications. The structure of Google Play Store is such that different categories of data can be retrieved from mining it. For example, the store asks developers to give an app description. Referring to the app description, researchers have attempted to understand whether the respective app is secure or violates security policies. Google Play Store accepts app name, app description (maximum 4000 characters), application type, category, organization type, and other related information. Along with this textual content, the store can also store graphics related to features of the application [2]. With all this information available, it is no wonder that researchers mine the content available on these stores.

The information-heavy Google Play Store page has motivated the present work to explore the possible use cases that can be addressed with play store content. The primary investigation of the Google Play Store page led authors to pay attention to content quality ratings and reviews. Further exploration of content reviews offered a unique use case of the identification of apps which are not safe for children.

The structure of the Google Play Store page recommends developers to mention the age group for which the hosted app is suitable. But there are reported cases where an innocuous-looking app is playing a role in the creation and distribution of child abusive material, typically, chatting and gaming applications. Though Google Play Store has its policies to restrict the spread and reach of unsafe applications, there are thousands of such applications present on the play store. Thus, there is a need to acknowledge that developers with malicious intent have successfully circumvented Google Play Store policies. The present work attempts to bring attention that the information mining from content reviews could prove helpful to identify child unsafe applications, and the results obtained strengthen the adopted approach.

The following sections of the paper are as follows: Sect. 2 discussed the research approaches attempted and suggested with reference to user reviews present on app stores. Section 3 covers in detail the methodology exercised with the inclusion of data collection and result presentation, and further the work is summarized in Conclusion.

2 Literature Review

It is established through the literature that user's reviews about any application on the play store offer myriad information. The researchers are curious about comprehending app stores not only for mining the content but also for encouraging repositories of apps from the perspective of software engineering. The work [1] conducted a survey on app stores for comprehending research opportunities by exploring technical and non-technical attributes of an app. The work presented a timeline that covered key research topics regarding app store analysis like time to deploy apps on the store, feature equivalent among apps, etc.

Given the popularity of Google Play Store, it is researched extensively by researchers to address its limitations and challenges. In work [3], the authors deployed a crawler named PlayDrone to crawl the store pages. PlayDrone was successful to circumvent Google's anti-crawling mechanism and evaluated the play store against characteristics of apps hosted, the impact of time over content evolution of hosted apps, code decompilation of hosted apps, and content plagiarism among the apps. The work concluded that over 25% of hosted apps suffered plagiarism. An alarming number of apps lacked effective secure coding practice and could lead to unauthorized access to users who are not legitimate.

With the literature review, it is clear that user reviews on the play store have been refereed for various use cases such as finding popular features of apps and fixing bugs reported in reviews. In continuation, work [4] discussed the usefulness of reviews to estimate user experience in terms of accessibility though the conclusion of work is not encouraging for hypotheses under investigation. The work concluded that user reviews were not reflective of accessibility issues related to the app. The users offered higher rating irrespective of reporting issues about app accessibility.

In work [5], WisCom framework proposed to identify inconsistencies in reviews and estimate reasons why an app is not liked by users. Also, the framework offered a dashboard to reflect the evolution of reviews over the period. Descriptive statistics were run on 13,286,706 user reviews. The framework also included Text-Rating-Inconsistency feature to address discrepancies between content reviews and ratings.

The work [6] examined 5,06,259 reviews of 148 apps that sprawled over 14 categories with supervised learning algorithms like Naive Bayes, Logistic regression, and Random forest, concluding that logistic regression offered better accuracy. In work [7], the framework IDentify Emerging App (IDEA) was proposed to point out emerging threats through reviews removing noisy and non-informative words from reviews.

Though the reviews seem more appealing to extract the frequent features mentioned in the review, the work [8] extracted infrequent features from reviews. The proposed approach derived features from reviews using a text processing approach along with feature extraction from the app description mentioned on the play store page. Further, the method looked for similarity measures attributed to a single term, synonyms, and the sentence measure similarity approach. The work concluded that the proposed approach was more efficient to extract irrelevant features from reviews. The results were put up by referring to 36 apps (along with app description pages) and in total 1327 reviews. In work [9], the reviews were examined to be labeled according to issues written in it. The authors discussed proof of concept analytics to address research questions:

1. Whether one user review involved more than one issues?
2. Is it possible to label issues rose in reviews?
3. In what way associated stakeholders find the proposed approach beneficial?

The work confirmed the unstructured noisy nature of user reviews which challenges the multi-labeled approach adopted by authors. Still, the work concluded satisfactorily that it is possible to label issues related to apps referring to user reviews.

The work [10] mentioned user reviews as ‘user-facing crowd-sourced indicator of app quality.’ The work attempted to analyze reviews to investigate the position and negative sentiment expressed toward an app. The other two research questions undertaken were as follows: first proportionate the relationship between the usage of English vocabulary used in reviews and average vocabulary used in speaking English, and second was the relationship between review text and ratings given to apps. The work concluded that the usage of vocabulary for negative sentiment is more than positive sentiments. Also, reviews reflected 37% of vocabulary from England dictionary, and reviews were more expressed through sentiment expressing words.

An attempt is made in work [11] to identify the non-functional requirements of an app from the reviews posted by users. Typically, work focused on non-functional requirements like reliability, performance, portability, and usability. The text feature classification algorithm like Bag of Words (BoW), TF-IDF (Term Frequency-Inverse Document Frequency) along with Chi-Square statistical techniques were applied along with supervised learning algorithms on collected data. The work further proposed a variant of the BoW algorithm titled ‘AUR-BoW’.

The Google app store is required to ensure the quality of hosted apps in terms of security. The play store security policies and anti-malware solution tend to ensure whether an app is legitimate or has security concerns. But this approach has a dependency on app features and execution threads. Thus, the other informal approach of analysis reviews the app to confirm whether it has security-related concerns. In the work [12], CIAA-RepDroid proposed sentiment analysis on reviews using a Naive Bayes classifier. The work is carried out on 13 applications. There was a consideration of 1,050 reviews on security and functionality issues from 7,835,322 reviews. The work documented 23% of examined applications to be having security concerns mentioned in the reviews. But the work has a limitation on identifying fake reviews.

The work [13] addressed one of the types of use cases to be attempted through learning user reviews. The authors proposed to examine whether the approach of gamification in the mHealth app could garnish positive response from the user. The inclination of users toward mHealth apps due to gamification has proven to be improved in the discussed study. The work concluded findings based on a sample study of 1000 apps out of which 772 apps had some types of gaming mechanism.

The literature survey on app review analysis concludes that the app reviews are analyzed for the use case of identifying functional requirement or detecting bugs/issues from reviews. In a few cases, reviews are examined to establish security concerns associated with an app or the popularity of apps due to certain features. Thus, this makes the proposed work one of the kinds of attempts to examine user reviews to identify child unsafe content on Google Play Store. The following section discusses the methodology opted for the mentioned use case.

3 Methodology

The present work attempted the unique approach to recognize the child unsafe apps on the play store. The following discussion covers the methodology to collect user reviews from the play store and keywords selected for the categorization of reviews. The result of this work is included further in the discussion. The pseudo-code (see Algorithm 1) explains in brief the procedure carried out. With the pseudo-code in mind, the code for the module has been developed and implemented.

```

Algorithm 1
start
define keywords and words under each keyword_group
for each application:
    Extract application data and store in a csv.
    Extract reviews and store in another csv with app name.
    Check reviews for all the keywords specified
    if keyword present in review:
        Increase keyword_group value by 1
        Tag application with keyword
    end if
    Calculate user density (see eq. 3)
end for
end

```

3.1 Data Collection and Processing Phase

A crawler was created to crawl the Google Play Store web page for each app, find links for similar apps, and visit them. While accessing respective web pages, the tool obtains the App Name, Package Name, Developer Name (if available), App Description, and 40 most recent Reviews which include User (providing the review), Text of the actual review, number of Upvotes, and date of posting the review.

This data was collected through a set of pre-defined keywords. The keywords were selected to categorize a review if it mentions content related to Child, Media such as pics or videos, and Indian Content or Origin. The keywords like ‘Child,’ ‘CP,’ ‘Young,’ ‘Girl,’ and ‘Teen’ were used to validate if the App contains data related to children whereas keywords like ‘Porn,’ ‘Videos,’ ‘Pics,’ and ‘Nude’ were referred to verify if the App content may contain media. Apps containing media content can become a tool for creating and sharing child abusive content. Not every app with media content may host such content but these apps can be abused by pedophiles. All the media-related keywords were used to filter the text in each review.

The Child Keyword filter fetched apps that had content meant for children. Thus, fetched data also contained apps like Disney+ and Child Care Apps along with some browsers. The combined filters of Child Keywords and Media Keywords have fetched a considerable amount of Chat Apps followed by Social Media Apps.

Further, to confirm the reference of Indian user or Indian content for the App, keywords like ‘Indian,’ ‘Ganda,’ ‘Mera,’ ‘Hai,’ ‘Desi,’ and ‘Hota’ were used.

The proposed method aims to extract bulk data and observe patterns regarding the App content specifically related to Children and Media. The data for 1,600 apps was extracted along with 60,620 reviews.

Following are a few sample reviews extracted:

“It’s a fun sexy game I guess. I wasn’t looking for sexy video game it’s more like a hot or not app it’s light and a fun sexy game with sweet messages to make you feel good. Saw a hot girl video talking about it from someone in India and thought it was cool. Would like some more pictures”

“PLEASE READ ALL REVIEWS BEFORE UPLOADING!! **This app needs reporting to Google.** Download at your own risk!! As there are children on this ADULT site. The most visually complicated app I have ever known. If you like easy, clean, simple to use controls this site is not for you. If you like constant flashing content, loads of crass childish filters and men pretending to be women. Then this app is for you. I deleted /uninstalled.”-D (A google user)

“This was horrible. I got this app thinking, “I could vent to strangers, and they don’t know me so they can’t talk about it to people I know.” Then people act like total creeps. I can’t believe people are like this. They ask for things they really shouldn’t. There was even a few pedophiles. One guy actually said, “I am looking for a young female.” -Dia Moodley

The crawled data was collected in a .csv file to categorize reviews in child, media, and India categories. The categorization is accomplished with the help of the K-means unsupervised learning algorithm. The unsupervised algorithm must attempt the recognition of optimal clusters for the given data set. The Elbow method is preferred to find suitable clusters. The K-means algorithm divided the collected reviews data into 3 clusters decided with reference to the categories of keywords.

3.2 Results

From the data obtained, child keywords were used to cluster apps to narrow down the search for Child Abusive Apps. Higher amount of Child Keywords present can indicate that the App is either meant for consumption by children or the content is related to minors. Besides a few Games and Educative Apps, Online Chat Apps contribute majorly to child-related content (see Fig. 1). If the app does not properly maintain the child restrictions, it can become a major platform for grooming victims.

From a global perspective, the trend of Chat Apps and Dating Apps being prone to child abusive practices continues. For an app with a high number of Child Keywords and Media Keywords, these activities are more prone to occur.

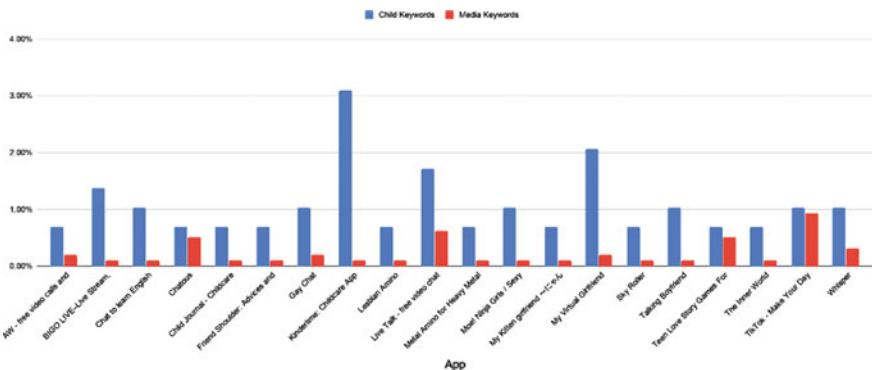


Fig. 1 The chart depicts the apps containing the keywords—child and media. The respective percentages have been determined based on the count of feedback with the keywords

Further extending, a comparison of the reviews with all the three keyword categories—Child, Media, and India—is performed, which results in only a few Apps standing out (see Fig. 2). On a deeper look into some applications on the list, comments that suggested the presence of pornographic material were identified, proving the applications to be potentially harmful. While the figure only gives a rough estimation using the data from the reviews, these applications warrant a second look by the policy-makers.

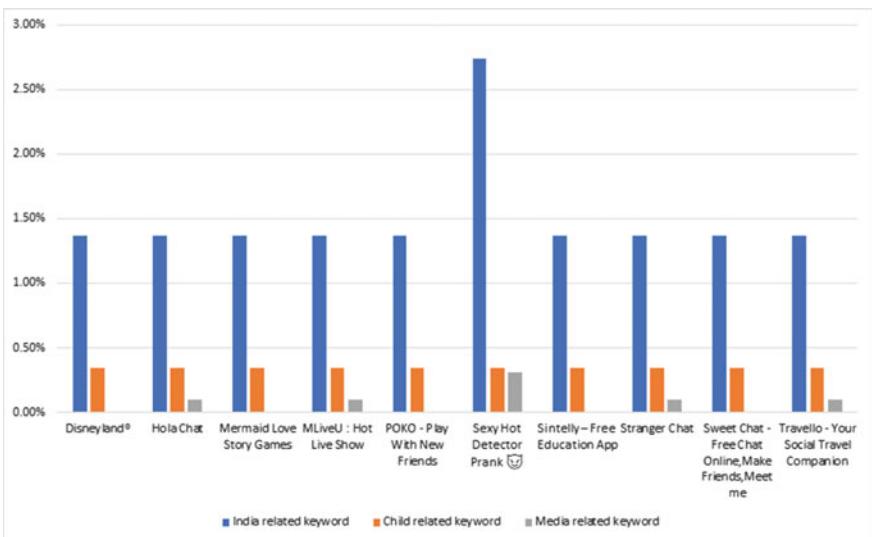


Fig. 2 Some applications that exhibit the keywords—Child, Media, and Indian. The respective percentages have been determined based on the count of feedback with the keywords

3.3 Calculations of User Density

Google Play Store does not provide a clear number of downloads or the user base. To address this challenge, the present work attempted to calculate user density for a respective app.

The calculation of user density was done by comparing the difference in the timestamps of the most recent and the least recent reviews that were scraped, and this was divided by the sum of total reviews and the review upvotes. This aggregated data was normalized with respect to all other apps. The relative score is then generated by subtracting the minimum value among all the apps that were scraped.

The user density helped in establishing user engagement with apps. Reference to the least recent review helped to underline the fact that whether an app is active or in a dormant stage. If the review timestamp reflects the earlier date of a year or months in comparison with the current date, then it is safe to assume that the app is in a dormant stage. This is one of the hypotheses assumed in the undertaken work.

$$\text{TimeGap}_{\text{app}} = (\max(\text{date}_{\text{app}}) - \min(\text{date}_{\text{app}})) / \left(\sum \text{Comments}_{\text{app}} + \sum \text{FeedbackUpdates}_{\text{app}} \right) \quad (1)$$

$$\text{UserDensity}_{\text{app}} = \left(\frac{\text{TimeGap}_{\text{app}}}{\sum \text{TimeGap}} \right) \times 100 \quad (2)$$

$$\text{RelativeUserDensity}_{\text{app}} = \min(\text{UserDensity}) - \text{UserDensity}_{\text{app}} \quad (3)$$

This relative density distribution shows that apps vulnerable to child abusive content are not far behind popular apps such as WhatsApp messenger and TikTok, making their reach dangerous (see Fig. 3). These apps have been chosen as they have text matching the three keyword categories of Indian, Child, and Media.

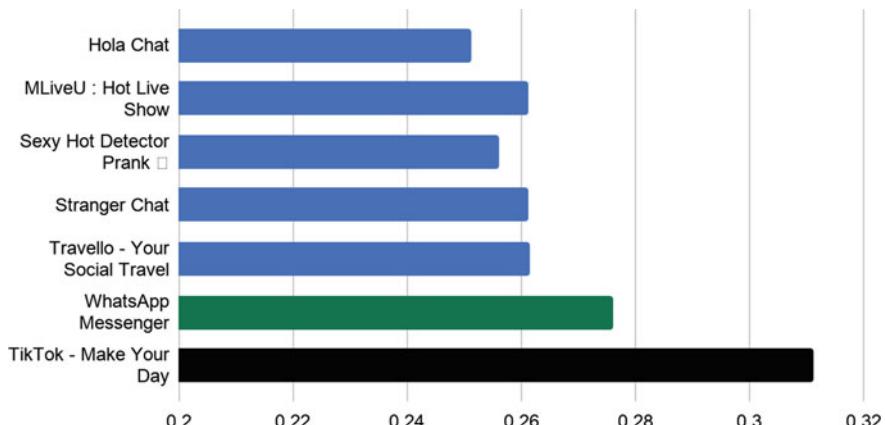


Fig. 3 The figure demonstrates the relative user density of applications in comparison with WhatsApp and TikTok, two well-known and highly downloaded applications

The presented result strengthens the fact that the user reviews on Google Play Store could be served as a source up to a certain extent to identify child unsafe apps.

4 Conclusion

The reviews on Google Play Store have drawn the attention of researchers and proven useful in various contexts. This work attempts to present a unique use case of recognizing child unsafe apps by categorizing user reviews of the respective app on the Google Play Store page. The reviews were analyzed for keywords related to child, media, and India. The presence of all keywords in user reviews acknowledges the app is likely vulnerable to child abuse material. Apps containing media content can become a tool for creating and sharing child abusive content. Not all apps with media content may host such content, but these apps can be abused by pedophiles. Further, the work attempted to find out whether the app is in the context of India. The context to India is evaluated in two ways—one is whether an Indian user has put up a review or there is a mention of an Indian child. The work also tried to present user engagement of apps with the unique concept of user density. The presented results satisfactorily validated that the proof of concept analysis of user reviews is useful in recognizing child unsafe apps on Google Play Store.

References

1. Martin, William., Sarro, Federica., Jia, Yue., Zhang, Yuanyuan, Harman, Mark: A survey of app store analysis for software engineering. *IEEE Trans. Softw. Eng.* **43**(9), 817–847 (2016)
2. Create and Set Up Your App—Play Console Help (2020). Accessed 24 Dec 2020. <https://support.google.com/googleplay/android-developer/answer/9859152>
3. Viennot, N., Edward, G., Jason, N.: A measurement study of google play. In: The 2014 ACM International Conference on Measurement and Modeling of Computer Systems, pp. 221–233 (2014)
4. Eler, M.M., Leandro, O., Alberto, D.A.O.: Do Android app users care about accessibility? an analysis of user reviews on the Google play store. In: Proceedings of the 18th Brazilian Symposium on Human Factors in Computing Systems, pp. 1–11 (2019)
5. Fu, B., Jialiu, L., Lei, L., Christos, F., Jason, H., Norman, S.: Why people hate your app: making sense of user feedback in a mobile app store. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1276–1284 (2013)
6. Karim, A., Azhari, A., Samir, B.B., Ali, A.Q.: Machine Learning Algorithm’s Measurement and Analytical Visualization of User’s Reviews for Google Play Store (2020)
7. Gao, C., Jichuan, Z., Michael, R.L., Irwin, K.: Online app review analysis for identifying emerging issues. In: Proceedings of the 40th International Conference on Software Engineering, pp. 48–58 (2018)
8. Sutino, Q.L., Siahaan, D.O.: Feature extraction from app reviews in google play store by considering infrequent feature and app description. *J. Phys.: Conf. Ser.* **1230** (1), 012007 (2019). IOP Publishing

9. McIlroy, Stuart., Ali, Nasir., Khalid, Hammad, Hassan, Ahmed E.: Analyzing and automatically labelling the types of user issues that are raised in mobile app reviews. *Empir. Softw. Eng.* **21**(3), 1067–1106 (2016)
10. Hoon, L., Rajesh, V., Jean-Guy, S., Kon, M.: A preliminary analysis of vocabulary in mobile app user reviews. In: Proceedings of the 24th Australian Computer-Human Interaction Conference, pp. 245–248 (2012)
11. Lu, M., Peng, L.: Automatic classification of non-functional requirements from augmented app user reviews. In: Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering, pp. 344–353 (2017)
12. Tchakounté, F., Athanase, E.Y.P., Jean, C.K., Marcellin, A.: CIAA-RepDroid: a fine-grained and probabilistic reputation scheme for android apps based on sentiment analysis of reviews. *Future Internet* **12**(9), 145 (2020)
13. Schmidt-Kraepelin, M., Scott, T., Ali, S.: Investigating the relationship between user ratings and gamification—a review of mHealth apps in the apple app store and google play store. In: Proceedings of the 52nd Hawaii International Conference on System Sciences (2019)

Brain Tumor Classification Using Deep Learning and Big Data Analytics



V. P. Arathi, Gayathri Suresh, T. H. Harikrishna, and A. G. Hari Narayanan

Abstract A brain tumor could be a bunch of cells in your brain that are unusual. The cranium is exceptionally inflexible, which encases the brain. Any development can cause issues inside such a restricted space. There may be malignant or benign brain tumors. Radiologists can assist in tumor diagnostics without invasive measures by developing technology and machine learning. CNN is the foremost common and broadly utilized machine learning algorithm for visual learning and picture acknowledgment errands. Similarly, in our paper, in conjunction with Big Data Analysis and Picture Processing, we present the convolutional neural network (CNN) approach to classify brain MRI pictures into cancerous and non-cancerous. The experimental analysis on a very limited dataset shows our model has high accuracy and has a very low complexity rate. The newly created CNN engineering may be utilized as an imperative decision-support strategy for radiologists in therapeutic diagnostics, with good generalization capabilities and good execution speed.

Keywords Deep learning · K-means clustering · Brain tumor classification · Convolutional neural network (CNN) · Big data analysis

1 Introduction

The brain is one of the human body's greatest and most complex organs. The aggregation of unusual cells within the brain could be a brain tumor. The cranium is exceptionally unbending, which encases the brain. Any development can cause issues inside such a constrained space. Brain tumors can be cancerous (malignant) or non-cancerous (benign). They may cause the weight inside the cranium to extend as benign or harmful tumors develop. This could cause harm to the brain, and it may be life-threatening. Tumors of the brain are evaluated as essential or auxiliary. Your brain is where a primary brain tumor begins. Numerous primary brain tumors are generous. Brain cells, the layers covering the brain, which are called meninges, nerve

V. P. Arathi (✉) · G. Suresh · T. H. Harikrishna · A. G. H. Narayanan

Department of Computer Science and IT, Amrita School of Arts and Sciences, Kochi, Kerala, India

cells, and organs, may develop out of them. When cancer cells relocate to the brain from another organ, such as the lung or breast, an auxiliary brain tumor, moreover alluded to as a metastatic brain tumor, happens. There are often secondary brain tumors that are malignant. Tumors that are benign should not move from one part of the body to another.

A brain tumor diagnosis starts with a physical exam and a look at the patient's medical history. In diagnosing brain tumors, medical imaging plays a key role. Hence, medical imaging is performed after the physical exam. Obtrusive and regularly unsafe early imaging procedures have been deserted in support of non-invasive, high-resolution methods, particularly magnetic resonance imaging (MRI) and computed tomography (CT), although the reference standard is usually used for MRI. A special dye is used to detect tumors in order to take an MRI of the head. Like CT, a two-dimensional representation of a thin "slice" of the body is traditionally created by MRI and is therefore considered a tomographic imaging technique. In the form of 3D blocks, modern MRI instruments can generate images that can be considered a generalization of the single-slice, tomographic, principle. MRI does not require the use of ionizing radiation, unlike CT, and is thus not related to the same health hazards.

In recent years, major developments have been made in medical science such as the Medical Image Processing technique due to technology and machine learning, which lets doctors detect disease early and quickly, before it becomes complicated and time-consuming. Computer-aided technology is therefore much needed to overcome such constraints because the Medical Field needs accurate and efficient techniques to diagnose life-threatening diseases such as cancer, which is the world's leading cause of mortality for patients. Subsequently, in our research, we contribute with Big Data Investigation to the classification of brain MRI into threatening and generous utilizing of the Machine Learning algorithm, particularly Convolutional Neural Network.

2 Literature Review

In [1], a CNN design for cerebrum tumor order was performed utilizing a T1-weighted difference improved MRI picture dataset which contains three tumor types. As information, entire pictures were utilized, so it was not important to play out any pre-processing or division of the tumors. This organization has a generally excellent execution speed of 15 ms for each picture.

In [2], the Picture net database is utilized for classification utilizing CNN. It is one of the pre-prepared models. In this way, the preparation is performed for the final layer. At long final, the Angle decent-based loss work is connected to achieve high accuracy.

In [3], the creator displayed the CNN approach alongside Data Expansion and Picture Processing to arrange cerebrum pictures into damaging and non-dangerous utilizing the transfer learning approach showing low multifaceted nature rate by fulfilling 100% precision. In this framework, the picture edge discovery strategy was

utilized to discover the region of interest in MRI pictures and trimmed them at that point; the data augmentation technique was utilized for expanding the size of the training data. The normal preparing time per epoch is 205 s for this model.

In [4], a Recurrent Neural Network plan was proposed for the acknowledgment of tumor cells with high precision. RNN may be a kind of counterfeit neural organization in that the affiliations between centers structure a facilitated chart along a common course of action. By examining CNN and RNN, RNN gives more precision when differentiated with CNN, however, planning time taken by RNN is higher than CNN.

In [5], the programmed cerebrum tumor characterization calculation utilizing an exceptionally precise calculation is introduced. To begin with, the brain portion is portioned by a thresholding approach taken later by a morphological action. The picture enlargement methods are utilized to create the data falsely by various procedures, for example, pivot, scaling interpretation, flipping, resizing, point of view change, and so on to get a freed of restricted accessibility and inconstancy of the clinical picture dataset. Likewise, this procedure assists with defeating the overfitting issues. The AlexNet is utilized due to the limitation of the cerebrum MRI dataset. The AlexNet is supplanted by the SoftMax layer with generous and harmful pictures. The test shows that this system is superior to the old techniques.

3D magnetic resonance images are implemented in [6] to enhance efficiency and minimize the difficulty in segmenting the MRI data, which can provide better precision in detecting the region of the tumor. Using 3D pictures, subspace clustering techniques were proposed to identify brain tumors. It is possible to conduct a 3D assessment of brain tumor detection using a 3D slicer that reliably finds brain tumors.

In [7], tensor flow was engaged to attain brain tumor division, in which the anaconda structures are utilized to execute high-level numerical capacities. Using the CNN architecture, brain tumor segmentation applies to both local and global characteristics, so it helps to reliably conduct segmentation. By lessening the features within the fully associated layer, the speed is expanded. The reduction of parameters often allows overfitting to be minimized. The result shows that the method implemented helps to detect tumor enhancement and to specify tumor only for the actual tumor area.

3 Proposed System/Materials and Methods

In this, we want to classify a patient's brain MRI scan obtained in the axial plane as to whether a tumor is present or not. We will use Big Data Analytics and Deep Learning to identify the images as tumors or not. A Deep Learning class, often used to analyze visual images, is Convolutional Neural Networks (CNN) or ConvNet. There are numerous systems in Python to apply CNN, for example, TensorFlow, PyTorch, etc. to prepare the model. We're going to train this model using the Keras library with the PyTorch backend.

3.1 Dataset

We will analyze the data from the MRI in the first stage. We have a dataset containing brain MR images along with manual segmentation masks for FLAIR abnormality in this problem. The photographs were taken from TCIA. They correspond to 110 patients included within the lower-grade glioma and usable genomic cluster results. There are two separate files in the data collection that are Yes or No. The files both contain various MRI images of the patients. The Yes folder has brain tumor patients, while the No folder does not have MRI images of brain tumor patients.

3.2 Pre-processing

Similar to most real-world records, medical images suffer from problems that can increase inaccuracy if not treated. Adjustment of contrasts, reduction of noise, elimination of physiological artifacts, and the management of the missing data are some of the reasons for conducting pre-processing steps prior to analysis in order to validate hypotheses about the model. Pre-processing is conducted using median or Gaussian filtering techniques where the image noise is eliminated by default structuring. The picture is segmented after removing the noise using Big Data Analytics clustering algorithms such as subspace clustering and K-means clustering. It is utilized to segment a picture into K clusters. The 3D image was divided into 2D images and clusters with K-means and finally all the subspaces. This algorithm is basically pointed at decreasing Squared Error Function.

3.3 Data Augmentation

In medicine, data augmentation is a very significant factor where there would be several cases of data imbalance. The imbalance of data is where the number of observations per class is not distributed equally. In most cases, the number of sick patients will usually be much smaller than the number of healthy patients. We take a single MRI image in Data Augmentation and perform different kinds of image enhancements such as spinning, mirroring, and flipping to get more pictures. To get roughly equivalent numbers of images for both classes, we can add more augmentation to the class with a smaller number of images. So, after applying Data Augmentation to our dataset, in both classes, we can get almost the same number of images.

3.4 Splitting the Data

In the next step, we divide our data into two sets, namely training and test set. 80 percent of images will be in the training set that our neural network will be using to get trained. The remaining 20% will go to the test set to apply our qualified neural network and identify it to verify the accuracy of our neural network.

3.5 CNN Model

CNN is in deep learning which is mostly applied for image analysis. Compared to other algorithms, CNN uses very little pre-processing implying that the network knows the filters that were hand-engineered in conventional algorithms. The convolutional neural network is made of three layers, namely input, hidden, and output layers. Here, the middle layers are called secret since the activation function and final convolution hide their inputs and outputs. The secret layers contain layers that perform convolutions. It normally involves ReLU as its activation function. This is accompanied by pooling completely linked, and normalization layers.

There are different CNN architectures available such as LeNet, AlexNet, VGG, GoogLeNet, ResNet, U-Net, and more. In our work, we used the U-Net architecture (see Fig. 1). In general, convolutional neural network focuses on classifying images with an image as input and simple tag as output, but medical science demands locating

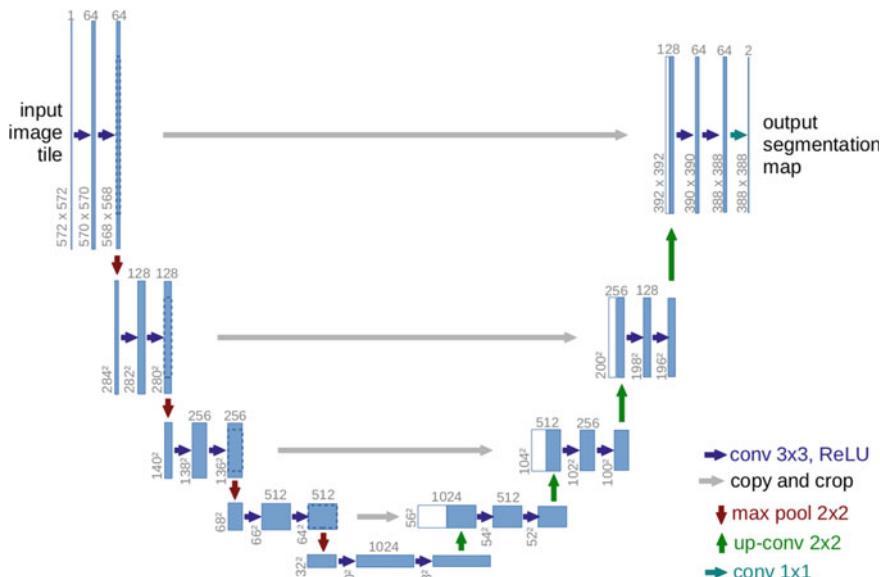


Fig. 1 U-Net Architecture

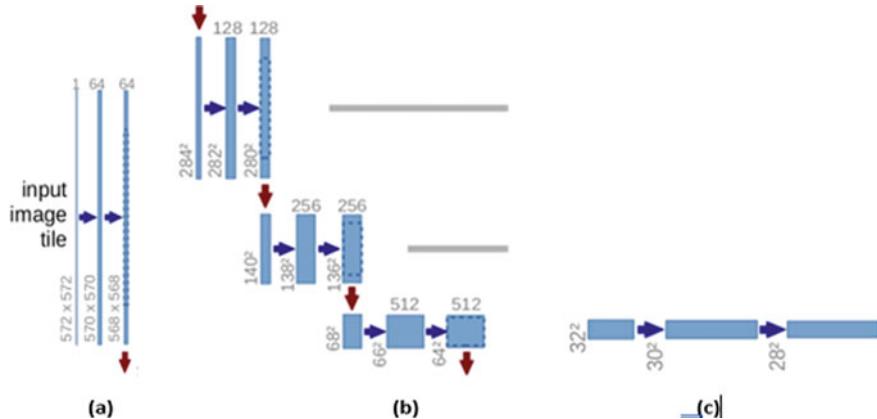


Fig. 2 Contracting path

the region of abnormality along with discerning whether there is a disease or not. U-Net is committed to resolving this issue. The explanation why boundaries can be localized and separated in U-Net is by classifying each pixel, thus the output and input share the same dimension. For instance, the output with size 2×2 will have input size 2×2 .

U-Net has a “U” structure at first sight. The design is symmetrical and has two main parts: the left part is the common convolutional method called the contracting path and the right part is the 2D convolutional layers transposed called the expansive path.

The equation followed by the contracting path:

$$\text{convoltional_layer1} \rightarrow \text{convolutional_layer2} \rightarrow \text{max_pooling}$$

Note that each phase consists of two convolutional layers (see Fig. 2.), and the number of channels varies from 1 to 64, as the depth of the picture would be increased by the convolution process. The red arrow is the maximum pooling mechanism that cuts in half the scope of the image. The mechanism is replicated 3 more times. There are 2 more convolutional layers at the bottom with no max pooling. The image has been resized to $28 \times 28 \times 1024$ at this stage.

In the expansive direction, the image is upgraded to its original size by the formula:

$$\begin{aligned} & \text{convolutional_2D_Transpose} \rightarrow \text{concatenation} \rightarrow \\ & \text{convolutional_layer1} \rightarrow \text{convolutional_layer2} \end{aligned}$$

Transposed convolution is a form of up-sampling that increases the image size. Followed by a convolution operation, it does some padding on the original image. Then the picture is concatenated with the equivalent picture from the opposite contracting direction. This is done to incorporate the info from former layers with

the aim of getting a more accurate estimate. This step is repeated 3 more times, the same as before. The last move is to reshape the picture at the top of the architecture to fulfill our prediction requirements.

By predicting image pixel by pixel, U-Net can do image localization, and it is powerful enough to do good prediction using data augmentation methods based on a small amount of data.

4 Experiment and Result Analysis

To Pre-train the CNN model “train_generator” and “validation_generator” are generated to store the split images of the training set and the test set into two groups (yes and no). At the final stage to train the CNN model, the image data is fitted to the trained neural network. The image data is trained for many epochs. For the Neural Network to be better equipped with the training images, an epoch can be thought of as an iteration in which we feed the training images again and again. For an image, the training set keeps improving with each iteration. This means that it is possible to enhance the Neural Network model by classifying the picture as a tumor or not a tumor. At the end of the last epoch, the trained CNN model has a high validation accuracy. The “accuracy” and “loss” of both “train generator” and “validation generator” were plotted after the training for all the epochs (iterations) (see Fig. 3).

Feature extraction steps are not separately needed in the proposed Big Data Analytics and CNN-based classification. Thus, the complexity and the time of analysis are minimal, and the accuracy is excessive. Many studies use the same database for brain tumor categorization. We selected the paper which had designed neural networks to identify the images to equate our findings with those of previous studies. Table 1 provides a contrast with the currently available approach that uses an engineered neural network and augmented data.

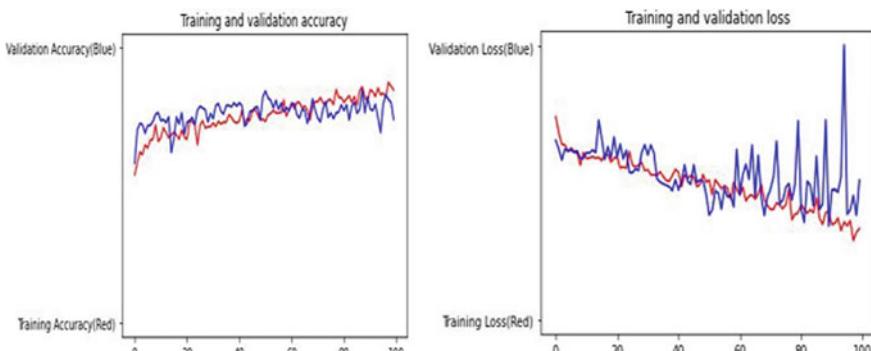


Fig. 3 Training accuracy and loss

Table 1 Comparison table

Method	Precision (%)	Recall (%)	F-measure (%)
Existing method [1]	97.15	97.82	97.47
Proposed method	100	98.76	98.43

5 Conclusion

In this research, we divided brain tissues using MR images of the brain into normal and tumor-infected tissues; to eradicate the impact of unnecessary noise, pre-processing is used. Brain MRI segmentation tends to make detection and classification simpler. Brain MRI segmentation is challenging due to noisy input images. Different segmentation approaches have been suggested, with varying complexities, to address these challenges. In the past few decades, these techniques have resulted in more precise results. To further improve the efficacy of the segmentation process, we used Big Data Analytics clustering algorithms such as subspace clustering and K-means clustering are used. By integrating Big Data Analytics, we have an effective technique for brain tumor classification by proposing a simple CNN network. Our findings suggest that the approach proposed will facilitate to diagnose brain tumors reliably and quickly, along with determining their exact location. Thus, the proposed concept is effective for the detection of brain tumors from MR images.

References

1. Milica, M.B., Marko, C.B.: Classification of brain tumors from MRI images using a convolutional neural network. MDPI (2020)
2. Seetha, J., Selvakumar Raja, S.: Brain tumor classification using convolutional neural networks. Biomed. Pharmacol. J. (2018)
3. Hassan, A.K., Wu, J., Muhammad, M., Muhammad, U.M.: Brain tumor classification in MRI image using convolutional neural network. Math. Biosci. Eng. (2020)
4. Suganthe, R.C., Revathi, G., Monisha, S., Pavithran, R.: Deep learning based brain tumor classification using magnetic resonance imaging. J. Crit. Rev. 7(9) (2020)
5. Sunita, M.K., Sundari, G.: Brain MRI classification using deep learning algorithm. Int. J. Eng. Adv. Technol. 9(3) (2020). ISSN: 2249–8958
6. Padmavathi, V., Sudha, T.: Big data analytics for brain tumour detection using subspace clustering. Asian J. Comput. Sci. Technol. 8(S3), 150–153 (2019). ISSN: 2249–0701
7. Malathi, M., Sinthia, P.: Brain tumour segmentation using convolutional neural network with tensor flow. Asian Pac J. Cancer Prev. <https://doi.org/10.31557/APJCP.2019.20.7.2095>
8. Arun, K.R., Har, P.T.: Image analysis for MRI based brain tumor detection and feature extraction using biologically inspired BWT and SVM. Int. J. Biomed. Imaging (2017)
9. Nagaraj, Y., Jae, Y.C., Bumshik, L.: MRI segmentation and classification of human brain using deep learning for diagnosis of alzheimer's disease: a survey. Biomed. Sens. (2020)
10. Mahboob, Al., Mohd, A.: Segmentation and classification of brain MR images using big data analytics. ResearchGate (2018)

11. Amirhessam, T., Anahid, E., Behshad, M., Amir, H.G., Katja, P., Anke, M.-B.: Big data analytics in medical imaging using deep learning. ResearchGate (2019)
12. Haimiao, Z., Bin, D.: A Review on Deep Learning in Medical Image Reconstruction (2019)
13. Usman, K., Rajpoot, K.: Brain tumor classification from multi-modality MRI using wavelets and machine learning (2017)
14. Parveen: Detection of brain tumor in MRI images, using combination of fuzzy c-means and SVM. Signal Process. Integr. Netw. (2015)
15. Sajjad: Multi-Grade brain tumor classification using deep CNN with extensive data augmentation. J. Comput. Sci. (2018)

Drone Stability Simulation Using ROS and Gazebo



Rajesh Kannan Megalingam, Darla Vineeth Prithvi,
Nimmala Chaitanya Sai Kumar, and Vijay Egumadiri

Abstract In the recent year, Unmanned Aerial Vehicles (UAV) which are generally called Drones or Quadcopters have gained a lot of attention towards researches and companies because of their wide range of capabilities in different sectors such as in military and public. The PID control algorithm which runs in the ROS environment controls the 3-D modelled quadcopter in Gazebo. The main focus is to make the quadcopter stable and to hover at the desired point in Gazebo by tuning PID values in the ROS environment. A plot juggler has been used to obtain the trajectory analysis and to analyse the tuning errors by the trial and error method for each quadcopter motion such as roll, pitch and altitude. Through this framework, certain parameters of the drone are obtained such as settling and rise time with PID tuning.

Keywords PID · ROS · Gazebo · Pitch · Roll · Yaw · Altitude · Rise time · Settling time

1 Introduction

In olden days where transmitting information is quite often a problematic task and also takes a large amount of time to communicate. Today's technology has made communicating things simpler as it takes place in a fraction of seconds. In this jam-packed busy schedule, knowing what's happening around us and how things are going to change is extremely important. In the field of communicating information like disaster and safety management, transport, monitoring traffic updates, aerial photography, agriculture, traffic monitoring, logistics, meteorological purposes and

R. K. Megalingam (✉) · D. V. Prithvi · N. C. S. Kumar
Department of Electronics and Communication Engineering, Amrita Vishwa Vidyapeetham,
Amritapuri, India
e-mail: rajeshm@am.amrita.edu

V. Egumadiri
Department of Electrical Engineering, University of South Florida, Florida, USA
e-mail: egumadiri@usf.edu

in the military sector, an unmanned aerial vehicle (UAV) simply known as Drone (or) Quadrotor has a wide range of applications.

Mainly in the field of survey and rescue, drones have a greater advantage due to their wireless technology. Predominantly, drone stability performs a significant role in real-time video assistance, but in general, due to its unstable nature, execution and testing in the real world, it can lead to loss or damage to equipment and the environment. In this paper, our main goal is to analyse the stability and various errors to stabilize the drone at a destination coordinate autonomously using a Proportional–Integral–Derivative (PID) control system in Gazebo open source, where the position is detected using WhyCon markers by the overhead camera. PID algorithm controls the position feedback error and the drone. Gazebo is a 3D simulator consisting of a physical engine for illumination, inertia, gravity and all types of common sensors, and Robotic Operating System (ROS) serves as the interface for the drone.

2 Problem Statement

Drones have been used in the research field and also in commercial applications. An experienced pilot is needed to control the drone. It is a difficult task to hover a drone steadily at a fixed point by manual operation. This paper presents the simulation and implementation of a UAV in a Gazebo environment using a robotic operating system. The drone simulation in the paper helps to understand the tuning of PID flight controllers with less stability error and the characteristics of drones using plot juggler in ROS.

3 Related Works

In the research work [1], the authors did the stability simulation analysis for a 3-D modelled drone. CATIA software has been used to obtain the parameters such as mass and the moment of Inertia in x, y and z axes. A detailed study about the stability simulation for basic drone motions such as roll, pitch, yaw and altitude has been mentioned with parameters such as settling time and overshoot percentages. Settling time of roll and pitch is 1.419 s, yaw is 2.327 s and altitude hold is 6.339 s.

In the research paper [2], the gradient method has been proposed to tune the proportional–integral–derivative (PID) controller parameters for a quadcopter model named as AR Drone. This paper mentioned experimental results as well as simulations of the autotuning of PID for AR drones in two ways: one is the waypoint navigation and the other is the leader–follower formation control.

The authors in [3] use a PID-based drone available in MATLAB and Simulink for the simulation. The study deals with the basics of drone dynamics and external forces that act on the drone and also develop autonomous navigation and trajectory planning for drones. Newton–Euler equations have been used for the dynamics. The papers

also showed detailed simulations and results to showcase the difference between the PD and PID controllers.

In the research paper [4], the authors present the simulation, modelling and implementation of a quadrotor unmanned aerial vehicle. The study focuses on the quadcopter dynamics and to further develop the autonomous navigation system, autonomous trajectory planning for quadcopter, and efficient control algorithms. The Newton–Euler method has been used for quadcopter dynamics and the quadcopter was developed in MATLAB and Simulink. The author also compared the MATLAB model with the commercially available quadcopters such as the adopter and MultiWii.

The authors in [5] presented an unmanned aerial vehicle (UAV) which can detect humans. The video capture by the UAV will be collected and processed in MATLAB using a human detection algorithm. A Pixhawk flight controller has been used as a quadcopter controller.

In [6], the study focuses on making a quadcopter from scratch. Arduino Mega, Raspberry Pi, GPS module, Flysky CT6B transmitter and FS-R6B receiver are used in the making of the drone and OpenCV with optical flow and Canny Edge Detection algorithms were also used for the location and the size of the objects. The author successfully implemented and tested a quadcopter which is able to fly from one point to another point autonomously and able to detect and award obstacles in its path.

The authors in [7] present a simulation framework for the position control and trajectory tracking of the Gazebo model of the Crazyflie 2.0 quadcopter. The control algorithm which runs in the MATLAB and Simulink® environments controls the position of the quadcopter in the Gazebo simulator, through Robot Operating System (ROS) interfaces. The goal here is to establish the connection between the ROS-enabled Gazebo quadcopter model and the Simulink environment and thereby sending and receiving the control commands between both the environments. The main focus of this paper is to establish a connection between ROS and Gazebo to control the drone.

The research paper [8] presents an area for testing the quadcopter capabilities and also presents a proper environment setup to train the users. The developed arena consists of different setups with certain objects such as rings and poles and with different paths and conditions for the user to gain knowledge on stability and on the performance of the quadcopter. In total, there are five arena setups mentioned. In the results, the average time taken by the drone to complete each arena is also mentioned.

In the research work [9], the authors implemented, simulated and proposed a mathematical model to tune the wheels encoders, PID controllers and inertial measurement unit (IMU) for multi-terrain robots. The authors used a robotic operating system to interface with Gazebo to simulate the multi-terrain robot. The authors also simulated the robot in different conditions with and without PID values to showcase the system stability.

In [10], the authors developed a 3-D model bot in SOLIDWORKS and simulated it in Gazebo of the Robot Operating System (ROS). A gamer's steering wheel is used to control the movement and speed of the robot. The robot was tested in different environments such as rough terrain, asphalt planes and ramps.

4 Drone Modelling

4.1 System Architecture

Figure 1 represents the block diagram of the entire system of this research work. The master controller for the whole system is ROS which subscribes and publishes various nodes and also communicates with the drone and other Gazebo parameters. As the drone changes its momentum to reach the destination and to acquire stabilized state, the overhead camera module detects the WhyCon markers placed above the drone as shown in Fig. 2. According to the sensor, camera and physical parameters of the drone, the correction value is calculated based on PID terms and feedback

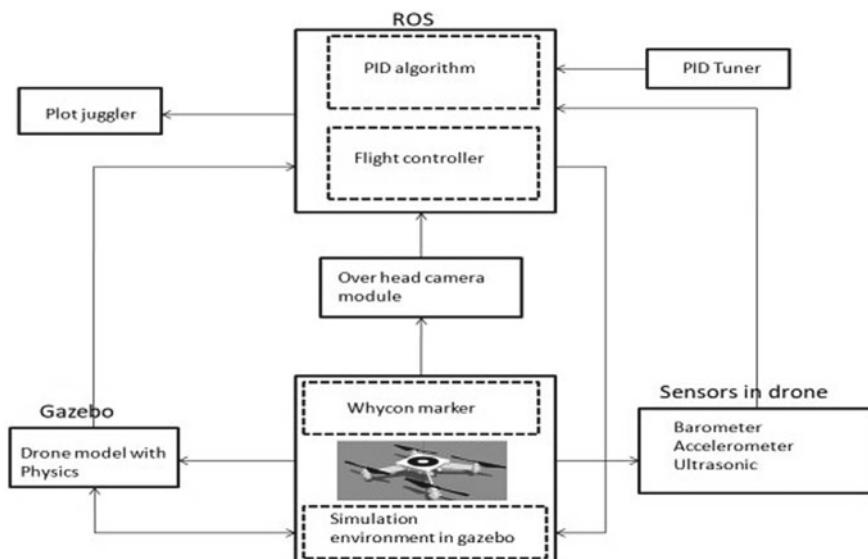
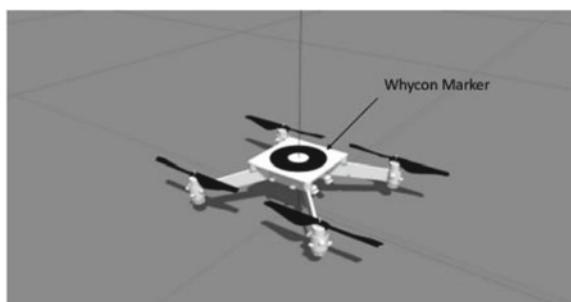


Fig. 1 System overview—block diagram

Fig. 2 Drone in the Gazebo simulation environment



to the system to reduce the error value. PID algorithm and flight controller script communicate with the drone. Using plot juggler, graphs with respective altitude, pitch, roll errors and time are taken. By tuning appropriate values of K_p, K_i and K_d of Roll, Pitch, Yaw and Altitude parameters, the drone can reach the destination point with less settling time and less stability error.

4.2 Drone Motion

Quadrotors have a major asset over other types of unmanned aerial vehicles; they can float without difficulty in confined space within a short period of time, can easily take off and land. They maintain the concept of the thrust and steering function by four propellers attached with independent motors at the respective edges of diagonals from the centre of the drone as shown in Fig. 2.

There are four degrees of freedom in which the drone can acquire stable movement. Rotating the four propellers in the below-mentioned directions results in the following movements:

- (1) **Thrust/Throttle:** By increasing or decreasing the angular velocity of four motors, drone height can be increased or reduced. To attain vertical lift of the drone, each motor spins in the opposite direction of the adjacent motor, as shown in Fig. 3.
- (2) **Pitch:** Changing the angular velocity of front and back motors as shown in Fig. 4 results in the drone movement to move forward or backward. Simply increasing the speed of backside motors as compared to the front side effects the drone to move forward and vice versa.

Fig. 3 Throttle movement

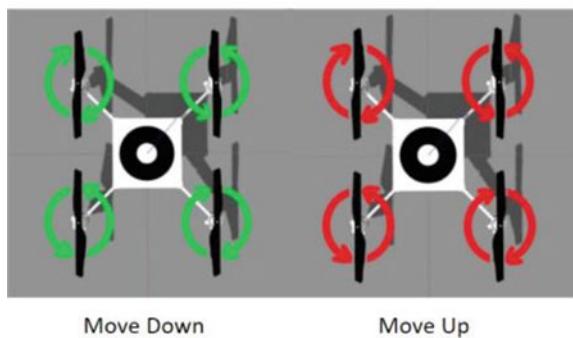
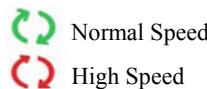
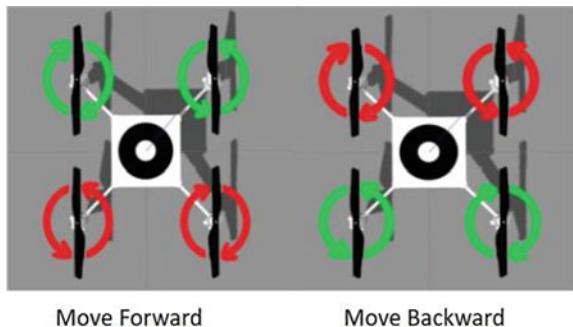
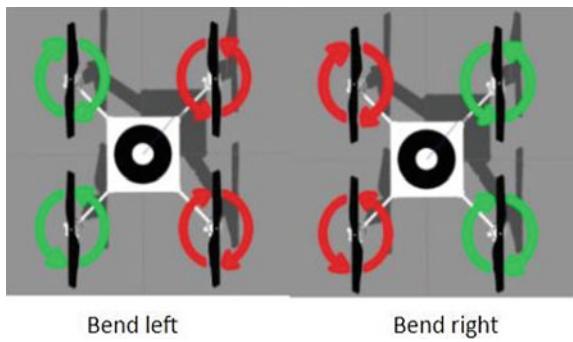
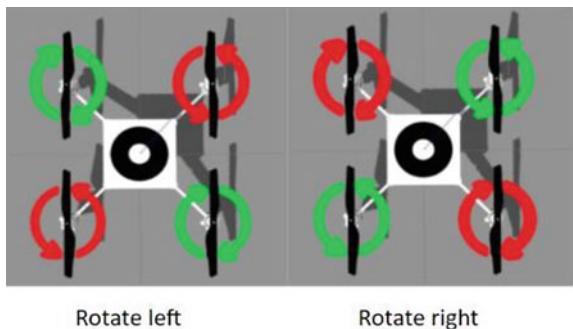


Fig. 4 Pitch movement

- (3) **Roll:** To move the drone left or right, we have to change the angular velocity of the right side or left side motor. In this rolling movement, the pair of left-sided motors rotates with high speed compared to that of the right-sided motors which effects the drone to bend to the right side as shown in Fig. 5.
- (4) **Yaw:** The drone yaws to the right or the left by changing the speed of the alternating motors as shown in Fig. 6.

Fig. 5 Roll movement**Fig. 6** Yaw movement

4.3 Sensors

There are mainly three sensors used in this drone, in order to maintain its ‘pose’ inflight and stability. A drone highly depends on sensors that continuously monitor the drone’s ‘attitude’. Based on the feedback from sensor values, the motor spin speed adjusts and shifts the drone in flight to remain stable.

- (1) **Accelerometer and Gyroscope:** The accelerometer is employed to sense the acceleration of the drone within the X, Y and Z-axis. The gyroscope is used to measure the tilt in pitch, roll and yaw of the drone. MPU6050 sensor consists of a 3-axis accelerometer and a 3-axis gyroscope. It helps us to measure acceleration, velocity, orientation, displacement and other motions related to drones. Mostly, this sensor is used in self-balancing robots and hand-gestured robots.
- (2) **Barometer:** Barometers are used to measure air pressure and detect the height of the drone from the ground. The barometer can easily detect large enough heights.
- (3) **Ultrasonic sensor:** Ultrasonic sensors emit ultrasonic sound waves and convert reflected waves into electric signals, which are extremely useful when flying less than 50 cm from the ground, whereas the barometer gives an accurate distance beyond 50 cm.

5 PID Controller

In this paper, the drone model in Gazebo uses a PID flight controller. PID is a combination of three control systems such as proportional, integral and derivative. These control systems are combined in such a way that they produce a control signal. As shown in Fig. 7, the PID controller uses a control loop feedback system to continuously calculate the difference between the desired set point and measured

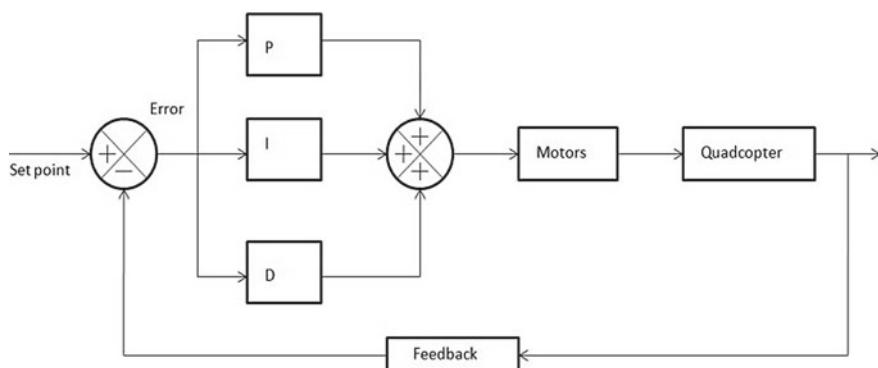


Fig. 7 The overall system block diagram with PID

value or current value to generate an error signal. Based on the generated error signal, the controller will alter the system output in such a way that it will minimize the error. Otherwise, it will reduce the error to zero. The controller in the drone gets the error signal by calculating the difference between the gyroscopic sensor data and the rotation speed of motors. The controller will minimize the error signal by changing the speed of the motors to keep constant output based on the input of the sensor data.

From Fig. 7, it is clear that the error is being continuously altered by three controllers P, I and D. The summation result of the three controllers are being fed to motors and thus enable the drone to fly. Due to the feedback loop in the controller, this process continues until the drone attains stability.

5.1 P Controller

Figure 8 shows the P controller block diagram. P controller gives an output proportional to the measured error at that time. Proportional gain (K_p) determines how intense the flight controller works to correct the error and to reach the set point. In the P controller, a fine-tuned proportional gain (K_p) is responsible for a snappy and sharp control of the drone. The more is the K_p , the lesser is the steady-state error but if the K_p is too high, it causes overshoot in the drone and thus causes instability. The downside of K_p is even when the drone reaches a stable state if the error is too small, the output will be negligible and the drone will never reach the set point (Fig. 8)

$$\text{Output} = K_p \times \text{error} \quad (1)$$

5.2 PD Controller

The PD controller block diagram is shown in Fig. 9. The PD controller can predict the future error by analysing the rate at which the drone is approaching a set point

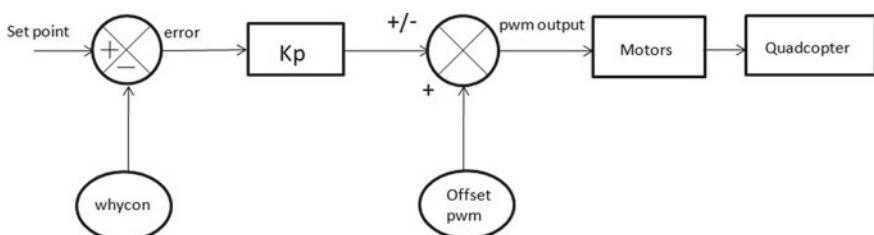


Fig. 8 Depicts the P controller block diagram interfacing with the quadcopter

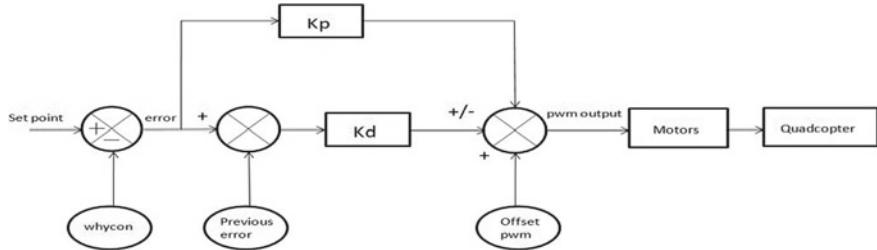


Fig. 9 Depicts the PD controller interfacing with the quadcopter

and changes controller output with the rate of change of error with respect to time. The PD controller helps to decrease the output of P and I when the drone is near to the point to decrease the aggressive changes in the drone. Here, differential gain will work as a shock absorber to drones, and it reduces overshoots and oscillation caused by P and I. If the differential gain is too high, I will also cause vibrations in the drone due to the non-linearity in the system and it also causes overheat in drone motors. Overall, a fine-tuned gain gives a soft control on the drone (Fig. 9)

$$\text{Output} = K_d \times (\text{error} - \text{previouserror}) \quad (2)$$

5.3 PID Controller

Figure 9 shows the entire PID controller flow diagram. The PID controller will eliminate the offset and steady-state error by integrating the error for over a period until the error gets nullified and the I controller also determines the stiffness and sturdiness of a drone to hold the set point against external forces such as wind. If integral gain (K_i) increases, the steady-state error will also decrease, but when K_i is too high overshoot will increase and the stability of the system will also degrade. Just like the P controller flight path, I controller flight path is not efficient and takes a lot of time to settle (Fig. 10).

$$\text{Output} = (\text{sum of errors}) \times K_i \quad (3)$$

$$\text{PID output} = K_p \times \text{error} + (\text{sum of errors}) \times K_i + K_d(\text{error} - \text{previouserror}) \quad (4)$$

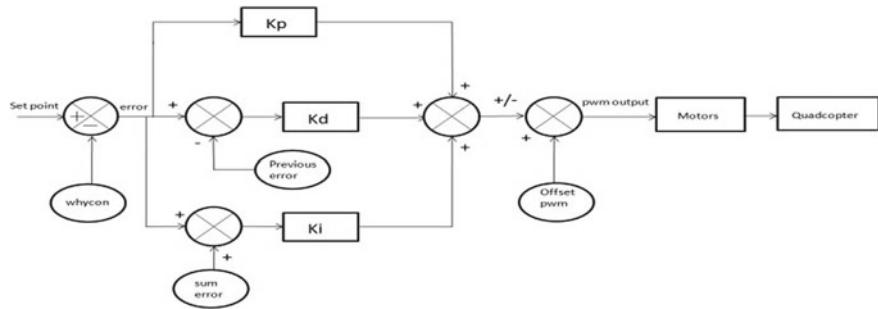


Fig. 10 Depicts the PID controller interfacing with the quadcopter

6 Results and Discussion

Initially, the drone is situated at the origin [0, 0, 0] as shown in Fig. 2. When we run the PID script, the overhead camera detects the WhyCon marker attached at the top of the drone. To reach the destination coordinate, the drone gradually changes its four motors' speed as explained in the drone motion section. The PID script starts calculating errors from the current position coordinates to the previous coordinate position. Based on the position errors, the altitude, pitch and roll values change. By the trial and error method of tuning, the simulation parameters such as K_p, K_d and K_i values of Roll, Pitch, Throttle and settling time are obtained as shown in Table 1.

The graph below in Fig. 11 shows where the drone starts at 19 s and reaches the destination at nearly 24 s and hovers steadily with minimal error. For testing, the destination point in this experiment is [2, 2, 20]. In Figs. 12 and 13, that is altitude errors and pitch, error stabilizes in less than 4 s. In Fig. 14, it takes less than 6 s. it is clearly observed that both the altitude and pitch graph look similar because both the coordinate points of the X and Y-axis are the same. Here, the Yaw error is neglected because there is no clockwise or anti-clockwise rotation required to reach the destination coordinate.

Table 1 Simulation parameters

S. no	Parameters	K _p	K _i	K _d	Settling time (Ts) (s)
1	Throttle	1083	7	1410	4
2	Pitch	262	13	2555	4
3	Roll	157	5	393	6

Fig. 11 Overall error versus time

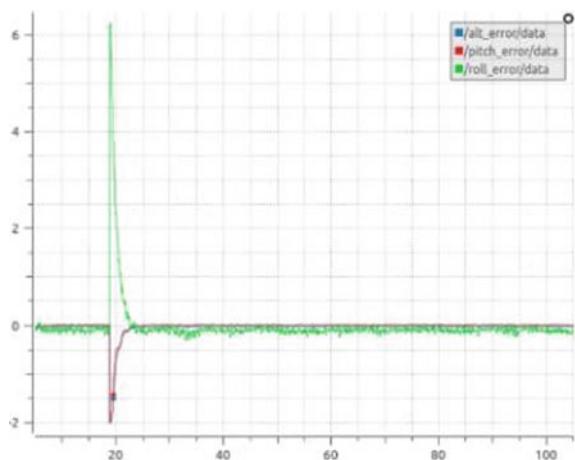


Fig. 12 Altitude error versus time

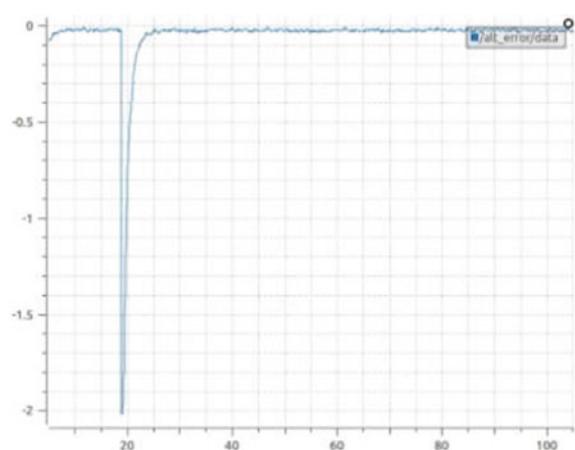


Fig. 13 Pitch error versus time

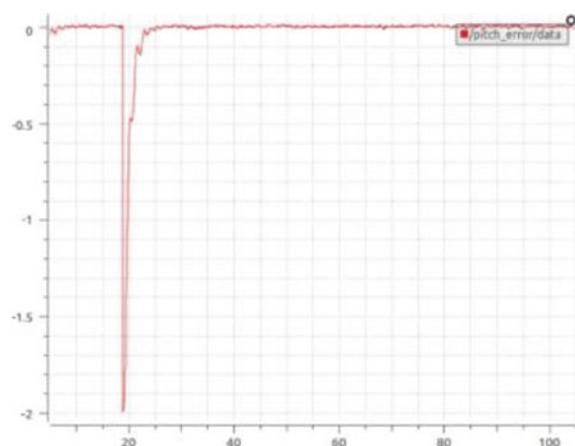
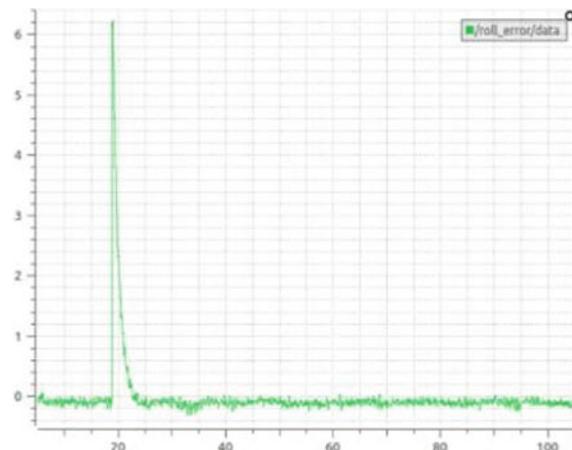


Fig. 14 Roll error versus time



7 Conclusion

In this paper, we presented the simulation of drone interfacing with ROS and the Gazebo simulator, in the tests of the drone to reach the destination point and hover steadily. Due to the physical and structural characteristics of the drone it is not easy to hover the drone with comfortable stability. But required stability can be achieved by tuning the kp, ki and kd values precisely. In a fraction of seconds, the drone reached its destination point. However, there is much scope for this system improvement in the future to increase the wide range of communications by adding different types of communicating devices like cameras for telecasting. This model can be further developed by integrating with GPS for survey and rescue operations.

Acknowledgements We take this opportunity to express our profound gratitude and deep regards to HuT Labs, Amrita Vishwa Vidyapeetham which provided guidance and space for us to complete this work.

References

1. Lukmana, M.A., Nurhadi, H.: Preliminary study on Unmanned Aerial Vehicle (UAV) Quadcopter using PID controller. In: 2015 International Conference on Advanced Mechatronics, Intelligent Manufacture, and Industrial Automation (ICAMIMIA), Surabaya, pp. 34–37 (2015). <https://doi.org/10.1109/ICAMIMIA.2015.7507997>
2. Babu, V.M., Das, K., Kumar, S.: Designing of self-tuning PID controller for AR drone quadrotor. In: 2017 18th International Conference on Advanced Robotics (ICAR), Hong Kong, pp. 167–172 (2017). <https://doi.org/10.1109/ICAR.2017.8023513>
3. Vamsi, D.S., Tanoj, T.V.S., Krishna, U.M., Nithya, M.: Performance Analysis of PID controller for Path Planning of a Quadcopter. In: 2019 2nd International Conference

- on Power and Embedded Drive Control (ICPEDC), Chennai, India, pp. 116–121, doi: 1109/ICPEDC47771.2019.9036558
- 4. Fernando, H.C.T.E., De Silva, A.T.A., De Zoysa, M.D.C., Dilshan, K.A.D.C., Munasinghe, S.R.: Modelling, simulation and implementation of a quadrotor UAV. In: 2013 IEEE 8th International Conference on Industrial and Information Systems, Peradeniya, pp. 207–212 (2013). <https://doi.org/10.1109/ICIInfS.2013.6731982>.
 - 5. Mandal, S., Biswas, S., Balas, V.E., Shaw, R.N., Ghosh, A.: Motion prediction for autonomous vehicles from lyft dataset using deep learning. In: 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, pp. 768–773 (2020). <https://doi.org/10.1109/ICCCA49541.2020.9250790>
 - 6. Sharma, V.N.V.A., Rajesh, M.: Building a quadcopter: an approach for an autonomous quadcopter. In: 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, 2018, pp. 1252–1258 (2018). https://doi.org/10.1109/ICA_CCI.2018.8554718
 - 7. Nithya, M., Rashmi, M.R.: Gazebo—ROS—Simulink framework for hover control and trajectory tracking of crazyflie 2.0. In: TENCON 2019–2019 IEEE Region 10 Conference (TENCON), Kochi, India, pp. 649–653 (2019). doi: 1109/TENCON.2019.8929730
 - 8. Megalingam, R.K., Raj, R.V.R., Masetti, A., Akhil, T., Chowdary, G.N., Naick, V.S.: Design and implementation of an arena for testing and evaluating quadcopter. In: 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, pp. 1–7 (2018). <https://doi.org/10.1109/ICCA.2018.8777578>
 - 9. Megalingam, R.K., Nagalla, D., Nigam, K., Gontu, V., Allada, P.K.: PID based locomotion of multi-terrain robot using ROS platform. In: 2020 Fourth International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, pp. 751–755 (2020). <https://doi.org/10.1109/ICISC47916.2020.9171152>
 - 10. Megalingam, R.K., Nagalla, D., Pasumarthi, R.K., Gontu, V., Allada, P.K.: ROS based, simulation and control of a wheeled robot using gamer’s steering wheel. In: 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, pp. 1–5 (2018). <https://doi.org/10.1109/ICCA.2018.8777569>.

Knowledge Management Framework for the Supervision of IT Postgraduate Research in Sri Lanka



W. M. J. H. Fernando and M. P. A. W. Gamage

Abstract The private sector higher education industry is increasingly attracting a knowledge-based community that depends critically on Knowledge Management (KM) and Knowledge Sharing (KS) activities to expand the quality of supervising postgraduate research students. Using the KM approach to share good research supervision knowledge will help junior research supervisors to conduct quality research with students and thereby help the supervision process to be more successful. The objective of this study is to suggest a conceptual framework that fits in the supervision process. This is conducted to investigate how KM and Information Technology (IT) can be used to develop a model for the supervision process. The framework highlights the critical KM activities in the research supervision process, and it is based on the Task/Technology Fit theory. Using this framework, the knowledge of the more experienced supervisors will be captured and used by junior supervisors in their supervision process.

Keywords Knowledge Management · Postgraduate research · Task/Technology Fit Theory · Information Technology · Conceptual framework

1 Introduction

In the present world of knowledge revolution, knowledge has been passed down generation by generation. Amin [1] has pointed out that knowledge is a critical component of global competitiveness. It supports nurturing knowledge-sharing mechanisms and learning abilities in and out of the workplace. Understanding the status of knowledge is an important element in universities that look after

W. M. J. H. Fernando (✉)

Faculty of Graduate Studies and Research, Sri Lanka Institute of Information Technology,
Malabe, Sri Lanka

e-mail: jeewafdo91@yahoo.com

M. P. A. W. Gamage

Faculty of Computing, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka
e-mail: anjalie.g@slilit.lk

knowledge-sharing mechanisms and learning skills within an organization and across organizations.

KS is a major component of KM and its strategies are the solutions to structural and individual progress. It is mainly defined as the method of distributing information and knowledge within the workplace [2]. The method of exchanging and shifting existing knowledge and ideas among postgraduate research students creates new knowledge and ideas to help higher education institutions to reach their objectives. This aims to provide a better communication link between the students and the supervisor and helps to align and complete research work in the field of IT more successfully. The findings from this study will bring advantages to IT postgraduate research supervision. The KM system (KMS) attempts to integrate the relevant information technology into the research supervision process that would support supervisors' knowledge creation, knowledge sharing, applications, and knowledge management. The proposed framework would assist the institutions in higher education to identify the critical KM activities associated with the KM Technology and plan for the postgraduate research supervision KMS for the supervision process.

Education can be measured as one of the most important effective investments that are involved all the time. Learning through a lifetime helps to search, protect, and broadcast knowledge. There is a close relationship in the fundamental roles between education in higher education institutions and KM. Excelling in higher education is about changing students, and universities should provide students with a transformative learning environment. The problem identified is that postgraduate research studies are challenging and sometimes students may experience problems making progress. Currently, Sri Lankan private sector education institutions face a lot of problems during their postgraduate research supervision and in their profession, which requires more attraction for management. Some of these challenges consist of miscommunication among project supervisors and students, ambiguities in research doings, and lack of effecting the following processes for the position of different research work. The top management at the institutions must play their role and put more effort to ensure that their project supervisors have the proper platform and support to share their knowledge among students.

Mainly, this study focuses on state universities and private higher education institutions in Sri Lanka. Currently, state universities rank higher in postgraduate research than private institutions, and the research supervision process is at a standard level. It was noted that KM can be used to increase the ability of private higher education institutions to reach their targets by using and sharing knowledge in different ways which progress to success as a team. Part of KS relative to individual performance among research students is to accomplish excellent supervising and learning methods. This also provides standard education and postgraduate qualification which are required by the staff. KS is also a vital component when it comes to achieving the vision and missions of private higher educational institutions.

2 Literature Review

Higher education institutions of Sri Lanka under the private sector are registered under the company act. They provide IT undergraduate degrees, limited postgraduate degrees, and M.Phil and Ph.D. programs [3]. Some even offer postgraduate degree programs at foreign universities [4].

The main role of private education is generating and moving valuable knowledge among the followers themselves to make them competent [5]. Currently, private higher education institutes have several years of experience in their academic work where many professors are employed in their institutes. The main objective of this research is to suggest a conceptual framework that fits in the supervision process, using KM activities and IT for designing such a model for supervision [6]. The main aim is to provide a better communication link between the students and the supervisor and help to align and complete research work in the field of IT more successfully. It is often evident that junior supervisors and students face difficulties in initiating research activities. The differences among supervisors and students' stages of knowledge and skills augment the problems in the research supervision process [7].

Research is a communicating process and needs the development of social as well as theoretical skills. A school's managerial function is normally taken to means as referring to handling, operating, or directing an institute to help students toward the accomplishment of a Masters qualification. A graduate student sometimes has previous knowledge and problems in their research project [3]. As a result, this requires the effective use of resources in research and supervision. Developing skills for effective research supervision are essentials to be engaged in various ways [6].

Swanson and Watt [3] show theoretical guidelines for research supervision. This theoretical guideline aims to help supervisors in research development to make effective monitoring of research supervision [8]. This work shows that life-long learning and life-long research are created on the main 4 stand framework of learning.

- Learning to know
- Learning to do
- Learning to live with others
- Learning to do.

Yew et al. [9] state that research supervisors' performances are progressively being determined in the form of education experts' procedures and protocols that drive supervision service distribution. Project supervisors specifically use the research supervision knowledge that joins with their know-how and experience to carry supervision services to supervision.

According to previous works, there is no clear analysis of the processes of post-graduate research supervision planned by researchers to plan the supervision activity based on a dynamic process rather than a static process. They propose various methods and systems to manage supervision activities.

Gatfield and Alpert [10] have proposed the supervisory management styles (SMS) model. It recommends a four-quadrant supervisory style management grid that emphasizes the understanding of supervisory styles and adjustments in the supervisory style during the supervisory phase. The authors have said that supervisors tended to think that they knew which parts of the supervisory process and management styles were more suitable for success. This SMS model highlights the following things:

- Supervisor's help and
- Managing skills of students to handle research and their ability to communicate and work together.

Maor and Currie [11] have proposed the blended postgraduate supervision (BPS) model which is offered in three styles through the classroom, the virtual classroom, and online courses. They have shown that postgraduate supervision is blended learning in these delivery methods, with the exclusion that there is hardly any classroom education but rather face-to-face communication between supervisors and students. The core thing is to offer students with links on web pages that agree with students to access several topics either through the Internet or Web CT. This is a chance for students to discuss matters and consult with their project supervisor and use different kinds of technologies to make and share information. Research is a significant method for learning for postgraduate students. It allows them to improve their experiences and range the limits of their knowledge. The practice of IT-based communication, social media, online databases, and e-learning systems can expand their skill to detention, transmission, and to share information [11]. The BPS shows the supervisors to signify the key foundations of information to be retrieved by students in face-to-face discussions and via the Internet and libraries and assess what the postgraduate student has gained from it. Therefore, the BPS model highlights the following facts:

- Using web technology
- Using IT-based communication
- Supervisor's support.

The KM model for the supervision process is a better model that applies to detect the knowledge-sharing impact factors since it is just a research supervision model that is specifically designed for KM and KS. Using this method, one can select the factors which are specific to the research supervision environment. This model has been suggested by Paul [12]. According to that, for students to graduate successfully, the supervisory process is very important. The method is complicated and subtle.

A systematic knowledge-sharing approach is required to help both supervisors and students to achieve, share, and apply knowledge. A knowledge-sharing method proposed by Paul [12] shows that the supervisor focuses on helping students to improve the knowledge-sharing capability in research supervision. This knowledge-sharing capability means not only the expertise of using enhanced technological resources to manage the information but also the potential to decide about choosing

and using the information. Additionally, Zhao has insisted that postgraduate supervision involves KM and thought that the effectiveness of supervision can be improved by incorporating its concepts.

KM is an educational benefit and financial resource that supports the quality of a university and operates the research supervision as acquiring knowledge to add benefit to the university. It is also measurement to evaluate the quality of academic work of a specific university. The entire process of education occurs when people share their knowledge because knowledge is built into the minds of individuals. The important success tips of knowledge sharing are highlighted as a fact in Zhao's KM model.

3 Methodology

This research is conducted as authors' quantitative research. It is based on the authors' collection of data-based theories, hypotheses, and experiments. It is followed by descriptive and inferential statistical methods. According to the nature of the analytical research, it is extended to a descriptive approach to explaining the way something that has happened or is happening.

3.1 Scope of the Study

This study is mainly focusing on state and private sector higher education institutions, under the Sri Lanka University Grant Commission. The targeted participants are the project supervisors in the Faculty of Graduate Studies. The research aims to provide a better communication link between the students and the supervisors and help to align and complete research work in the field of IT more successfully. The findings from this study will bring an advantage to the supervision of IT postgraduate research. This study helps to cultivate and justify a framework that helps to examine the project supervisor's behavior toward KM and KS with students. The outcome of this research will be a framework for proper project supervision which can be applied for IT postgraduate research supervision.

3.2 The Study Population and Sample

This study focuses only on the IT postgraduate degree programs offered at state and non-state universities. There are six state universities and approximately ten non-state universities in Sri Lanka. According to statistics of IT postgraduate research, the staff population of these state universities is nearly 300 under the categories of academics. According to the statistics of each private university, it consists of approximately 60

staff under the categories of academics in one university. Therefore, the total staff of all the non-state universities is nearly 300. In this study, the target population is 300. The sample population was taken from 07 universities, 02 state and 05 non-state universities.

The purposive sample was applied to examine the hypotheses of this study. In the end, researchers made a verdict of the study depending on the chosen sample. Therefore, the sample required to be accepted is based on the aimed population. The findings can be generalized based on selected received samples.

The existing study comprised 277 responses within seven (both state and non-states) universities. In compliance with the current study, selected state university respondents (40%) were University of Moratuwa and University of Colombo. Chosen non-state university respondents (60%) were SLIIT, APIIT, ESoft, IIT, and ICBT campus.

3.3 Conceptual Framework

This research proposes that KM can be used to increase the ability of private higher education institutions to reach and use supervisor perception and organizational culture in different ways which progress to success.

The conceptual framework for this research is to design from the above literature review. It includes two fields such as independent variables and dependent variable.

Independent Variables

- The ability of a supervisor to share knowledge.
- Culture of the university.
- Supervisor support.
- Learning strategy.
- IT system.

Dependent Variable

- Quality postgraduate research supervision.

3.4 Proposed Hypotheses

- A supervisor's ability to share positively affects KS in quality IT postgraduate research supervision.
- The culture of a university has an important and positive effect on KS in quality IT postgraduate research supervision.
- Supervisor support positively affects KS in quality IT postgraduate research supervision.

- The learning strategy has an important and positive effect in KS in quality IT postgraduate research supervision.
- IT system has an important and positive effect in KS in quality IT postgraduate research supervision.

4 Discussion and Finding of the Research

This study is done to develop a KM framework for the supervision of IT postgraduate research in Sri Lankan higher education institutions that play an essential role in the academic industry. Mainly, the supervision process, Knowledge Management (KM) activities, and technologies are the elements investigated in the research.

This aims to provide a better communication link between the students and the supervisor and help to align and complete research work in the field of IT more successfully. The findings from this study will bring an advantage to the supervision of IT postgraduate research. This study helps to cultivate and justify a framework that helps to examine the project supervisor's behavior toward KM and KS with students. The outcome of this research is a framework for quality project supervision which can be applied for IT postgraduate research supervision.

4.1 *Finding of the Study*

The general objective of the study is to identify the sources of quality supervision IT postgraduation research supervision. Multiple regression analysis was used to determine whether the ability of a supervisor to share knowledge, the culture of a university, supervisor support, and IT system they use for quality research supervision in Sri Lanka. The results from the regression analysis are presented in Table 1.

Table 1 The results from regression analysis

Model		Unstandardized coefficients		Standardized coefficients	t	Sig.
		B	Std. error			
1	(Constant)	0.027	0.103		0.267	0.790
	SSK	0.281	0.056	0.280	5.074	0.000
	CU	0.184	0.093	0.205	1.980	0.050
	SS	0.741	0.086	0.692	8.600	0.000
	LSU	0.004	0.047	0.005	0.082	0.935
	ITS	0.200	0.077	0.173	2.602	0.010

Table 1 presents the results of SPSS linear regression output. The p-value for the ability of the supervisor to share knowledge, the culture of a university, supervisor support, and IT system is less than 0.05. Hence, quality postgraduate research supervision of the research supervisors depends on the ability of supervisors to share knowledge, the culture of a university, supervisor support, and IT system. The R-square value was 0.889 ($F = 85.694$, $p < 0.05$), which means 9.665% of the variation in quality postgraduate research supervision can be explained by the ability of the supervisor to share knowledge, the culture of a university, supervisor support, and IT system identified in this study.

The VIF values are less than 5. As Denis [13] mentioned if VIF for one of the variables is around or greater than 5, there is multicollinearity associated with that variable. Hence, there is no problem with multicollinearity in this model. The Durbin-Watson (DW) statistic should fall within the acceptance range from 1.53 to 2.50 to ensure that there is no autocorrelation problem in the data [14]. The DW statistics of the model is 2.145, where values around 2 indicate no problem of autocorrelation. In residual diagnostics, the residuals were independent and normally distributed. Between the culture of university and supervisor support, resources recorded the highest beta value ($\beta = 0.741$, $p < 0.05$). As a conclusion, the result of multiple regression analysis is supported by the data: H1 (a supervisor ability to share positively affect KS in quality IT postgraduate research supervision), H2 (the culture of a university has an important and positive effect in KS in quality IT postgraduate research supervision.), H3 (supervisor support positively affects KS in IT postgraduate research supervision), and H4 (IT system has important and positive effects in KS in quality IT postgraduate research supervision.).

Taking a deeper look, the effects of the ability of the supervisor to share knowledge, culture of a university, supervisor support, and IT system on quality postgraduate research supervision were identified separately using stepwise regression analysis. The R-square value of the model indicated 79% of the variability in quality research supervision. The R-value of the model indicated 88% of the variability in quality postgraduate research supervision explained by the ability of the supervisor to share knowledge, the culture of a university, supervisor support, and IT system.

The overall results of the multiple regression analysis and the study hypotheses were presented in Table 2 illustrating the direct effect model of the study.

4.2 Discussion

The proposed framework highlights the critical KM activities in the research supervision process, and it is based on the Task/Technology Fit theory. Using this framework, the knowledge of the more experienced supervisors will be captured and used by junior supervisors in their supervision process.

According to Lin [15], quality research supervision depends on knowledge sharing and the skill of the research supervisor. The study resulted based on several factors of

Table 2 Hypothesis results of multiple regression model

Hypotheses	Result
A supervisor's ability to share positively affects KS in quality IT postgraduate research supervision	Supported
The culture of a university has an important and positive effect on KS in quality IT postgraduate research supervision	Supported
Supervisors' support positively affects KS in quality IT postgraduate research supervision	Supported
The learning strategy has an important and positive effect in KS in quality IT postgraduate research supervision	Not supported
IT systems have an important and positive effect in KS in quality IT postgraduate research supervision	Supported

the outcome, among which it identified different research supervision improvement factors for quality research supervision.

This study integrates perspective from the ability of the supervisor to share knowledge, the culture of a university, supervisor support, the learning strategy, and IT system to propose a suitable framework to identify the factors influencing to improve quality of IT postgraduate research supervision to private higher education institutions in Sri Lanka. The study is based on the Task/Technology Fit theory which has primarily focused on individual performance and the skills of the IT level. Supported by the practical indications, this study originates that factor settings have the most significant influence on the university culture and the way that they help students to complete their research work. Among the above factors, the ability of the supervisor to share knowledge, culture of a university, supervisor support, and IT system has also made a positive impact on quality research supervision. It should be pointed out that the findings of this study are based on data collected from state and non-state universities in Sri Lanka.

4.3 *Proposed Framework*

Private sector higher education institutions are progressively attracting a knowledge-based society which varies critically on KM and KS actions to increase the excellence in supervising postgraduate research students (see Fig. 1).

Expanding the KM approach to share quality research supervision knowledge will help junior research supervisors to conduct quality research with students and thereby help the supervision process to be more successful.

According to the suggested framework for KM in the IT postgraduate research supervision, it introduces the research supervision process with a few inputs,

1. Research student.
2. Research environment.

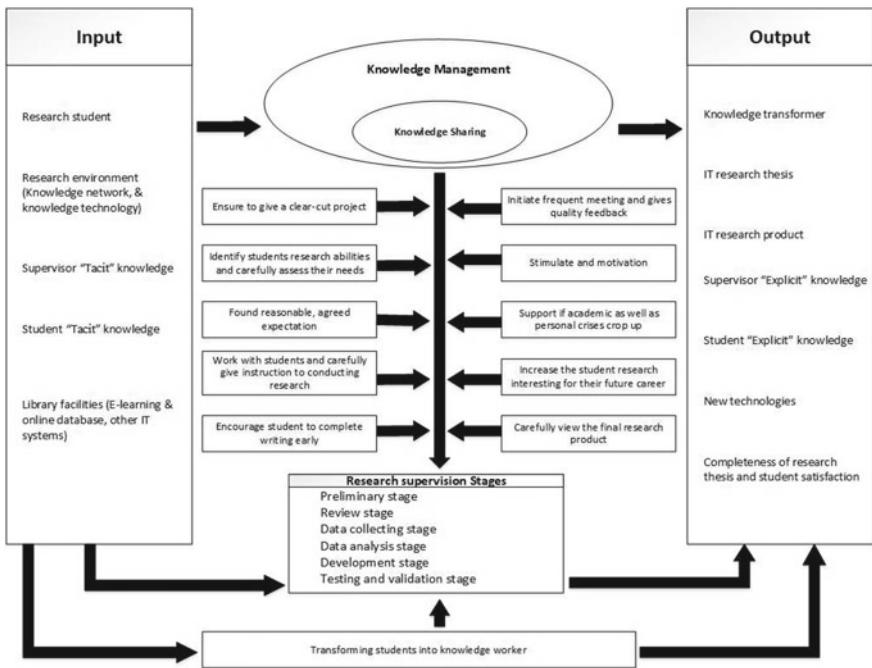


Fig. 1 Proposed framework to improve quality of IT postgraduate research supervision

3. Supervisors' and students' tacit knowledge.
4. Library facilities (E-learning and online database, other IT system).

Apart from the above, the framework also identifies the following outputs:

1. Knowledge transformer.
2. IT research thesis.
3. IT research product.
4. Supervisors' and students' explicit knowledge.
5. New technology.
6. Completeness of research thesis and student satisfaction.

The study process has a few steps such as defining the research problem and developing research methodology. Throughout this process, the tacit knowledge of students and supervisors is transformed into explicit knowledge, and a study on the product and theory is being produced. KM and KS actions increase the excellence of postgraduate research supervising students. Knowledge Sharing (KS) plays a significant role in this process as it helps and expands the process by improving the quality of IT postgraduate research supervision. The identified factors are listed as follows:

1. Ensure to give a clear-cut project.
2. Identify students' research abilities and carefully assess their needs.

3. Found reasonable, agreed expectation.
4. Work with students and carefully give instructions for conducting research.
5. Encourage the student to complete writing early.
6. Initiate frequent meetings and gives quality feedback.
7. Stimulation and motivation.
8. Support during an academic or personal crisis.
9. Increase the student's research interest for their future career.
10. Carefully view the final research product.

Depending on the framework, the organizational, individual, and technological factors influence the sharing of knowledge in this process.

This factor aims to help supervisors in research supervision to make effective monitoring and management of supervision. According to the framework, the following process stages are needed to be followed by supervisors.

1. Preliminary stage.
2. Review stage.
3. Data collecting stage.
4. Data analysis stage.
5. Testing and validation stage.

This framework is tested by research supervisors of private universities and final year research students in Sri Lanka. One point that they have mentioned in the feedback was impenetrability measuring individual knowledge level. The main reason they emphasize was that there is no correct mechanism to measure current junior supervisors' knowledge and skills of private higher education institutions. By improving the quality of IT postgraduate research supervision in private higher education institutions, the rank of the institutes and the demand for their degree programs can be increased.

5 Results

In this study, the factors and requirements were gathered by using standard techniques. It has carefully analyzed the gathered data using several methods. Requirements' prioritization methods were used to categorize and give the prioritization for requirements. Also, it analyzes whether there are trends and patterns in the requirements. It also checks if there is a statistical equivalence in the questionnaires which were taken from the users. By using these standard techniques for validating requirements, as a result, it comes up with a requirement specification with expected changing requirements in the future.

The developed framework as a result of the study was evaluated to check its validity. In the validation process, it considered taking the finished framework and assessed it for aspects of the usability, and consisted with carrying out experiments after implementation, and the feedback was taken to find how far the framework is

successful. Here, online questionnaires were distributed among a selected group of students to verify the developed framework which applies to the IT postgraduate research process. The framework diagram was distributed with the questionnaires. The sample student group size was 130. 83% of students very much liked the final framework and features. Here, it is conducted to mainly get the responses from the students about the quality research supervision of the proposed framework.

6 Conclusion and Future Research Work

This research helps university and postgraduate research supervisors to realize which factors are important for knowledge sharing in the research supervision process. Also, it shows the way that supervisors improve their performance and the way to increase the quality of the research supervision process. Mainly, this research is focused on effective KS that can increase the performances, productivity, and usefulness in the postgraduate research supervision process. It helps to improve inventions in private sector higher education institutions. The main objective of the KS approach for postgraduate research supervision is to increase the equality of the postgraduate research training, through the students' knowledge and research experience in the revolution. In the university, KS helps in reducing the cost and is a cheaper path to conduct postgraduate research. It helps the student to complete their MSc degree program within a given period without delays. The outcomes of this study could encourage the university higher management to give structures that would encourage students and research supervisors to share their knowledge. This encouragement may come in the form of incentives, identifications, rank, or development of the university.

This research was conducted only for postgraduate research supervision on the state and non-state universities in Sri Lanka. The discovered result outcomes were found by collecting data gathered by selected seven universities and 120 research supervisors.

Therefore, it is recommended that this statistic of this study can be applied against evaluating related trends in a wider context, like evaluating quality postgraduate research supervision of any higher education institution.

In this study, each of the KS effect aspects that fit from the literature for the proposed framework was from several iterative factors that were applied in various areas. The proposed study framework investigates these precursor factors in new areas. The proposed framework in this research can be applied to more universities with various cultures. In the research supervision domain, individual, organizational technological factors are very important on KS in the university environment. But it is recommended that future work should review not only the impact of KS factors (individual, organization technological) but can also review the impact of KS on the excellence of both experiences of learning and research stages. The results of such an analysis might improve the achievement rate of students and encourage critical thinking.

References

1. Amin, S.H.M., Zawawi, A.A., Timan, H.: To share or not to share knowledge: observing the factors. In: 2011 IEEE Colloquium on Humanities, Science and Engineering, Penang, Malaysia, pp. 860–864, Dec. 2011. <https://doi.org/10.1109/CHUSER.2011.6163859>.
2. Dooba, I.M., Downe, A.G., Mahmood, A.K.: If professors knew what professors know: A technique for capturing university teachers' tacit knowledge of research supervision. In: 2010 International Symposium on Information Technology, vol. 1, pp. 1–5, Jun 2010. <https://doi.org/10.1109/ITSIM.2010.5563117>
3. Phang, F.A., Sarmin, N.H., Zamri, S.N.A., Salim, N.: Postgraduate Supervision: Supervisors versus Students," in 2014 International Conference on Teaching and Learning in Computing and Engineering, Kuching, Malaysia, Apr. 2014, pp. 251–255, doi: <https://doi.org/10.1109/LaTiCE.2014.55>.
4. Sri Lanka Qualifications framework
5. Sheng, D., Qingli, L., Liang, S.: Study on improving the service level of engineering supervision of China. In: 2010 7th International Conference on Service Systems and Service Management, Tokyo, Japan, pp. 1–4, Jun 2010. <https://doi.org/10.1109/ICSSSM.2010.5530125>
6. Jassim, O.A., Mahmoud, M.A., Ahmad, M.S.: A Framework for Research Supervision, p. 10 (2015)
7. Jamal, H., Shanaah, A., Wingkvist, A.: The Role of Learning Management Systems in Educational Environments: An Exploratory Case Study, p. 67
8. Lubega, J.T., Niyitegeka, M.: Integrating E-supervision in Higher Educational Learning, p. 8
9. Yew, K.-T., Ahmad, W.F.W., Jaafar, J.: A framework for designing postgraduate research supervision knowledge management systems. In: 2011 National Postgraduate Conference, pp. 1–6, Sep 2011. <https://doi.org/10.1109/NatPC.2011.6136310>
10. Gatfield, T., Alpert, F.: The Supervisory Management Styles Model, p. 11
11. Maor, D., Currie, J.K.: The use of technology in postgraduate supervision pedagogy in two Australian universities. Int. J. Educ. Technol. High. Educ. **14**(1), 1 (2017). doi: <https://doi.org/10.1186/s41239-017-0046-1>
12. Paul, S., Verma, J.K., Datta, A., Shaw, R.N., Saikia, A.: Deep learning and its importance for early signature of neuronal disorders. In: 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, pp. 1–5 (2018). <https://doi.org/10.1109/ICCCA.2018.8777527>
13. Simple and Multiple Linear Regression: ResearchGate (2020). https://www.researchgate.net/publication/326876947_Simple_and_Multiple_Linear_Regression. Accessed 18 May 2020
14. Diana-Rose, F., Zariyawati, M.A., Norazlina, K., Annuar, M.N., Manisah, O.: Consumers' Purchasing Decision towards Food Products of Small and Medium Enterprises, vol. 6, no. 4, p. 8 (2016)
15. Lin, H.: Knowledge sharing and firm innovation capability: an empirical study. Int. J. Manpow. **28**(3/4), 315–332, Jun 2007. doi: <https://doi.org/10.1108/01437720710755272>

RFM-Based Customer Analysis and Product Recommendation System



Rahul Krishnan and Prashant R. Nair

Abstract In this current Covid-19 pandemic scenario, most of the supermarkets and retail stores that are present even in small towns have started giving out products online. This system classifies the customers into categories based on their Recency, Frequency and Monetary (RFM) values and recommends products to them to make sure the potentially valuable customers are retained. Based on the RFM scores, each of these customers will be divided into segments based on the scores that they obtain. To make sure that the customers are retained, better offers should be given to them, and they must also get proper product recommendations. The product recommendation system that is used is based on a concept called cognitive similarity. It is a hybrid recommendation model by taking Collaborative filtering as the base model. Comparison is also done with the existing methods to make sure that there is an improvement in the method put up. For performance evaluation, precision@k was considered. The proposed method using cognitive similarity shows an improvement of 5% in comparison with the existing methods.

Keywords RFM analysis · Customer segmentation · Cognitive similarity

1 Introduction

The application area of this project is mainly in direct marketing. The example of supermarkets has been considered for the study of existing systems in the area [1]. Other application domains have also been taken into consideration such as digital footprints from mobile phone data [2] on which Exploratory Data Analysis and Predictive Analysis have been used in the domain of grocery shopping. An important factor in any marketing field is finding out the customer lifetime value [3]. Most

R. Krishnan · P. R. Nair (✉)

Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India
e-mail: prashant@amrita.edu

R. Krishnan
e-mail: cb.en.p2cse19019@cb.students.amrita.edu

of the profit for a supermarket comes from a very few percent of customers. So, it is important to find out who those customers are and they should be retained. Apart from that, finding out new customers and making them loyal is also necessary for building the sales. A feasible way for making this possible is by segmenting the customers. Thereby, the most valuable customers can be found and appropriate measures can be taken to make sure they are retained. The RFM method [4] has been used for segmenting the customers. After segmentation, product recommendation can be done. Different strategies exist in making a recommendation of products for a customer. General awareness of some of the major machine learning algorithms has been done on the domain [5]. Like various implementations in the area [6], implementation of various classification methods was also considered [7]. K-Means clustering [8] was a method widely used in segmentation. As far as the recommendation is concerned, collaborative filtering [9] is the most used model. Another important thing for retaining customers and also for allowing new customers to be a part is the product recommendation system. The product recommendation system is recommending products to the customers based on their past purchase history. There are different types of product recommendation systems. Normally, the recommendation system works like it will check your purchase history and recommends products of the same trend. In the collaborative filtering method, this will take into account two things. First is taking into account all the purchase history of the customer. Then is to find out other customers with the same purchase history and find out what other items the second customer has bought and suggest the same to the first customer. Studies have proven that all the methods have their own advantages and disadvantages, so combining more than one method will help in reducing the demerits.

2 Related Work

2.1 RFM Analysis

In supermarkets, there will be different types of customers and it needs to be made sure that you retain your high-valued customers. For checking which customers are of the highest value, the RFM method [4] is used. RFM is Recency (how recently a person visits the supermarket), Frequency (the number of times the customer visits the supermarket) and Monetary (how much amount does the customer spend on purchase). Based on these three, a ranking system would be given to the customers. A separate ranking system will be deployed. That is one for Recency, one for Frequency, and one for Monetary. Based on this ranking system, the top rank of each category will be given equal scores. Then we will add the points of all three categories for a single customer. Based on the scores that the customer gets, they would be clustered. The ones with the highest scores will be the most important customers. They can be said to be the best customers. Then there would be categories like highly valued customers, and it is needed to make sure that those customers would be retained.

Table 1 Details of customer segmentation

Name	RFM_segment	Detail
Best customers	444	Highest frequency as well as monetary value with least recency
Loyal customers	344	High frequency as well as monetary value with good recency
Potential loyalists	434	High recency and monetary value, average frequency
Big spenders	334	High monetary value but good recency and frequency values
At risk customers	244	Customers shopping less often now who used to shop a lot
Can't lose them	144	Customers shopped long ago who used to shop a lot
Recent customers	443	Customers who recently started shopping a lot but with less monetary value
Lost cheap customers	122	Customers shopped long ago but with less frequency and monetary value

Better offers should be given to those customers. Similarly, a range would be provided and customers will be divided into different clusters based on the final scores that are obtained. Each person who belongs to the same cluster will get the same offer. Top offers will be given to the customers in the first cluster (the ones with the highest score), but other customers which are of value will also be given subsequent offers.

Based on the RFM scores obtained, they were divided into different segments as shown in Table 1. The RFM scores for each category range from 1 to 4 with 1 being the lowest score and 4 the highest. Specific ranges are provided for each category according to which the scores are given. The customer with a value of 4 in all categories is classified as the best customer.

2.2 Recommender Systems

There are mainly 4 basic recommendation strategies [10]. There are advantages and disadvantages to all these basic models. The most prominent method is collaborative filtering in which a user's purchase history will be analyzed in relation to others. The system will check for the items which user 2 has brought but which user1 hasn't and those will be recommended to user 1. This method has a disadvantage though. It suffers from the cold start problem [9]. Cold start problem is when a new customer comes and he/she has no purchase history, then there won't be any product to recommend. Next comes the content-based filtering [10] method in which products will be recommended based on similarity. Here also, the user's purchase history would be taken into consideration and the properties of those items which are similar to the ones purchased by the user are found out and thereby those products are

recommended. This method also suffers from the cold start problem. Another filtering method is the Knowledge-based filtering method [10] in which the products would be recommended based on the knowledge about users. But for that, it requires some feedback or information about the user first which sometimes the users might not be ready to provide. Unless there is information about the users, this method is also incapable of solving the problem. A method which doesn't suffer from the cold start problem is demography-based recommendation [10]. This takes into consideration gender, age and educational qualification and recommends products based on that. This too has a disadvantage because as everyone's tastes are not alike, this would mostly lead to wrong recommendations which might cause dissatisfaction among customers. As any single method cannot solve the problem, combining these methods will lead to useful solutions. There comes the hybrid recommender [9]. As the best method is collaborative-based filtering [9], it is best to combine collaborative filtering with any other filtering models. The block diagram of the proposed recommender system is shown in Fig. 1.

2.3 Cognitive Similarity

Cognitive Similarity is used for finding out the similarity between the users. If there are 4 customers say customer1, customer2, customer3 and customer4, for recommending a product to customer1, his relationship with all other users will be calculated. If two customers are found with the same purchase history, then products bought by the second customer and not bought by the first customer, will be recommended to the first customer and vice-versa. That is, based on how a customer selects an item, a product would be recommended to him/her. There will be cases in which a similar user does not exist. In such a case, content-based product recommendation would be deployed. This method solves the cold start problem that will occur if only collaborative filtering is used.

3 Results

Precision@K is used for evaluating how well the recommendation system performed. Precision is one of the most commonly used metrics for evaluating the recommender system. Precision@K can be defined as out of k recommendations made, how many of them were useful. If the value of K is taken as 10, then precision@10 will be calculated. That is out of 10 recommendations made, how many were useful. If 7 of the 10 recommendations were found useful, then the precision will be 70%. Selecting the value k is based on top-n recommendations rather than taking all of

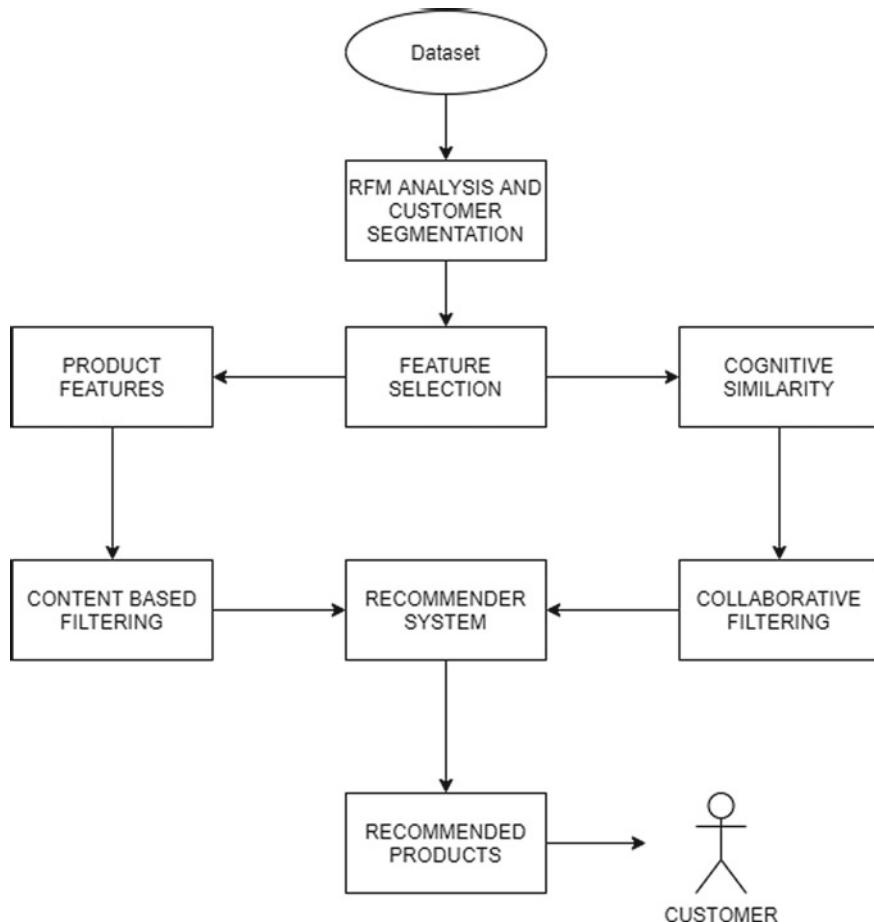


Fig. 1 Block diagram of the recommender system

the recommendations. Based on the calculation of Precision@K, the conventional collaborative filtering method showed a precision of 61.8% and the method proposed using cognitive similarity showed a precision of 66.8%. There has been an increase of 5% in precision for the proposed method (Table 2).

Table 2 Performance comparison of the methods

	Precision@k
Existing method (collaborative filtering)	61.8
Proposed method (cognitive similarity)	66.8

4 Conclusion and Future Enhancements

The customers have been properly segmented based on their purchase history, and appropriate recommendations have also been given to them. It could be inferred that it is better to divide customers into segments and analyze them instead of considering them as a whole. Compared to the traditional approach, the method proposed gets a better precision and thus it can be used to recommend products in a better way. This system also overcomes the cold start problem which was there in some of the recommendation systems. If it is possible to get implicit feedback, it would be better than using explicit feedback. With the help of user feedback, further methods can be deployed with the help of deep learning and NLP techniques. The deep learning concept of multilayer perceptron can be added as part of the collaborative filtering method. Here the important method being used is with the help of a feedback network which takes user feedback at regular intervals and applies the model based on that. So, this method can be used as a future work to try and get even better recommendations.

References

1. Subathra, P., Ghanapathy, S.V., Chidambaram, A.R.V., Ganesh, K.R.: Exploratory data analysis and predictive analysis on grocery shopping . *J. Adv. Res. Dyn. Control Syst.* **10**(5), 1803–1809 (2018)
2. Jisha, R.C., Krishnan, R., Vikraman, V.: Mobile applications recommendation based on user ratings and permissions. In: International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, India (2018)
3. Ahmad, A., Floris, A., Atzori, L.: OTT-ISP Joint service management: a customer lifetime value based approach. In: 2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM), IEEE (2017)
4. Gustriansyah, R., Suhandi, N., Antony, F.: Clustering optimization in RFM analysis based on k-means. *Indonesian J. Electr. Eng. Comput. Sci.* **18**(1), 470–477 (2020)
5. Dhanya, N.M., Veerakumar S.: Performance analysis of various regression algorithms for time series temperature prediction. *J. Adv. Res. Dyn. Control Syst.* 175–194 (2018).
6. Subbulakshmi, S., Ramar, K., Shaji, A., Prakash, P.: Web service recommendation based on semantic analysis of web service specification and enhanced collaborative filtering. In: Intelligent Systems Technologies and Applications, Cham (2018)
7. Baskar, A., Gireesh K.T.: Facial expression classification using machine learning approach: a review. In: Advances in Intelligent Systems and Computing, vol. 542, pp. 337–345 (2018)
8. Bhavani, K.D., Radhika, N.: K-means clustering using nature-inspired optimization algorithms-a comparative survey. *Int. J. Adv. Sci. Technol.* **29**(6), 2466–2472 (2020)
9. Chen, R., Hua, Q., Chang, Y., Wang, B., Zhang, L., Kong, X.: A survey of collaborative filtering-based recommender systems: from traditional methods to hybrid methods based on social networks. *IEEE Access* **6**, 64301–64320 (2018)
10. Cano, E., Morisio, M.: Hybrid recommender systems: a systematic literature review. *Intell. Data Anal.* **21**(6), 1487–1524 (2017)

Efficient Hardware Trojan Detection Using Generic Feature Extraction and Weighted Ensemble Method



Vaishnavi Sankar and M. Nirmala Devi

Abstract Usage of smart devices has tremendously grown with the advent of Internet of Things (IoT). Like software, Integrated Circuit (IC) being the core component of hardware is also highly un-trusted. Outsourcing of ICs to off-shore foundries makes Hardware threats inevitable in IC design. Hardware threats are of five types, namely Intellectual Property (IP) piracy, IC over-building, counterfeiting, reverse engineering, and Hardware Trojan [2]. One of the most sought-after hardware security breaches is Hardware Trojan insertion. Many methods have been proposed in the last decade for Hardware Trojan detection. Recently, machine learning techniques are used to perform efficient Hardware Trojan detection. Prevailing machine learning approaches based on circuit learning achieve good performance but face some critical challenges. The most pertinent challenges being the complex feature extraction, time-consuming training process, and compromised performance metrics due to high data imbalance in the training dataset. The proposed work aims to overcome the aforementioned challenges. The work proposes a voting ensemble method, which detects Hardware Trojans from the gate-level netlist. Firstly, 15 generic features are extracted from the circuit netlist, which leverages the rich circuit knowledge present in synthesis reports. An ensemble of Random Forest Classifiers is trained and soft voting is performed to get the final prediction. The Random Forest Classifiers are class-weighted and are so configured to handle the effect of data imbalance. Finally, testing of the model is performed using Leave-One-Out Cross Validation. Experimental results on Trust-Hub circuits show that the proposed methodology achieves 99.73% accuracy, 98.25% recall, 96.81% precision, 97.44% F-measure, and 99.97% ROC-AUC score on an average of 16 circuits under test (CUT).

V. Sankar (✉) · M. Nirmala Devi

Department of Electronics and Communication Engineering, Amrita School of Engineering,
Amrita Vishwa Vidyapeetham, Coimbatore, India
e-mail: s_vaishnavi@cb.students.amrita.edu

M. Nirmala Devi
e-mail: m_nirmala@cb.amrita.edu

Keywords Hardware trojans · Random forest classifier · Weighted ensemble method · Cross validation · Circuit under test

1 Introduction

The immense growth in traffic in mobile communication increased development of new services, and massive connection requirements have led to the emergence of 5G communication. Cyber-physical systems have made a huge leap in terms of growth. This in turn facilitated the growth of Internet of Things (IoT) [1]. Integrated Circuit (IC) forms the core component of the hardware. The booming growth in demand is met by outsourcing some parts of the design or the manufacturing process itself to off-shore foundries, thus making Hardware threats an inevitable part of IC design. Hardware threats are of five types, namely Intellectual Property (IP) piracy, IC over-building, counterfeiting, reverse engineering, and Hardware Trojan [2]. IP piracy is the process of acquiring the functionality of the IP and selling it in the black market; IC over-building is the process of procuring the design adding additional functionalities and then selling it illegally; IC counterfeiting happens when the un-trusted design house sells without proper authorization by producing more than the required number of ICs. One of the most sought-after hardware security threats is the Hardware Trojan attack. Malicious modifications made in a circuit are known as Hardware Trojans. Despite its tiny structure, a Hardware Trojans can have catastrophic effects on the IC. The effects a Hardware Trojan can have on the IC depend on the motive of the attacker. The effects of Hardware Trojans can be broadly classified into four types: leakage of information, change of functionality, degradation of performance, and denial of service [2].

Each step involved in IC design faces the threat of Hardware Trojan insertion. The process of IC design can be mainly divided into two: design process and fabrication process. In the design step, the IC design uses third-party Intellectual Properties (IP) that are prone to the insertion of Hardware Trojan. Third-party vendors can make changes in the design by simple but efficient code modifications. Logic synthesis is then performed on the design. The process involves the use of third-party tools and hence the tools can be manipulated by third-party vendors. The entire design step is extremely vulnerable to Hardware Trojan insertion. Hence Hardware Trojan remains an impending issue. Hardware Trojan insertion in the fabrication phase requires expertise knowledge; this in turn causes attackers to focus on the design stage. The research community has made great advancements in the development of Hardware Trojan detection methods. Hardware Trojan detection methodologies can be broadly categorized into pre-silicon and post-silicon detection methods. Post-silicon detection techniques include destructive reverse engineering, side-channel analysis, and logical testing. Destructive reverse engineering (also known as Physical inspection or optical reverse engineering) consists of inspection of the Integrated Circuit (de-packaging, de-metallization, micro-photography) through reverse engineering. Side-channel analysis is designed to measure and analyze the side-channel

parameters of a device, such as quiescent current or path delay, to discover manipulations made to the IC. Logical testing triggers the Trojan to observe its effect on the output. The optimal test vectors are obtained using various approaches such as the method suggested in [17] or using transition probability [18]. Post-silicon methods are sensitive to process variation noise, require a golden circuit, and are not an appropriate choice when IC is intensely integrated. Pre-silicon approaches are suitable in such cases. Pre-silicon approaches detect Hardware Trojans that are inserted in the design step. With high levels of IC integration, Machine learning approaches [12–15] exhibit better performance. As a result, machine learning approaches have emerged as the state-of-the art methodology. The existing machine learning approaches face four major challenges.

- Circuit learning for feature extraction.
- Complex feature extraction process.
- High data imbalance in dataset causing high False Positives (FP) and False Negatives (FN).
- Unavailability of a generic Hardware Trojan Detection methodology.

The work aims to solve the aforementioned issues. The work develops a generic feature extraction methodology by choosing the features that are commonly impacted by Trojan insertion. The new developed methodology handles the problem of data imbalance using the Weighted Ensemble method. The results show the effect of data imbalance is significantly reduced leading to higher performance metric values. The following are the significant contributions:

- The features are generic.
- Simple feature extraction process.
- Usage of the Weighted Ensemble technique and Leave-One-Out Cross Validation to handle imbalanced dataset.
- Improved model performance metrics. The proposed method achieves 99.73% accuracy, 98.25% recall, 96.81% precision, 97.44% F-measure, and 99.97% ROC-AUC score on an average of 16 circuits under test (CUT).

2 Related Work

The section aims to establish the current status of research pertaining to Hardware Trojan detection in general. The discussion then progresses to focus on gate-level netlist Hardware Trojan detection methods based on machine learning.

2.1 *Hardware Trojan Detection*

Numerous Hardware Trojan detection methods have been developed during the past decade. Hardware Trojan detection techniques can be broadly categorized into two

categories: pre-silicon detection techniques and post-silicon detection techniques. The pre-silicon techniques can be further categorized into dynamic and static detection techniques. The post-silicon techniques include physical inspection, logic testing, and side-channel analysis.

2.2 *Pre-Silicon Detection*

2.2.1 Dynamic Detection

Reference [9] deploys machine learning to detect Hardware Trojans from the gate-level netlist. Salmani [10] proposed a method to detect Hardware Trojans using SCOAP measures extracted from the netlist. The method was highly accurate and robust. The only disadvantage was that the k-means clustering model is sensitive to the initial clustering center, affecting the detection rate.

2.2.2 Static Detection

Reference [11] marks suspicious circuits in the netlist by performing scoring mechanism or trust verification. The circuit then undergoes Formal verification or is functionally simulated to identify if the suspicious circuit is a Hardware Trojan. The methods exhibited poor generalization capability. Chen and Liu [12] proffered a Hardware Trojan identification method that targeted single-triggered Hardware Trojans. Structural features based on the characteristics of the gate-level circuit structure were deployed for Hardware Trojan detection by the machine learning model. This method had benefits of low false positive rate and low overhead, but failed to identify Hardware Trojans with a different structure of the Trojan.

2.3 *Post-Silicon*

2.3.1 Physical Inspection

Physical inspection (also known as destructive reverse engineering or optical reverse engineering) consists of the examination of the manufactured chip (de-packaging, de-metallization, micro-photography) through reverse engineering. Reference [13] uses fast scanning electron microscope (SEM) to detect HTs. The method had the severe drawback of the destruction of the IC under test.

2.3.2 Side-Channel Analysis

Side-channel analysis is designed to measure and analyze the side-channel parameters of a device, such as quiescent current or path delay, to unleash manipulations made to the circuit. Chi et al. [14] use circuit characteristics and test vectors to detect HTs. The method accurately detected Trojans of 0.3% size but under-performed for smaller HTs.

2.3.3 Logical Testing

Logical testing triggers the Trojan to observe its effect on the output. The optimal test vectors are obtained using various approaches such as the method suggested in [17] or using transition probability [18].

2.4 Machine Learning Approaches

In general, machine learning used for Hardware Trojan detection consists of the following steps. First is the learning phase where the detection model is given features that efficiently distinguish normal nets from Trojan nets. The features are hand-crafted in other words formulated manually by analyzing the various Trojan templates present in Trust-Hub [15]. The extracted features are then split as training and testing sets, respectively. The training set is used to teach the model to distinguish between normal and Trojan nets, and the testing set is used to identify how well the model has learned to perform the aforementioned task. The above steps form the generic methodology of machine learning techniques. The following section discusses some of the prominent works that utilize ML to detect hardware Trojans at the gate level.

Chen and Liu [12] analyze various Trust-Hub circuits and propose nine features that efficiently differentiate the Trojan nets from the normal nets. The nets which have these nets are termed as weak classification nets. Based on the number of features inherited by a net, the scores were assigned and used for classifying Trojan nets. Another feature that was used was the maximum number of cycles for which a constant value holds for each net. The proposed work efficiently classified Trojan nets that formed part of the training set. The major drawback of the proposed work was that features chosen to describe Trojan behavior were too specific to the Trojan templates. This problem was solved by [13] which proposed five features that were generic to various Trojan templates. The proposed five features are LGFI, PI, PO, FF_i, and FF_o. Logic gate fan-in (LGFI) refers to the fan-in count of the net; Primary Input (PI) refers to primary input to the level the net is from the primary input; Primary Output refers to the level the net is from the primary outputs; FF_i refers to the number of flip-flops from the input side of the net; FF_o refers to the number of flip-flops from the output side of the net. The features were then given to a Support Vector Machine (SVM) to classify Trojan nets and normal nets. The model yielded good

results but with mediocre detection accuracy. In [14], the Trojan feature extraction was a two-step process to efficiently select features that aptly differentiated trojan nets from normal nets. At first, 51 trojan features were extracted. For each feature, F-measure was calculated. F-measure is the product of precision and recall divided by its sum. 11 features with high F-measure were selected. The 11 features are the following fan-in value of the net in the 5th and 4th levels from the net, number of flip-flops 4th level away from input side of net, number of flip-flops 3rd and 4th level away from the output side of the net, loops in and out 4th and 5th level away, the minimum level the net is from primary input, primary output, multiplexer, and flip-flop. The model yielded high accuracy on average. The proposed model faced the data imbalance issue which leads to a low true negative. Another approach [15] which used a multi-layer neural network for classification addressed the issue of data imbalance. The neural network had 11 inputs (11 features) which were the same as that of [14] and two outputs (Trojan or normal), and three middle layers with 200, 100, and 50 neurons, respectively. In order to overcome the issue of data imbalance, the leave-one-out cross-validation method was adopted for testing of the model.

3 Preliminaries

3.1 Gate-Level Netlist

Initially, the circuit which is depicted in Fig. 1 is present as a Verilog code. The Verilog code contains the complete description of the circuit. The Verilog code format is then converted to the netlist format using Synopsys dc [16]. The netlist on the other hand contains only the structural description of the circuit. As illustrated in Fig. 2, it describes the various gates and interconnections. The netlist starts with **module** and ends with **endmodule**. The module contains the inputs and outputs. Intermediate nets are termed as wires. Each gate is coded as the gate name, its technology file library name, followed by the output and input nets written within parentheses.

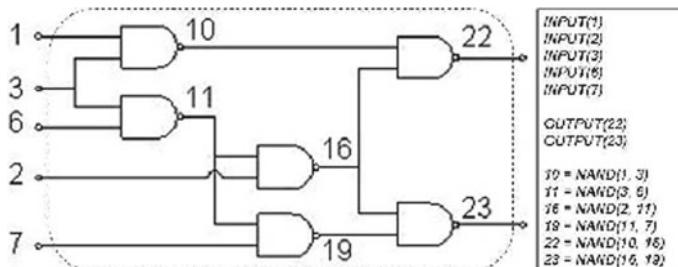


Fig. 1 Circuit diagram of c17 benchmark circuit

```

module c17 (N1,N2,N3,N6,N7,N22,N23);

input N1,N2,N3,N6,N7;

output N22,N23;

wire N10,N11,N16,N19;

nand NAND2_1 (N10, N1, N3);
nand NAND2_2 (N11, N3, N6);
nand NAND2_3 (N16, N2, N11);
nand NAND2_4 (N19, N11, N7);
nand NAND2_5 (N22, N10, N16);
nand NAND2_6 (N23, N16, N19);

endmodule

```

Fig. 2 Netlist example

3.2 *Hardware Trojan*

Hardware Trojans can be broadly categorized into two types: functional and parametric. In the former type, Trojans are realized by the addition or deletion of gates into or from the golden circuit. The latter type, the golden circuit, is manipulated by reducing the thickness of the connecting wire, subjecting the chip to radiation or weakening the flip-flops. In general, a Hardware Trojan consists of two parts: Trigger and the payload as shown in Fig. 3. In accordance with the motive of the attacker, the Hardware Trojans can exhibit any of the four effects, such as leakage of information, denial of service, degradation of performance, and change of functionality. They are tiny and stealthy in nature. Hence, Hardware Trojan detection is a challenging issue.

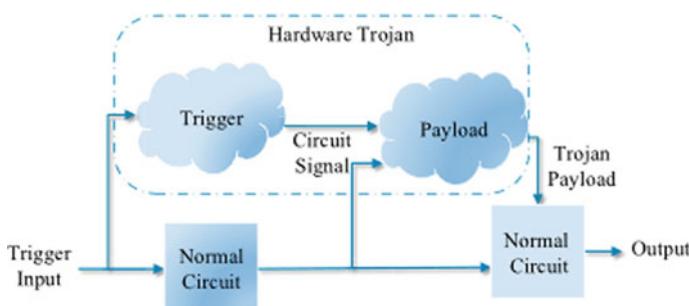


Fig. 3 Architecture of Hardware Trojan

3.3 Random Forest Classifier

It is an example of an ensemble learning method. A random forest [14] is a family of classifiers $h(x|\Theta_1), \dots, h(x|\Theta_k)$ with a decision tree as the base classifier. The parameters are Θ_k randomly chosen from a model of random vectors Θ . In the process of bagging, the original dataset is sampled with replacement into various bootstrap samples. These bootstrap samples are handled by different decision trees. The labels predicted by each tree are such that the miss-classified data is given higher weight. That is the generalization error, as given in Eq. 1, is minimized. The various decision trees are then bagged to arrive at a final decision. For the final classification, each tree votes for the most popular class for input x . The majority votes were taken as the final prediction.

$$e = P_x, y(\hat{m}_x, y < 0) \quad (1)$$

In Eq. 1, x is the input data and y is the labels, and P_x, y is the probability of prediction.

3.4 Weighted Ensemble

Ensemble methods are meta-algorithms that combine several machine learning approaches into a single predictive model. The models are combined and are termed as base learners. The method of combination of models produces three types of Ensemble methods as shown in Fig. 4. The different models have different outcomes such as bagging decreases variance, boosting reduces bias, and stacking improves predictions.

A weighted ensemble is where the week learners are given weights in accordance with its base models' individual performance on the dataset.

4 Methodology

The methodology as shown in Fig. 5 primarily consists of four steps: feature extraction, model training for class weighting, training and testing ensemble, and model evaluation using Leave-One-Out Cross Validation. In the feature extraction process, the circuit present in Verilog is converted to a netlist. Structural and Power reports are generated from the netlist. The Random Forest Classifier is trained and tested using the developed features, and the optimal class weights are obtained using an exhaustive grid search. A voting classifier is created using multiple configurations of Random Forest Classifiers. The model is used to classify Trojan nets and normal nets. The model performance is evaluated using Leave-One-Out Cross Validation.

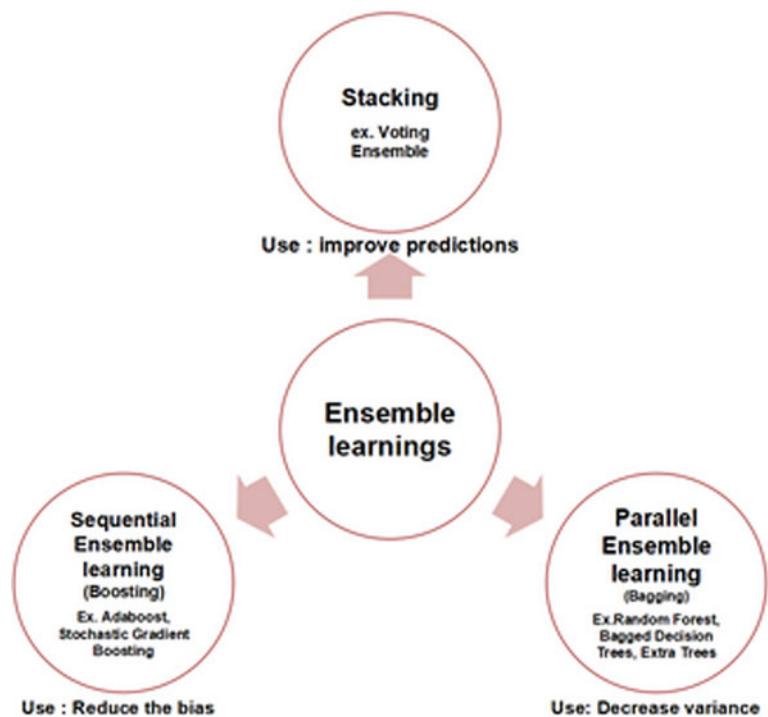
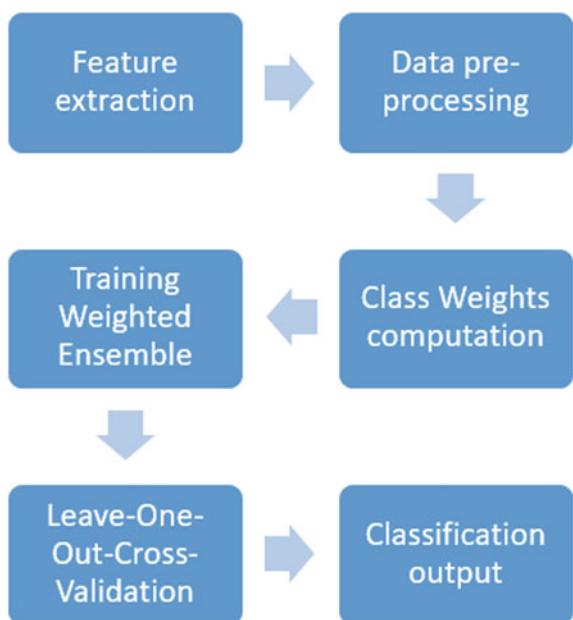


Fig. 4 Types of ensemble algorithms

Fig. 5 Methodology block diagram



4.1 Dataset

Circuits which are present in Verilog form in the Trust-Hub circuits [15] platform is converted to netlists using Synopsys Design Compiler (DC) [16]. Netlists is given as input to Synopsys IC compiler [17] to extract power and structural reports. The reports are then imported to a .csv file. The data is then pre-processed to create the final dataset.

4.2 Feature Extraction and Pre-processing

The circuit is read using Synopsys Design Compiler [16] and netlist is synthesized. The netlist is then used for feature extraction. Features for each net are extracted from the reports that are generated using Synopsys IC compiler[17]. Net features are selected according to their prominence (showing noticeable variation) and generic nature. Data pre-processing is performed on the obtained feature dataset, and the nets are labeled as 0 for normal nets and 1 for Trojan nets.

4.2.1 Features

The structural and power reports obtained from Synopsys IC compiler [17] contain the following features which are Fan-out, Fan-in, Capacitive load, Resist Pins, Rise transition, Fall transition, Static Probability, Toggle Rate, and Switching Power.

- **FAN-OUT:** The number of load gates connected to the output of the driving gate. Gates with large fan-outs are slower.
- **FAN-IN:** The number of inputs to the gate.
- **RESISTANCE:** The drive strength of a standard cell depends on output pull-up resistance also called output high-drive and output pull-down resistance also called output low-drive. This factor affects the overall system's timing. The addition or removal of gates affects path delay thereby causing variation in circuit timing.
- **PINS:** It is the sum of fan-out and fan-in of a net.
- **TOTAL NET LOAD:** The capacitive load of a net depends if the net is an input net or output net. For the former, capacitive load is input capacitance, and for the latter, the total capacitance is the sum of output capacitance of the driving cell, sum of all the capacitances of the wires, and sum of all the input capacitances of the cells it is driving. It affects the time the cell takes to switch from one logic state to another. The addition or removal of gates to a net affects the path delay to which the net belongs. This is in turn reflected in the load value of the net.
- **STATIC PROBABILITY:** It refers to the expected logic state a net will have. The value of this parameter can increase or decrease upon Trojan insertion.

- **TOGGLE RATE:** It is the rate at which the net switches from one logic state to another. It is also called the Transition rate. Trojans are normally inserted in nets having low toggle rates to go undetected in the conventional verification process.
- **SWITCHING POWER:** It is the power dissipated when charging or discharging internal net capacitances. The addition or removal of gates to a net affects the capacitive load of the corresponding net.
- **RISE AND FALL TRANSITIONS:** Rise transition is the time taken by the net to transit from logic low to logic high, and fall transition is the time taken to transit from a logic high to a logic low. The addition of a Trojan introduces additional delays to the aforementioned values.

5 Result and Discussion

5.1 Data Pre-processing

Data labeling is a manual process. The label assignment is in accordance with the Trojan information provided by the [15]. In the process of marking Trojan nets, the work considers all nets of the Trojan circuit as Trojan nets. The boundary network which is connected to the normal nets are also considered as Trojan nets. The Trojan nets are then labeled 1 and normal nets as 0.

5.2 Experiment Setup

Leave-One-Out Cross Validation is a special type of k-fold Cross Validation where k is 1. In LOOCV, out of the entire set of circuits all except one circuit is taken for training. In the next iteration, another circuit is kept aside for testing and the model is trained using the rest of the circuits. This process iterates until all the circuits in the set are considered as a testing circuit. In the experiment carried out, 15 circuits were taken as training sets and one circuit as the testing set. 16 iterations were carried out.

5.3 Model Selection and Parameter Setting

The implementation of the Random Forest Classifier has been performed using Python machine learning library scikit learn [18]. For the Random Forest Classifier the n_estimators is set to 200, and the impurity measure is gini. The ensemble methods have three configurations of Random Forest Classifier. One has the default Random Forest Classifier, the second configuration with the optimal class weights, and the final has a parameter n_estimator setting of 1500 and class weight of 11.

5.4 Performance Metrics

The model performance is evaluated using F-measure, Precision, Accuracy and ROC_AUC score. The following section explains the significance of each of the parameters. The following section explains the significance of each of the parameters, which are derived from confusion matrix as shown in Table 1.

- *True Positive Rate (TPR/Recall)*: The rate at which the true class (as per experiment, Trojan nets) is predicted as true class (Trojan nets). It is calculated as given by Eq. 2.

$$TPR = TP / (FN + TP) \quad (2)$$

In Eq. 2, True Positive (TP) gives the number of Trojan nets that are identified to be Trojan nets and False Negative (FN) gives the number of Trojan nets identified to be normal nets.

- *True Negative Rate (TNR)*: The rate at which the false class (as per experiment, Normal nets) is predicted as a false class. It is calculated as given by Eq. 3.

$$TNR = TN / (FP + TN) \quad (3)$$

In Eq. 3, True Negative (TN) gives the number of normal nets that are identified to be normal nets, False Positive (FP) gives the number of normal nets that are identified as Trojan nets.

- *Precision*: It is the proportion of Trojan nets that are rightly interpreted as Trojan nets. It is calculated as given by Eq. 4.

$$P = TP / (TP + FP) \quad (4)$$

In Eq. 4, True Positive (TP) gives the number of Trojan nets that are identified to be Trojan nets and False Positive (FP) gives the number of normal nets that are identified as Trojan nets.

- *F-measure*: Harmonic mean of precision and recall. It is calculated as given by Eq. 5.

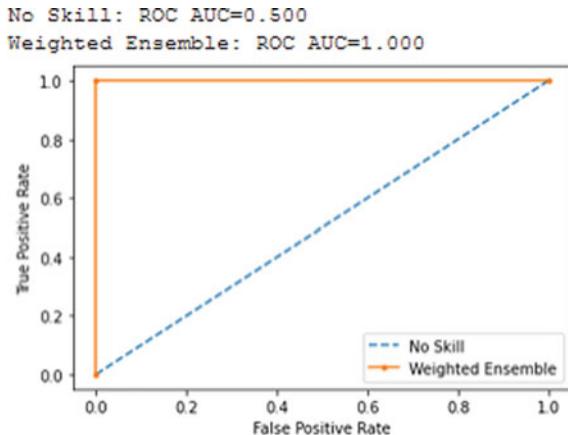
$$F = 2PR / (P + R) \quad (5)$$

In Eq. 5, P is precision and R is recall.

Table 1 Confusion matrix

Ground	Confusion matrix	
Truth	Trojan net	Normal net
Trojan net	True Positive (TP)	False Negative (FN)
Normal net	False Positive (FP)	True Negative (TN)

Fig. 6 ROC curve of s38584t200



Accuracy: It is the amount of correct prediction with respect to the total predictions. It is calculated as given by Eq. 6.

$$A = TP + TN / (TP + TN + FP + FN) \quad (6)$$

In Eq. 6, True Positive (TP) gives the number of Trojan nets that are identified to be Trojan nets, True Negative (TN) gives the number of normal nets that are identified to be normal nets, False Positive (FP) gives the number of normal nets that are identified as Trojan nets and False Negatives (FN) gives the number of Trojan nets identified to be normal nets. From Fig. 6, it can be seen that the ROC_AUC score compared to the no skill model achieves 100% ROC_AUC score, which indicates 100% detection of Trojan nets from circuit s38584t200. Figure 7 illustrates that the method achieves an ROC_AUC score in the range of 96%–100% which in turn shows the efficiency of the proposed Hardware Trojan detection method.

5.5 Comparative Analysis

From Table 3 compared to Table 2, it can be seen that the proposed method exhibits an increase in terms of Recall. This in turn shows the decrease in the presence of False Negatives, that is lesser number of Trojan nets are misinterpreted as normal nets. In the perspective of security, this misclassification needs to be reduced. In addition, in 10 Trojan circuits which accounts for about half of the total circuit exhibits 100% recall. It shows that the model has efficiently learned the Trojan nets using the 15 proposed features. Table 3 illustrates the evaluation results of the model in terms of Accuracy, Precision, Recall, and F-measure. Results show that despite the generic nature of the feature unlike that of [19] which is Trojan-template specific, the proposed techniques exhibit improved metrics. It shows that in more than 50% of

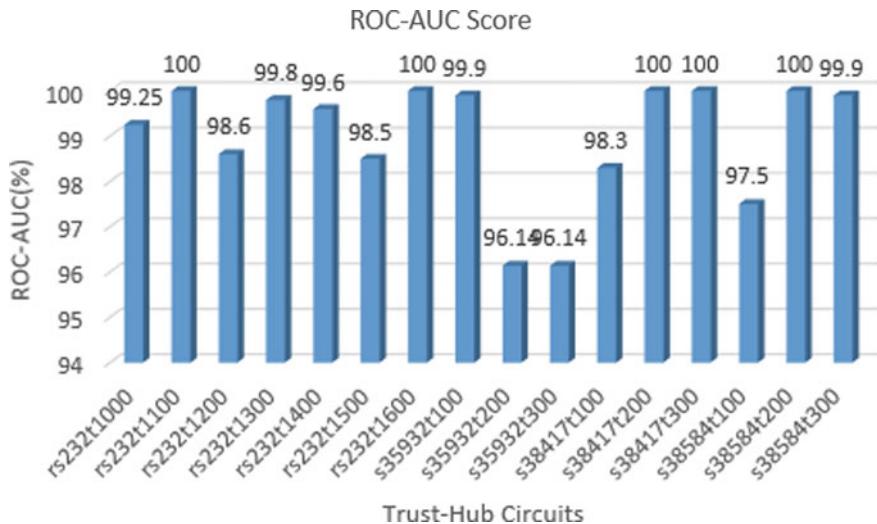


Fig. 7 ROC-AUC scores of Trust-Hub circuits

Table 2 Performance metrics results of [19]

Circuit name	Accuracy(%)	Precision(%)	Recall(%)	F-measure(%)
RS232T1000	99.04	81.82	90	85.71
RS232T1100	99.69	91.67	100	95.65
RS232T1200	100	100	100	100
RS232T1300	100	100	100	100
RS232T1400	100	100	100	100
RS232T1500	99.69	91.67	100	95.65
RS232T1600	100	100	100	100
S35932T100	99.98	100	92.31	96
S35932T200	99.84	100	16.67	28.57
S35932T300	99.94	92.31	97.3	94.74
S38417T100	99.81	50	100	66.67
S38417T200	99.97	100	81.82	90
Average	99.83.	92.29	89.4	87.75

the circuits, the F-measure lies between 97 and 100% which is indicative of the fact that the model does not show bias toward the majority class. Thereby the problem of data imbalance is dealt with efficiently using leave-one-out cross validation and class weighting. Table 4 is a comprehensive comparative analysis of the proposed scheme with the existing machine learning approaches. On comparison, the scheme

Table 3 Performance metrics results of the proposed method

Circuit name	Accuracy(%)	Precision(%)	Recall(%)	F-measure(%)
RS232T1000	98.7	92	100	96
RS232T1100	100	100	100	100
RS232T1200	99.4	98	98	98
RS232T1300	99.7	97	100	98
RS232T1400	99.4	96	100	98
RS232T1500	99	96	98	97
RS232T1600	100	100	100	100
S35932T100	99.9	87	100	93
S35932T200	99.9	95	92	94
S35932T300	99.9	95	92	94
S38417T100	99.9	100	97	98
S38417T200	100	100	100	100
S38417T300	100	100	100	100
S38584T100	99.9	100	95	97
S38584T200	100	100	100	100
S38584T300	99.9	93	100	96
Average	99.73	96.81	98.25	97.44

Table 4 Performance metrics results of the proposed method

Metrics	[9]	[8]	[19]	Proposed method(%)
Accuracy	89.7	90.6	99.83	99.73
Precision	88	92	92.29	96.81
Recall	90.7	86.6	89.4	98.25
F-measure	90.6	98	87.75	97.44

produced the highest metric value in terms of accuracy, precision, and F-measure. Figure 7 analyzes models' performance in terms of ROC-AUC scores and shows that the model can efficiently classify normal and trojan nets.

6 Conclusion

Machine learning has been showing significant growth in all the fields it has been applied to. Lately, it has emerged as the state-of-the-art Hardware Trojan Detection scheme. The proposed scheme proposes a generic Hardware Trojan detection scheme using the Weighted Ensemble classifier. The scheme proposes a new set of 15 generic

features for feature extraction. Weighted Ensemble and leave-one-out cross validation have been adopted to reduce the impact of the skewed dataset. The proposed methodology achieves 99.73% accuracy, 98.25% recall, 96.81% precision, 97.44% F-measure, and 99.97%, ROC-AUC score on an average of 16 circuits under test (CUT). 97.44% F-measure shows that the model has effectively handled the data imbalance issue. In future, further enhancement of the model is targeted by developing a feature set. Also developing models suitable to handle skewed datasets is targeted.

References

1. Guo, Zhu, W., Yu, Z., Wang, J., Guo, B.: A survey of task allocation: contrastive perspectives from wireless sensor networks and mobile crowd sensing. *IEEE Access*, vol. 7, pp. 78406–78420 (2019)
2. Tehranipoor, Mohammad, Koushanfar, Farinaz: A survey of hardware trojan taxonomy and detection. *IEEE Des. Test Comput.* **27**(1), 10–25 (2010)
3. Devi, N.M. et al.: Detection of malicious circuitry using transition probability based node reduction technique. *Telkommika* **16**.2, 573–579 (2018)
4. Ranjani, R.S., Devi, M.N.: Golden-chip free power metric based hardware trojan detection and diagnosis. *Far East J. Electron. Commun.* **17**, 517–530 (2017)
5. Oya, M., Shi, Y., Yanagisawa, M., Togawa, N.: score-based classification method for identifying hardware-trojans at gate-level netlists. In: 2015 Design, Automation & Test in Europe Conference & Exhibition (DATE), pp. 465–470. IEEE (2015)
6. Hasegawa, K., Oya, M., Yanagisawa, M., Togawa, N.: Hardware trojans classification for gate-level netlists based on machine learning. In: 2016 IEEE 22nd International Symposium on On-Line Testing and Robust System Design (IOLTS), pp. 203–206. IEEE (2016)
7. Hasegawa, K., Yanagisawa, M., Togawa, N.: Trojan-feature extraction at gate-level netlists and its application to hardware-Trojan detection using random forest classifier. In: 2017 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1–4. IEEE (2017)
8. Hasegawa, K., Yanagisawa, M., Togawa, N.: Hardware Trojans classification for gate-level netlists using multi-layer neural networks. In: 2017 IEEE 23rd International Symposium on On-Line Testing and Robust System Design (IOLTS), pp. 227–232. IEEE (2017)
9. Hicks, M., Finnicum, M., King, S.T., Martin, M.M.K., Smith, J.M.: Overcoming an untrusted computing base: Detecting and removing malicious hardware automatically. In: Proceedings of the IEEE Symposium on Security and Privacy, pp. 159–172, May 2010
10. Salmani, H.: COTD: reference-free hardware trojan detection and recovery based on controllability and observability in gate-level netlist. *IEEE Trans. Inf. Forens. Secur.* **12**(2), 338–350 (2017)
11. Zhang, Yuan, F., Wei, L., Liu, Y., Xu, Q.: VeriTrust: verification for hardware trust. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **34**(7), 1148–1161 (2015)
12. Chen, F., Liu, Q.: Single-triggered hardware Trojan identification based on gate-level circuit structural characteristics. In: Proceedings of the IEEE International Symposium on Circuits and Systems, pp. 1–4, May 2017
13. Courbon, F., Loubet-Moundi, P., Fournier, J.J.A., Tria, A.: A high efficiency hardware trojan detection technique based on fast SEM imaging. In: Proceedings of the Design, Automation & Test in Europe Conference & Exhibition (DATE), pp. 788–793, Mar 2015
14. Chi, G., Yu, Z., Zhou, Y., Lei, S.: A hardware trojan detection method based on region segmentation technology. *Semicond. Technol.* **42**(7), 555–560 (2017)
15. www.trusthub.org

16. Simulator, Synopsys. Design Compiler Reference Manual (1994)
17. Lin, Z-M.: DAVE: an automatic mixed analog/digital IC layout compiler. In: Proceedings of the IEEE 1991 Custom Integrated Circuits Conference, pp. 5–4. IEEE (1991)
18. Scikit, S.-L.D.: Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
19. Dong, Chen., Chen, Jinghui., Guo, Wenzhong, Zou, Jian: A machine-learning-based hardware-Trojan detection approach for chips in the internet of things. *Int. J. Distrib. Sens. Netw.* **15**(12), 1550147719888098 (2019)

An Approach for Offline Handwritten Character Shape Reconstruction Using Active Contour and Morphological Techniques



Anupam Garg, Amrita Kaur, and Anshu Parashar

Abstract Optical Character Recognition (OCR) models have increased their popularity in the past decade. The accuracy of OCR models has been increased with the new research models, which have improved the recognition rate of the OCR algorithms. In this work, the improved model for character shape restoration practices of handwritten documents is proposed. The proposed model involves the active contour selection (ACS) model to detect object locations, where the shape restoration method can be applied. The contour selection model utilizes the snake model for marking and localization of the characters in all possible shapes using polygonal point marking for the extraction of the object boundary. Then the distantia factor is applied over the detected contour regions to analyze the distance between two contour regions, which decides upon the selection of the target regions, where the opening needs to be closed. The combination of adaptive dilation and adaptive erosion has been utilized for filling or closing of the target region, which verifies the angle of the contour objects and thickness for removal of the smudge regions. The proposed model has undergone various experiments to assess the proposed model quality against the existing models. The proposed model has been found efficient and improved in terms of obtained performance parameters of peak signal-to-noise ratio (PSNR), mean squared error (MSE) and relative error. The proposed model has been recorded with 0.0006 percent relative error compared to 0.03, 0.26 and 0.03 in the case of the ring radius method, restoration through the medial axis and iterative midpoint method, respectively, which shows the robustness of the proposed model.

Keywords Image processing · Pattern recognition · OCR · Gurmukhi script · Multi-lingual character recognition · Broken character reconstruction

A. Garg · A. Kaur · A. Parashar (✉)

Computer Science and Engineering Department, Thapar Institute of Engineering and Technology, Patiala (147001), Punjab, India

e-mail: aparashar@thapar.edu

1 Introduction

OCR converts the printed copies to the corresponding text form, which can be effectively managed and processed for further use. The scanning process converts the hard copy of the sheet of paper to the digital form of an image. If the document contains a combination of image and text, then the whole document can be manipulated, but it does not manipulate the text present in it separately. So, the OCR technology can be utilized for the storage and manipulation of the information. The most important factor for analyzing the performance of the text recognition system is character accuracy which depends upon the quality and nature of the image. The OCR for Gurmukhi script (script used for the language spoken by Punjab state natives of India) become the research interest area in the 2000s when the OCR for offline handwritten Gurmukhi script has been evolved. Now, the popularity of OCR is increasing each year with new techniques leading to a better recognition rate [1]. The various applications of the OCR systems are receipt OCR, invoice OCR, check OCR and legal billing document OCR.

The prominent stages of OCR can be described as follows:

- *Preprocessing*

The input for the preprocessing stage is the scanned image on which a series of operations are performed to enhance the quality of the image to make it suitable for the next stage, i.e., segmentation. The different approaches are applied such as binarization, noise removal, alignment correction, hole filling and skew correction. In this, the binarization process generally converts a grayscale image into a binary image [2]. The noise removal techniques are used to remove the unwanted foreground pixels on the background. The alignment correction removes the presence of horizontal and vertical slants in the written text.

- *Segmentation*

It is the process to segment the region of interest from an image. In offline Gurmukhi handwritten text, the segmentation focuses on the partitioning of text into lines, then words and further into individual characters [2]. There are several approaches being used by various researchers to find character bounds.

- *Feature extraction*

Extracting the efficient features is a crucial step to classify the segmented region. It is a critical stage whose effectiveness improves the recognition rate and reduces misclassification [2].

- *Recognition*

The last step of OCR is recognition where recognition of the character is done. Efficient preprocessing, segmentation and feature extraction lead to a reasonable or



Fig. 1 The diagram depicts the removal of gaps in the broken character

higher recognition rate. Recognition rate generally lies between 80 and 99% to get efficient OCR.

1.1 Motivation and Problem Formulation

Recognition of the broken characters is a challenging problem in the field of Optical Character Recognition. Due to its assumption of having a negligible effect on the system's performance, it is often neglected in OCR. Any standard OCR system's performance working for acceptable documents decreases if tested on a document containing broken characters. The characters are generally degraded due to folds in documents, color degradation, water effect, etc., and receive the white cracks on various parts of the characters. The recognition of broken parts of the character hides from OCR systems. Numerous OCRs for the recognition of finely printed characters have been designed without any imperfections. But still the recognition of the degraded characters is unexplored due to its complexity. If the input image is degraded, then the efficiency of the OCR system decreases. In old documents, there is a prominent presence of broken characters at several places [3]. In this after dilation, generally, some Gurmukhi characters have large gaps that lead to a low recognition rate. To improve this, there is a need to fill these large gaps that lead to a better recognition rate. Hence, the present work focuses on removing larger gaps in the broken characters of the Gurmukhi script using contour selection to enhance the accuracy of further phases. The impact of the removal of larger gaps has been depicted in Fig. 1.

1.2 Major Contribution

- To implement the character reconstruction algorithm to fill the gaps of Broken Gurumukhi characters.
- The performance of the proposed algorithm is evaluated using three parameters such as Mean Squared Error (MSE), Relative Error and Peak Signal-to-Noise Ratio (PSNR).
- To compare the proposed methodology with the existing state-of-the-art techniques.

The paper is organized as follows. Section 2 describes the relevant literature, Sect. 3 discusses the methodology, Sect. 4 describes the experimental setup and results, and finally, Sect. 5 discusses the conclusions and future work.

2 Related Works

In Poovizh [2] the main objective is to explain the importance of the preprocessing steps for Optical Character Recognition. This paper shows that preprocessing generally improves the character recognition rate. Singla [7] presented preprocessing techniques to improve the quality of input images written in offline handwritten Gurmukhi script. The work proposed the Gmean filter for noise removal followed by hole filling and brokerage filling algorithms. But in this work, a problem of large gaps in Gurumukhi characters is found out as a limitation. Kumar et al. [10] explored different preprocessing techniques applied to various types of image ranges, from basic handwritten form-based documents to colorful and complex context documents with variable intensities. The other crucial preprocessing techniques which can be used for the enhancement of the images such as contrast stretching, binarization, noise removal techniques, morphological processing, normalization and segmentation are also analyzed. Ntirogiannis et al. [11] documented the importance of image binarization in the pipeline of document image analysis and recognition as it affects the performance of the recognition process in further stages. To study the algorithmic behavior, the evaluation of binarization methods aids in it. The effectiveness is also verified by analyzing its performance both qualitatively and quantitatively. Bieniecki et al. [12] described the digital cameras as fast, cheap and versatile image acquisition devices. But, the digital cameras endure constraints such as geometrical distortions that affect the performance of OCR applications. So, [12] performs the preprocessing on the images from a digital camera before the text recognition and experimental evaluation confirm the significance of preprocessing in OCR systems.

Thapar et al. [13] presented a brief introduction to morphology and its operations. Alginahi [15] focused on decreasing the variations present in the input images that lead to the recognition rate reduction and in complexities present with the use of preprocessing techniques. The data is transformed into a more effectively processed format. Shivakumara et al. [4, 6] introduced a reconstruction technique in video images for the improvement of character recognition rate by using novel Ring Radius Transform (RRT) and the concept of medial pixels on characters with broken contours in the edge domain. The inner and outer contour symmetry information is exploited to reconstruct the gaps of the broken character that outperforms the state-of-the-art methods in terms of relative error and character recognition rate.

Tian et al. [5] proposed an approach in finding the medial axis point from a given input character with the use of histogram gradient division and reverse gradient orientation in all directions. Then, the gaps in the video text are filled by extracting the candidate medial axis point to improve the recognition rate. Thilagavathy et al. [8]

presented the analysis of the inconsistent shape and irrespectively distorted characters. Gatos et al. [3] presented an adaptive approach with the amalgamation of various existing binarization techniques and effective edge information of grayscale input images. The proposed approach outperforms well-known approaches which are experimented on degraded machine-printed and handwritten documents. Raman Maini [9] analyzed and compares various techniques of edge detection, showing the outperformance of Canny's edge detection algorithm in all scenarios in comparison to other operators. But there is a limitation observed for Canny's edge detection algorithm that it is computationally expensive as compared to others. Maduria et al. [14] analyzed different edge detection approaches used for image segmentation using several computing approaches. Droettboom [16] proposed the graph-combinatorics-based technique to join the relevant broken character components in the historical printed text. Novak et al. [17] have worked on the distortions present in the text printed on the metal ingot due to the uneven surface of the ingot for improved recognition by using fuzzy logic and human-inspired approach. Kaur and Bathla [18] have studied different techniques used for the segmentation of broken and touching characters written in Gurmukhi script. Yu and Yan [19] have implemented the dilation algorithm for the reconstruction of handwritten digits.

3 Proposed Methodology

The proposed methodology combines novel image reconstruction techniques to improve the quality of the text in the text documents by filling up the cracks and line gaps. The proposed model performs two significant tasks:

1. Contour Selection;
2. Image Reconstruction.

The flowchart for the proposed methodology is shown in Fig. 2.

It uses Canny edge detection and the total variance method for selecting the contour of a broken part. After selecting the contour of a broken part, the image reconstruction tasks would be performed using the effective combination of the erosion and dilation techniques to reconstruct the brokage characters of text documents. The current work has used an active contour model for selecting the contour of a broken part of the image. After selecting the contour of a broken part, the image processing techniques, i.e., adaptive erosion and adaptive dilation techniques with disk shape structuring element are used for joining the parts of the broken characters, so that the shape of the characters can be maintained.

The technique has been designed around the reconstruction of the degraded text image documents, which contain the handwritten text in medium to a larger size. It involves multiple algorithms to produce the hybrid solution for the reconstruction. The idea has been designed around the pixel to pixel-based analytical models for the assessment of the missing pixels.

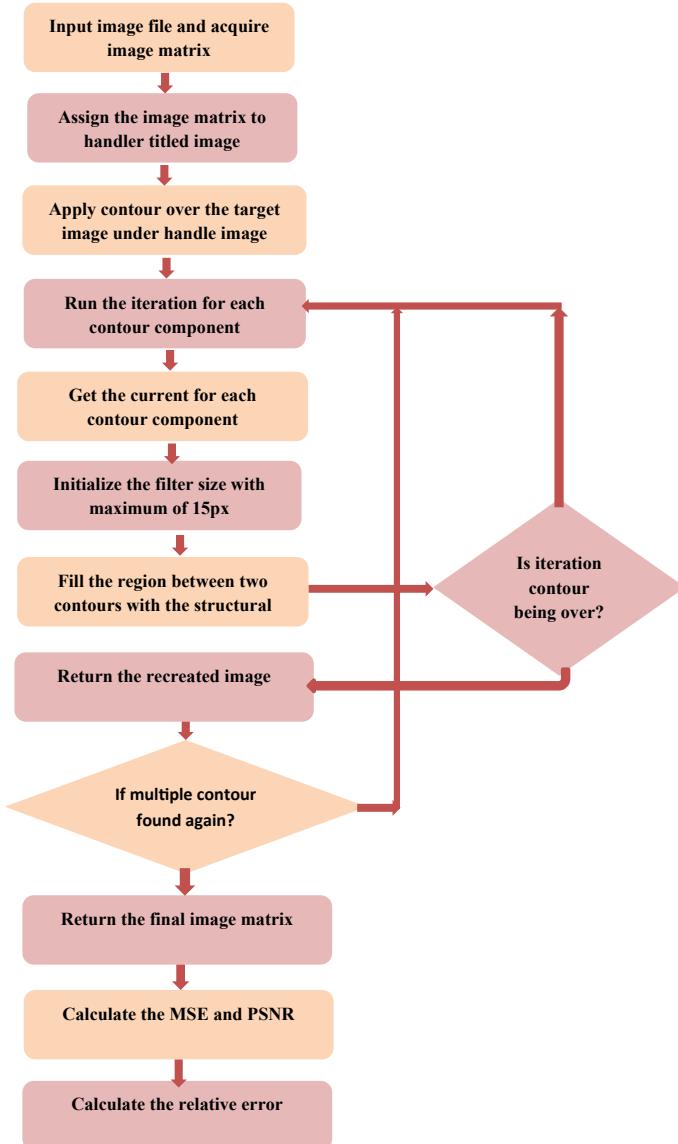


Fig. 2 The flowchart for the proposed methodology

4 Result and Discussion

The experimental analysis of the proposed model involves the various performance parameters, which includes the mean squared error, peak signal-to-noise ratio and relative error. PSNR generally defines the similarities of the input and output images,

whereas MSE defines the difference of the input and output images in term of pixels. MSE is the cumulative squared error between the compressed and original images, whereas PSNR is a measure of the peak error. The mathematical formulas for the two are

$$\text{MSE} = \frac{1}{MN} \sum_{y=1}^M \sum_{x=1}^N [I(x, y) - I'(x, y)]^2 \quad (1)$$

$$\text{PSNR} = 20 * \log_{10} (255 / \sqrt{\text{MSE}}) \quad (2)$$

where $I(x,y)$ is an original image, $I'(x,y)$ is the approximated version and M, N are the dimensions of the images.

To measure the quality of the reconstructed character, relative error is used. Relative Error is the absolute error divided by the magnitude of the exact value. The proposed algorithm has been tested qualitatively as well as quantitatively on grayscale images. The performance of the proposed algorithm has been conducted on some test images. To each test image, the character reconstruction algorithm is performed. This technique is compared with the existing state-of-the-art techniques.

The results are compared qualitatively and quantitatively using PSNR, MSE and Relative error as quality metrics. The PSNR and MSE values define the similarities and differences of the input and output images, whereas relative error defines the efficiency of the output image. The proposed approach exhibits the best performance concerning other techniques and gives effective results even for English and Hindi broken characters. The contour selection and reconstructed images of the input images for the proposed algorithm are shown in Table 1. The experimental analysis has been obtained from the image dataset containing the text from multiple languages that are evaluated on performance metrics as shown in Tables 2 and 3.

Table 1 shows the input images written in Gurmukhi, English and Devanagari scripts with brokage in the characters. It also gives the output of the contour selection technique, used to select the broken parts in the characters that are needed to be reconstructed. It provides the output of broken character reconstruction done to maintain the shape of the character to avoid information loss.

The maximum mean squared error (MSE) value has been recorded nearly at 0.1370. In contrast, the minimum value has been recorded at 0.0064 as per the results obtained from the experimental setup for image text reconstruction. This shows the robust performance of the proposed model in handling the image data and undergoing regeneration based upon the single and multiple pixel regeneration techniques. The maximum value of peak signal-to-noise ratio (PSNR) has been recorded nearly at 45.979 decibels, which defines the higher accuracy of the proposed image reconstruction model. The minimum value of the PSNR has been recorded at 32.695 decibels, which again describes the efficient minimum performance obtained from the proposed model, which justifies the efficiency of the proposed model in the regeneration of the missing pixels for the reconstruction of the character shapes.

Table 1 Output images of contour selection and character reconstruction techniques applied on broken parts of the input image

Input image	Contour selection of broken part	Character reconstruction technique
धन्ता	धन्ता	धन्ता
माहिराम	माहिराम	माहिराम
तुरा	तुरा	तुरा
शहरज़	शहरज़	शहरज़
Hello	Hello	Hello
Bye	Bye	Bye
हन्ता	हन्ता	हन्ता
आप	आप	आप

The results of PSNR and MSE have been recorded and shown in Table 2, which shows the MSE and PSNR values of the various images of the database that show that the proposed algorithm leads to good MSE and PSNR values to give effective results.

It also shows the relative error of the various images of the database that show that the proposed algorithm leads to less relative error to give effective results. The minimum value of 1.17×10^{-5} has been recorded from the experimental results, and

Table 2 Performance metrics (MSE, PSNR and Relative Error) values of output image

Output image	MSE value	PSNR value	Relative error
Bada.bmp	0.0268	39.778	5.094007E-05
Avikash.bmp	0.0064	45.979	1.26622E-05
Hunda.bmp	0.0276	39.641	5.129402E-05
Aarti.bmp	0.00745	45.339	1.17013E-05
Hello.png	0.0395	38.097	6.68066E-05
Bye.png	0.1370	32.695	0.00023
Hum.png	0.1043	33.881	0.00018
Aap.png	0.0747	35.331	0.00012

Table 3 Comparison with the state-of-the-art techniques based on the relative error

Method	Relative error
Ring radius transform method [4]	0.03
Restoration through medial axis [5]	0.26
Iterative Midpoint Method [6]	0.03
Proposed Method	0.0006

the maximum relative error value has been recorded nearly at 0.00023. Table 3 shows the comparison of the proposed work with different algorithms by using relative error as analysis metrics which show that the proposed algorithm leads to less relative error to give effective results. The computed relative error is less than 0.0294 which indicates that the proposed model is performing better as compared to the existing techniques. It concretely elaborates the improved performance of the proposed model, which has been found the relative error, less by the value of 0.0294, which shows it highly improved in comparison than the existing techniques.

5 Conclusion and Future Scope

The character shape restoration (CSR) model has been developed using the active contour model with adaptive erosion and dilation techniques for the restoration of the broken handwritten characters. The proposed model is aimed at improving the accuracy of the OCR models for different languages such as Gurmukhi (Punjabi), Hindi, English, etc. The proposed model design incorporates the active contour technique based upon the snake algorithm for polygon shape extraction; it uses a combination of dilation and erosion with an adaptive design for automatic character restoration and contour shape and angle estimation using the distantia factor for the detection of the opening regions in the text in the input image matrix. The image dataset contains the minor to severe variations in the case of openings or gaps in the given text. The performance has been analyzed under the various performance parameters, such as peak signal-to-noise ratio (PSNR), mean squared error (MSE) and relative

error. The readings of PSNR range from 32 to 46 decibels, whereas the MSE has been recorded between 0.0064 and 0.137. The relative error has been recorded in the range of 1.17×10^{-5} – 2.30×10^{-4} . The result comparison shows the robustness of the proposed model with 0.0006 percent relative error value against the minimum relative error value of 0.03 in the ring radius transform method and iterative midpoint method. The proposed model can be further improved by using the artificial intelligence models such as neural networks, convolution neural networks and other deep learning models. Additionally, the proposed model can also be further improved by using swarm optimization.

References

1. Bhagavati, C., Ravi, T., Kumar, S.M., Negi, A.: On developing high accuracy OCR system for Telugu and other Indian scripts. *IEEE Expl. Organ.* (2002)
2. Poovizh, P.: A study on preprocessing techniques for the character recognition. *Proc. Int. J. Open Inf. Technol.* **2**(12), 21–24 (2014)
3. Gatos, B., Pratikakias, I., Perantonis, S.J.: Improved document image binarization by using a combination of multiple binarization techniques and adapted edge information. In: *Proceedings of the 19th International Conference on Pattern Recognition ICPR* (2008)
4. Shivakumara, P., Phan, T.Q., Bhowmick, S., Tan, C.L., Pal, U.: A novel ring radius transform for video character reconstruction. *Patt. Recog.* **1**(46), 131–140 (2013)
5. Tian, S., Shivakumara, P., Phan, T.Q., Lu, T., Tan, C.L.: Character shape restoration system through medial axis points in video. *Neurocomputing* **161**, 183–198 (2015)
6. Shivakumara, P., Hong, D.B., Zhao, D., Tan, C.L., Pal, U.: A new iterative-midpoint-method for video character gap filling. In: *Proceedings of the 21st International Conference on Pattern Recognition*, pp. 673–676
7. Singla, A.: A novel approach of noise removal in offline handwritten gurumukhi words, M.Tech Thesis, Punjab Technical University, Jalandhar, India (2015)
8. Thilagavathy, S.K., Gandhi, R.I.: Recognition of distorted character using edge detection algorithm. *Int. J. Innov. Res. Comput. Commun. Eng.* **1**(4), 1056–1061 (2013)
9. Maini, R., Aggarwal, H.: Study and comparison of various image edge detection techniques. *Int. J. Image Proc. (IJIP)* **3**(1), 1–12 (2009)
10. Kumar, G., Bhatia, P.K.: Analytical review of preprocessing techniques for offline handwritten character recognition. *Int. J. Adv. Eng. Sci.* **3** 14–22 (2013)
11. Ntirogiannis, K., Gatos, B., Pratikakias, I.: Performance evaluation methodology for historical document image binarization. *IEEE Trans. Image Proc.* **22**, 595–609 (2013)
12. Bieniecki, W., Grabowski, S., Rozenberg, W.: Image preprocessing for improving ocr accuracy. In: *Proceedings of the International Conference on Perspective Technologies and Methods in MEMS Design*, pp. 75–80 (2007)
13. Thapar, S., Garg, S.: Study and implementation of various morphology based image contrast enhancement techniques. *Int. J. Comput. Bus. Res.* (2012)
14. Maduria, V.B., Vydehi, S.: Edge detection techniques using character segmentation and object recognition. *Int. J. Sci. Res. (IJSR) India* **2**(1), 523–526 (2013)
15. Alginahi, Y.: Preprocessing techniques in character recognition. INTECH Open Access Publisher (2010)
16. Drotetboom, M.: Correcting broken character in the recognition of historical printed document. In: *Proceedings of the Joint Conference on Digital Libraries*, Houston, pp. 364–366 (2003)
17. Novak, V., Hurtik, P., Habiballa, H.: Recognition of distorted characters printed on metal using fuzzy logic methods. In: *Proceedings of the Joint IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS)* (2013)

18. Kaur, K., Bathla, A.K.: A review on segmentation of touching and broken characters for handwritten gurmukhi script. *Int. J. Comput. Appl.* **120**(18), 13–16 (2015)
19. Yu, Donggang, Yan, Hong: Reconstruction of broken handwritten digits based on structural morphological features. *Patt. Recogn.* **34**(2), 235–254 (2001)

Stock Direction Prediction Using Sentiment Analysis of News Articles



Nipun Jain, Mohit Motiani, and Preeti Kaur

Abstract It is widely acknowledged that stock price prediction is a job full of challenges due to the highly unpredictable existence of financial markets. Many market participants or analysts, however, attempt to predict stock prices using different mathematical, econometric, or even neural network models in order to make money or understand the nature of the equity market. In the past few years, a lot of models based on deep learning have been gaining popularity for predicting the volatility of the stock market prices. In this paper, the outcomes of many classical deep learning models such as LSTMs, GRUs, CNNs, and their several common variants are contrasted with two distinct stock price prediction targets: absolute stock price and volatility. The aim of the comparative study is to find out which model is the best fit for stock market prediction. We also attempt to research the relationship between news and stock trends, believing that news stories have an impact on the stock market by incorporating sentiment analysis into our model. Our methodology was to scrape news articles of a particular stock and use the corpus gathered to generate a sentiment score which is further used as an input to the model.

Keywords Neural network · Sentiment analysis · Stock market · News scrapings

1 Introduction

A stock also referred to as equity, is a security representing the holding of a fraction of the company. This grants the shareholder the right to a share of the assets and earnings of the corporation equal to how much of the stock they own. The stock market refers to the set of stocks and trades in which shares in publicly listed companies are acquired, exchanged, and released on a daily basis.

A stock market prediction is an attempt to forecast the future trend of an individual stock, a particular sector of the market, or the market as a whole. These forecasts

N. Jain · M. Motiani · P. Kaur (✉)

Computer Engineering, Netaji Subhas Institute of Technology, University of Delhi, New Delhi, India

e-mail: preeti.kaur@nsut.ac.in

generally use fundamental analysis of a company or economy, or technical analysis of charts, or a combination of the two.

The prediction of stock prices is a popular and significant problem. We can gain insight into market behavior over time with a good model for stock prediction, identifying patterns that would otherwise not have been observed. Machine learning would be an effective way to solve this issue with the rising computing capacity of the computer.

Forecasting stock volatility is, therefore, an essential application of current knowledge and resources in the field of deep learning.

In the past few years, a lot of models based on deep learning have been gaining popularity for predicting the volatility of the stock market prices [1]. In this project, we try to implement several classical deep learning models like RNNs, CNNs, and their many variants by implementing them using the Tensor flow framework in Python [2].

Then we perform a comparative study of all the aforementioned models with our aim being to find out which model is the best fit for volatility prediction.

One of the most enticing realistic uses of sentiment analysis is stock market forecasting. Another motivation to use sentiment analysis is that, according to the Efficient Market Hypothesis (EMH) [3], stock volatility cannot be predicted by historical prices alone in the long term as investors are guided by fear and greed [4].

The next objective of this research is to observe how well the movements in a company's stock prices, the rises, and falls, are associated with the public views expressed in news articles regarding that company [5].

Therefore, we will improve the predictive model's potential to correlate the public meaning of stock markets with investor opinion by adding a sentiment analysis module for news records. In order to forecast stock price movement for the next day, we use news sentiment and the values of the previous day.

2 Literature Survey

In the paper “Neural networks for stock price prediction (2018)” by Han et al. [6], five neural network models, namely, back propagation (BP) neural network, radial base function (RBF) neural network, general regression neural network (GRNN), support vector machine regression (SVMR), and least square support vector machine regression were surveyed and compared with predictive capacity (LS-SVMR). They conclude that the BP neural network reliably and robustly outperforms the other four models by following mean square error and average absolute percentage error as parameters. Over all the given stocks, the results show that the Back Propagation Neural Network's performance exceeds that of the other four models in terms of both Mean Squared Error, i.e., MSE and Mean Absolute Percentage Error, i.e., MAPE.

Another paper “Stock Prices Prediction using Deep Learning Models (2019)” by Liu et al. [7], talks about the challenges of using deep learning models for predicting stock prices. This is a challenge, since there is a lot of noise and confusion in stock

price-related details. In order to denoise the data, this work utilizes sparse autoencoders with one-dimensional (1-D) residual convolutional networks. In order to forecast the stock price, long-short term memory, LSTM is then used. Prices, indexes, and macroeconomic factors in the past are the attributes used to estimate the expense of the next day.

“Stock Trend Prediction using News Sentiment Analysis” by Joshi et al. [8] focuses on the famous theory of stock prediction, i.e., the Efficient Market Hypothesis. This paper includes aggregation of unquantifiable information such as financial news reports from a company and estimating the potential trend in the market using a news sentiment categorization. This is an attempt to research the relationship between news and stock trends, believing that news stories have an impact on the stock market. They developed three distinct classification models to explain this which indicate that the polarity of news articles is positive or negative. Observations show that in all forms of testing, random forest and support vector machines perform well. Naive Bayes performs well, but not in contrast to the other two. The accuracy of the forecast model is over 80 percent and 50% accuracy relative to news random labeling; the model has improved accuracy by 30%.

“Sentiment Analysis of Twitter Data for Predicting Stock Market Movements (2016)” by Pagolu [9] highlights that social media nowadays are a true reflection of public perception and opinion on current affairs. A fascinating area of research has been the stock market forecasting based on public opinions shared on Twitter. Earlier studies have shown that Twitter’s aggregate public sentiment may well be linked to the Dow Jones Industrial Average Index. (DJIA). Two separate textual representations, Word2vec and N-gram, have been used in the present paper to examine public feelings in tweets.

3 Background

3.1 Neural Networks

Neural Nets are a cognitive model of the human brain that can handle complex problems and identify unknown correlations and patterns using statistical algorithms and mathematics. They are capable of interpreting input data by means of machine perception, clustering or labeling. The examples used for training these nets are typically hand-labeled beforehand, if supervised learning is employed. For example, an object recognition system could be fed thousands of marked images of vehicles, buildings, coffee cups, and so on, and visual patterns would be identified in images that correspond closely with similar labels.

3.2 Recurrent Neural Networks

RNNs are used in the creation of models that replicate neuron function in the human brain. In situations where the context is important for predicting the result, they are incredibly powerful and are distinguished from other types of artificial neural networks because they use feedback loops to process a data sequence that informs the final outcome, which may also be a data sequence. These feedback loops cause knowledge to continue; the effect is often represented as memory.

3.3 Long Short Term Memory

LSTM [10] networks are a category of Recurrent Neural Network which uses standard units and special units. LSTM units have a “memory cell” that can store data over long periods of time in memory. A gate series is used to track where info arrives at the memory when it’s produced, and when it’s discarded. There are three kinds of viz gates, gate input, gate exit, and gate forget. The input gate defines how much information is stored in memory from the last sample; the output gate governs the sum of data transferred to the next layer and the tearing rate of the stored memory is controlled by forgetting gates. This design helps them to consider longer-term dependencies.

3.4 Gated Recurrent Unit

GRU [3] is a specific variant of the RNN which is commonly used for machine learning tasks related to memory and clustering. Gated recurrent units help to alter the weight of the input of the neural network to overcome the vanishing gradient problem common to recurrent neural networks. The gated recurrent units include two gates namely, a reset gate and an update gate as an improvement of the overall recurrent neural network structure. The uses of GRU include but are not limited to audio recognition, text recognition, handwriting, OCR, and many more tasks. In stock market analytics and government service, some of these networks are also used.

3.5 Convolutional Neural Network

A Convolutional Neural Network [11] is a machine learning algorithm that takes input data, usually images, calculates weights that are defined on the pretext of some untold feature signals about several elements in the input given, and can easily carry out such applications over these inputs. In the areas of facial and object recognition,

image detection, etc., CNNs have shown outstanding results. CNNs also benefit from fewer restrictions than a fully linked network, making it simpler to train and house the same number of secret units. In an image, by converting them into a type suitable for achieving the desired goal, it can effectively capture spatial and temporal dependencies.

3.6 *Variations*

Bidirectional RNN The bidirectional recurrent neural networks, BRNN [7], links two hidden layers going in opposite directions to the same output, allowing the algorithm to receive feedback from both the forward and the backward states.

2-Path Networks 2-path networks [12] are an easy but efficient way to organize and model RNN layers in a deep structure. They break the long sequential input into smaller chunks and iteratively implement intra- and inter-chunk operations, where the input length can be made equal in each operation to the square root of the initial sequence length.

Sequence to Sequence Networks Sequence to Sequence [13] (Seq 2Seq) modeling is about training models that can translate sequences, such as Hindi to Russian, from one domain to sequences from another domain. The encoder and decoder carry out this Seq 2Seq modeling.

Bidirectional Sequence to Sequence Variant Combining both the above approaches of Bidirectionally and Sequence to Sequence models we also get the Bidirectional Sequence to Sequence variant [14] of the LSTM and GRU.

4 Classical Way-Methodology

4.1 *Historical Data Collection*

We use Yahoo Finance for collecting the historical stock price data which will be used in all of the models as a means to record and observe the previous trends. The market is a lot affected by the crowd psychology. The demand and supply factors which in turn affect the price are affected by the psychology of the market. And it's believed that these psychological patterns recur over time. Thus, the price also tends to move in patterns. With the help of data analysis one can understand these price movements and be on the correct side of the market most of the time. Using technical analysis or charting helps you find out the odds in which price could probably move.

4.2 Historical Stock Price Data Processing

Extraction of Closing Stock Prices We use the closing stock prices for each day in our further analysis and predictions.

Normalization The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. For machine learning, every dataset does not require normalization. It is required only when features have different ranges. So we perform MinMaxScalar Normalization across all columns. MinMaxScalar scales all the data features in the range [0, 1] or else in the range [1] if there are negative values in the dataset.

Train Test Split We take the stock prices of the past 250 days from Yahoo Finance for a particular company, first 220 of which are used in training and the last 30 day data is used for testing.

4.3 Models Considered

Categories Considered:-

- Vanilla Recurrent Neural Networks
- Long Short Term Memory(LSTM) Model
- Gated Recurrent Unit(GRU) Model
- Convolutional Neural Network(CNN) Model.

Variations Considered

- Vanilla
- Bidirectional
- 2-Path
- Sequence to Sequence
- Bidirectional Sequence to Sequence.

4.4 Model Construction

The following are the hyperparameters that have been used in construction of all the models.

Hyperparameters

- Number of Layers in each Model = 1
- Batch Size for training = 5
- Epoch = 10
- Learning Rate for Back Propagation = 0.01

- Size of Test data = 30 days
- Dropout Rate for each cell in the model = 0.8.

4.5 Training

For each Model mentioned above:

- After normalization of data we take the 4the column of the dataset, i.e., “Close” indicates the closing stock price data for each specific date. We will use only this stock price across all models to maintain uniformity.
- Out of the 252 data points 222 are used for training and the most recent 30 days are used for testing during the train test split.
- The 222 data points are divided into batch sizes of 5 stored in the variable batch_X and sent as a placeholder for X in the model constructed using the feed_dict object. batch_Y is also formed of 5 data points but the data starts from the position 1 ahead of the starting position of that of batch_X. These are also sent to the model object and are used in the calculation of loss function.
- The model is run on the batch_X and batch_Y values and other inputs given and it gives 4 outputs variables/vectors.
 1. logits = containing the predicted values using the batchX
 2. laststate = containing outputs for each data point from last layer only
 3. loss = containing the loss function calculated above.

4.6 Testing

For each Model mentioned above:

- From the total 252 data points final 30 points are taken for testing purposes.
- Again here the test data is input in batch size of 5, the answer calculated is stored in an array “output_predict”.
- The input in the model is given through the standard feed\dict object like done in training to fill the placeholder.
- Inverse Normalization is applied using the “minimax.inverse_transform” function to get the data according to the input scale, not the standard Gaussian scale.
- The output is now returned and this is later used to compare with the original output and calculate the accuracy accordingly.

We ran 10 simulations of the model to ensure the perfect training and testing accuracies. Each simulation trains over the training data for epoch(10) times, i.e., we train over the data 100 times and we calculate the accuracy and cost side by side in each simulation.

4.7 Results

A “results” vector contains the result of all the 10 training/forecasts for the 10 epochs. We have the two following evaluation metrics:

- **Direction Prediction Accuracy:** Calculated by taking the percentage of predicted prices that are predicted to swing in the correct direction, i.e., the same direction as the original stock price for that day.
- **Price Prediction Accuracy:** 100-(Mean of Relative error of each day using the predicted price and the true price).

As we can observe that classical models give average performance on Direction Accuracy (Table 1).

Vanilla RNNs which are simple deep learning models perform worse on price accuracy because of their basic architecture of using just the immediate past information.

In general GRUs perform better than Vanilla RNNs because they store past patterns in multiple levels of memory and uses 2 gates to control the information in memory.

CNN, a traditional model used in image processing gives good accuracy in price prediction but suffers greatly in prediction Stock directions.

Table 1 Price and direction accuracy prediction (in %)

Model name	Price accuracy	Direction accuracy
Vanilla RNN	90.49	56.55
Vanilla RNN Bidirectional	89.97	57.55
Vanilla RNN 2-path	89.62	56.67
LSTM	94.98	57.80
LSTM Bidirectional	93.32	56.70
LSTM 2-path	94.40	53.66
LSTM seq 2seq	94.81	56.67
LSTM seq 2seq Bidirectional	94.01	60.00
GRU	94.13	53.66
GRU Bidirectional	90.28	53.34
GRU 2-path	93.21	54.67
GRU seq 2seq	90.88	55.00
GRU Bidirectional seq 2seq	67.99	53.33
CNN seq 2seq	90.73	54.00
Dilated CNN seq 2seq	95.86	53.33

Therefore, we see that LSTM models give better results for both stock direction and stock price among all. It uses 3 gates to select the right information to be stored in the memory. So we will be continuing with LSTM model in the next part, i.e., News Sentiment Analysis of the paper.

We plot each of the forecasts on the same graph and we also plot the actual true trend plot on the same graph colored black. This graph can be used to check where the model has undergone overfitting and where its still underfit and thus also determine a perfect forecast which most closely matches the true trend. Here the black line indicates the true trend and others are forecast trends.

**The graphs for all models are displayed after the references.*

5 Adding Sentiment Analysis

5.1 News Article Generation

Web Scraper [15] made in Python is used to get the list of links of news articles. This uses the python module “scrapy”. We use the website “reuters.com” for this purpose. For each day, we have a corresponding page on the website which lists all the news articles of the given company on that day. We use scrapy to extract the links of the news articles from the HTML anchor tags on the page. Then we output it in a CSV file with two columns, the date and URL of the news articles.

5.2 Sentiment Score Generation

We use the “Rosette” API for text analysis for the contents for each article outputting a confidence label corresponding to each article [16]. The confidence column presents the probability/confidence with which the label is decided for that specific date, i.e., label defines the attitude/emotion of the whole article, while entity label defines the emotion towards a specific entity here in.

This case is the company name. Similarly, confidence is the probability for the label, while entity confidence is probability for entity label. This file outputs a csv file which contains the columns date, label, confidence, entity label and entity confidence. The column confidence can range from -1: very negative outlook to 1: very positive outlook.

5.3 Bullishness Score

Using the sentiment score calculated for each article in the previous step, We calculate the bullishness score of the selected stock on a per day basis. Bullishness score [17] represents the average perspective of the particular stock on the particular date. Afterwards, we combine it with the stock price and sort it in chronological order.

Bullishness = Sum of Confidence Score/Number of Articles In case, the sentiment score of a day is not available, mark the sentiment score as the score of the latest previous day available.

5.4 LSTM Model

For training our model, the sentiment score goes z score normalization followed by min-max normalization. Then, we construct a 2 layer LSTM followed by 2 Dense layers. Input of which will be previous k days stock price and it's corresponding sentiment score where k is the window size. The output is the prediction of stock price. Hyperparameters of this model are as follows batch size = 128, epochs = 200, validation split = 0, verbose = 0. We used RMSProp optimizer [9] to minimize mean square error. Our objective is to maximize stock volatility prediction accuracy.

5.5 Results

We can see that accuracy is best increased in Facebook stock by 0.056 to achieve an overall accuracy of 0.6048, while the minimum increase in accuracy is shown in Apple stock of 0.0082. Average increase in accuracy among these 5 stocks is 0.03386 (Table 2).

After completing the prediction, we can clearly observe that the results using sentiment analysis offer marginally better results as compared to predicting without them. Hence, we conclude that analyzing sentiments in order to predict the volatility in the stock market is not worth the effort.

Table 2 Accuracy and the best k for RNN + EMM and RNN

	AAPL	AMZN	FB	GOOG	MSFT
LSTM	0.6129	0.5968	0.5484	0.5968	0.5726
LSTM + SI	0.621	0.6129	0.6048	0.6371	0.621
k(LSTM)	4	7	6	7	6
k(LSTM + SI)	4	9	7	7	5

6 Conclusion

In this paper, we have tried to predict stock market price and volatility. We created 15 classical deep learning models. On analyzing the prediction results from the classical deep learning models, we found that all the models performed almost equally for volatility prediction but LSTMs outperform the others for predicting the stock market price. We try to incorporate sentiment analysis in our LSTM model to further increase volatility prediction. We observed that using sentiment analysis offered only a marginal increase in accuracy, and hence concluded that given the time and effort required to incorporate news sentiment analysis into the models, it is not worth the effort.

Some of the limitations of this paper were due to Lack of historical news articles. We were not able to find any news website available for getting historical news articles on a particular company due to which, we had to stick to the biggest known companies for our research. Also, Under the free version of Rosette API, we were allowed to make only 500 calls largely limiting our power to process all our articles in one go. Therefore, we had to spread this part to multiple days to process sentiment analysis of all the articles.

Our analysis outlines that there is still scope for future research work like Merging multiple social media sources (like Twitter, reddit) and unconventional sources (like Youtube) for sentiment analysis and building the sentiment score module from scratch instead of outsourcing it to Rosette. We can also create a mechanism that only recommend when there is a noticeable sentiment undercurrent in the market for stronger results.

References

1. Zhou, S. A stock prediction method based on LSTM, p. 24 (2021). <https://doi.org/10.1007/978-3-030-63784-2>
2. Wang, G., Wang, T., Wang, B., Sambasivan, D., Zhang, Z., Zheng, H., Zhao, B.Y.: Crowds on wall street: extracting value from collaborative investing platforms. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing (2015)
3. Malkiel, B.G.: Efficient market hypothesis. In: Eatwell J., Milgate M., Newman, P. (eds.) Finance. The New Palgrave. Palgrave Macmillan, London, p. 13 (1989). <https://doi.org/10.1007/978-1-349-20213-3>
4. Teoh, T.-T., Lim, W., Koh, K., Soh, J., Tan, T., Liu, S.Y., Nguwi, Y.-Y.: From technical analysis to text analytics: stock and index prediction with GRU, 496–500 (2019). <https://doi.org/10.1109/cis-ram47153.2019.9095772>
5. Shynkevich, Y., McGinnity, T.M., Coleman, S., Belatreche, A.: Predicting stock price movements based on different categories of news articles. In: IEEE Symposium Series on Computational Intelligence (2015)
6. White: Economic prediction using neural networks: the case of IBM daily stock returns. In: IEEE 1988 International Conference on Neural Networks, San Diego, CA, USA, pp. 451–458, vol. 2 (1988). <https://doi.org/10.1109/icnn.1988.23959>
7. Chiewhawan, T., Vateekul, P.: Stock Return Prediction Using Dual-Stage Attention Model with Stock Relation Inference, p. 42 (2020). 10.1007/978-3-030-41964-6

8. Liu, J., Chao, F., Lin, Y.-C., Lin, C.-M.: Stock Prices Prediction using Deep Learning Models (2019)
9. Zou, F., Shen, L., Jie, Z., Zhang, W., Liu, W.: A sufficient condition for convergences of Adam and RMSProp (2018)
10. Mootha, S., Sridhar, S., Seetharaman, R., Gopalan, C.: Stock price prediction using bi-directional LSTM based sequence to sequence modeling and multitask learning (2020). <https://doi.org/10.1109/uemcon51285.2020.9298066>
11. Joshi, K., Bharathi, N., Rao, J.: Stock trend prediction using news sentiment analysis. *Int. J. Comput. Sci. Inf. Technol.* **8**, 67–76 (2016). 10.5121/ijcsit.2016.8306
12. Zhang, R., Wu, Z., Wang, S.: Prediction of stock based on convolution neural network, 3175–3178 (2020). 10.1109/CCDC49329.2020.9164222
13. Shynkevich, Y., McGinnity, T.M., Coleman, S., Belatreche, A.: Predicting stock price movements based on different categories of news articles. In: 2015 IEEE Symposium Series on Computational Intelligence.
14. Liu, L., Finch, A., Utiyama, M., Sumita, E.: Agreement on target-bidirectional LSTMs for sequence-to-sequence learning. In: Proceedings of the AAAI Conference on Artificial (2016)
15. Najork, M.: Web Crawler Architecture (2017). https://doi.org/10.1007/978-1-4899-7993-3_457-3
16. Jia, M., Huang, J., Pang, L., Zhao, Q. Analysis and research on stock price of LSTM and bidirectional LSTM neural network (2019). <https://doi.org/10.2991/iccia-19.2019.72>
17. Song, Y.-G., Zhou, Y.-L., Han, R.-J.: Neural networks for stock price prediction (2018)

Analysis and Prediction of COVID-19 Confirmed Cases Using Deep Learning Models: A Comparative Study



Trisha Sinha, Titash Chowdhury, Rabindra Nath Shaw, and Ankush Ghosh

Abstract COVID-19 or Novel coronavirus is an infectious disease that was first noticed in December, 2019 and it eventually emerged as a pandemic as it is highly contagious in nature. It affected the economic and social structure worldwide and caused a huge loss of human life. Due to the scarcity of medical infrastructure, it has become nearly impossible to cure every case of COVID-19 and hence the loss of lives is exceedingly increasing. So, if the cases can be forecasted beforehand, proper precautions can be taken on time and thousands of human lives can be saved. In this paper, predictions of the number of coronavirus confirmed cases for the five topmost affected countries across the world have been made. Along with it, a comparative study of ANN (Artificial Neural Network) and RNN (Recurrent Neural Network) based LSTM (Long Short Term Memory) Model has been carried out. The countries taken into consideration for this paper are USA, India, Brazil, Russia, and France. The models have been used to train the dataset and validate the prediction results against the original data based on the predefined metric of MSE or Mean Squared Error. The prediction results have been visualized graphically and it was inferred that the LSTM model outperformed the ANN model.

Keywords Machine Learning · ANN · RNN · LSTM · COVID-19

1 Introduction

The novel coronavirus disease is spreading throughout the world like a fire. The scientific name of coronavirus was first accepted as a genus name by International Committee for the Nomenclature of Viruses in the year 1971. The name coronavirus was mainly due to the structure of the virus. They are usually large roughly spherical

T. Sinha · T. Chowdhury · A. Ghosh (✉)

School of Engineering and Applied Sciences, The Neotia University, Kolkata, West Bengal, India

R. N. Shaw

Electronics & Communication Engineering, Galgotias University, Noida, India

e-mail: r.n.s@ieee.org

particles but with time and years it has evolved in itself, and according to the recent study, Coronavirus is a group of related RNA viruses that has spike like structure outside representing crown shape that causes disease in mammals and birds.

The first coronavirus case was reported in the Wuhan region of Hubei Province, China, in December, 2019, according to the World Health Organization (WHO). On 11th February, 2020, the International Committee on Taxonomy of Viruses (ICTV) termed this virus as “severe acute respiratory syndrome coronavirus 2”(SARS-CoV-2). And on the same day, World Health Organization announced the name of this disease as “COVID-19”. This name was mainly chosen because it was found that it was genetically related to the coronavirus responsible for the SARS outbreak of 2003 and also many more related outbreak in 1900s. The COVID-19 was declared as a Public health emergency of International concern (PHEIC) on January 30 by World Health Organization (WHO) [1]. Initially, Social Distancing was a method of keeping the virus at bay throughout the world followed by night curfew, quarantines, and similar restrictions which also included personal hygiene, wearing masks, using sanitizer, etc. But as the cases started increasing, half of the world population by April 2020, was under lockdown with billions of people ordered to stay at home by the government. The initial lockdown included closing of all educational institutes, closing of all non-essential as well as essential things. But with passing months, the essential services were opened like medical services, banks, etc. Partial lockdown was then implemented in many countries, states, and regions that helped many people in resuming their daily life by following the basic protocols such as social distancing, wearing facial masks, using sanitizer, etc. As of 3rd December, 2020, there have been 66,729,375 million confirmed cases, 1,535,982 million total deaths, and 43,645,100 million recovered cases [2].

Machine learning, especially its subfield of Deep Learning, had advanced to a great extent in the recent past. It has been widely used for various research work specially for predicting data due to its ability to automatically learn and improve from experience without being programmed explicitly. It has led to quite a breakthrough in technological field which is being used by billions of people. It also provides future opportunities to work on the model as well as to change accordingly for getting more accurate results. Deep Learning models provide real-time insights. It helps healthcare professionals diagnose patients faster and with more accuracy and reduce medical and diagnostic errors. Moreover, it helps the government in getting a brief idea about the future cases and what precautions they should take accordingly for keeping the number of cases in bay.

In this paper, we have used deep learning approaches of RNN (Recurrent Neural Network) based LSTM (Long Short Term Memory) model and ANN (Artificial Neural Network) to predict the number of COVID-19 cases in five top most affected countries throughout the world. Using both LSTM and ANN models, matrices like loss function were plotted. Even the real data and predicted data comparison was made by LSTM and ANN models and were plotted and it can be clearly seen that LSTM model gave much better results as compared to ANN model.

In our work, a comparative study is presented using LSTM model and ANN model for predicting the number of COVID-19 cases based on data from 31st December, 2019 to 3rd December, 2020.

2 Literature Review

D. Sarkar and M. Biswas used ARIMA models for predicting cases in West Bengal India. It was clear from the models that the cases have been increasing even during lockdown and the conditions are going to degrade due to the upliftment of Lockdown and initiating Unlock 1.0 [3].

In the paper by ‘S K Tamag, P D Singh, and B Datta’, they made prediction using curve fitting technique of artificial intelligence. In this study, they presented various nonlinear models with the help of curve fitting technique unlike any other conventional method. They employed this method as it does not involve complex calculations. And this might help other countries in decreasing the cases by maintain the protocols [4].

In the study by S Dutta and Dr. S K Bandhopadhyay, three different models of machine learning technique were used for the Confirmation of COVID-19 cases. The three models used were based on RNN models of LSTM, i.e., LSTM, GRU, and LSTM-GRU. And from the three models, they concluded that the combined LSTM-GRU-based RNN model provides comparatively better results as compared to others [5].

A Mollalo, K M Riveria, and B Vahedi in their study used Artificial Neural Network for modeling of novel coronavirus (COVID-19) across the continental United States. They used multi-layer perceptron ANN in modeling cumulative incidence of COVID-19. It was concluded by them that though the model employed gave results, but it did not have consistency [6].

There is another research article by Fenglin Lu, who predicted and analyzed the COVID-19 epidemic in China, using SEIRD, LSTM, and GWR models. They concluded that all the three models gave great results and by modifying all these models, it could be used for higher studies [7].

3 Methodology

Data Description

In this study, we have collected data for COVID-19 worldwide or on country level from 31st December, 2019 to 3rd December, 2020 from European Centre for Disease Prevention and Control. We have worked on the top five most affected countries reported by WHO (World Health Organization) as of 3rd December, 2020, which includes USA on top with a total of 14,570,523 cases followed by India with

9,703,770 cases, then Brazil with 6,603,540 cases followed by Russia with 2,488,912 cases. And on the fifth position is France with 2,252,282 cases. The data is organized as sequential time series data [8].

Data Preprocessing and Analysis

The dataset contained number of confirmed cases and number of deaths per day for 339 days starting from 31st December, 2019. For this study, the number of confirmed cases per day is taken into consideration. Data preprocessing is carried out and all the null and erroneous values have been taken care of. The graphical visualization of the number of cases per day for all five countries have been shown in Fig. 1a–e.

Figure 1a shows the graphical representation of the number of cases in USA from 31st December 2019 to 3rd December 2020. Figure 1b shows the variation in confirmed cases in India from 31st December 2019 to 3rd December 2020. Figure 1c depicts the number of COVID-19 confirmed cases in Brazil during the same period of time. Figure 1d shows the graphical representation of the number of confirmed cases in Russia from 31st December, 2019 to 3rd December, 2020. Finally, Fig. 1e is the graphical representation of the cases in France for the same duration. The above-mentioned countries have been plotted in the decreasing order of the number of confirmed COVID-19 cases, USA having the maximum cases among all other countries as per 3rd December, 2020.

4 Artificial Neural Network (ANN)

Artificial neural networks are built like the human brain, with neuron nodes interconnected like a web. It has hundreds or thousands of artificial neurons called processing units, which are interconnected by nodes [9, 10]. The processing units are made up of input and output units. The input units receive various forms and structures of information, and the neural network tries to learn the information fed to it to produce output report.

The ANN is usually made up of input layer, output layer and one or more hidden layer. The hidden layer usually sees the activity of the input units and also the activity of each hidden layer includes the internal weight between input layer and hidden layer. ANN have the ability to learn and model nonlinear relationships [11–13]. Due to this unique feature, ANN models are preferred more than other algorithms. Due to the hidden layers in a neural network, since more is learned and thus the pattern detection is more accurate.

For the calculation of neural network, there are two aspects—weight and activation. Weight is a connection between different neurons which carries a value. Higher the value, larger is the weight. The neurons involved in the process are supposed to make a small decision on the output, which, in turn, returns another output. This process is known as activation. Activation is represented by $f(z)$, z is the aggregation of all the input. When $f(Z) = z$ then $f(z)$ is known to be a linear activation function.

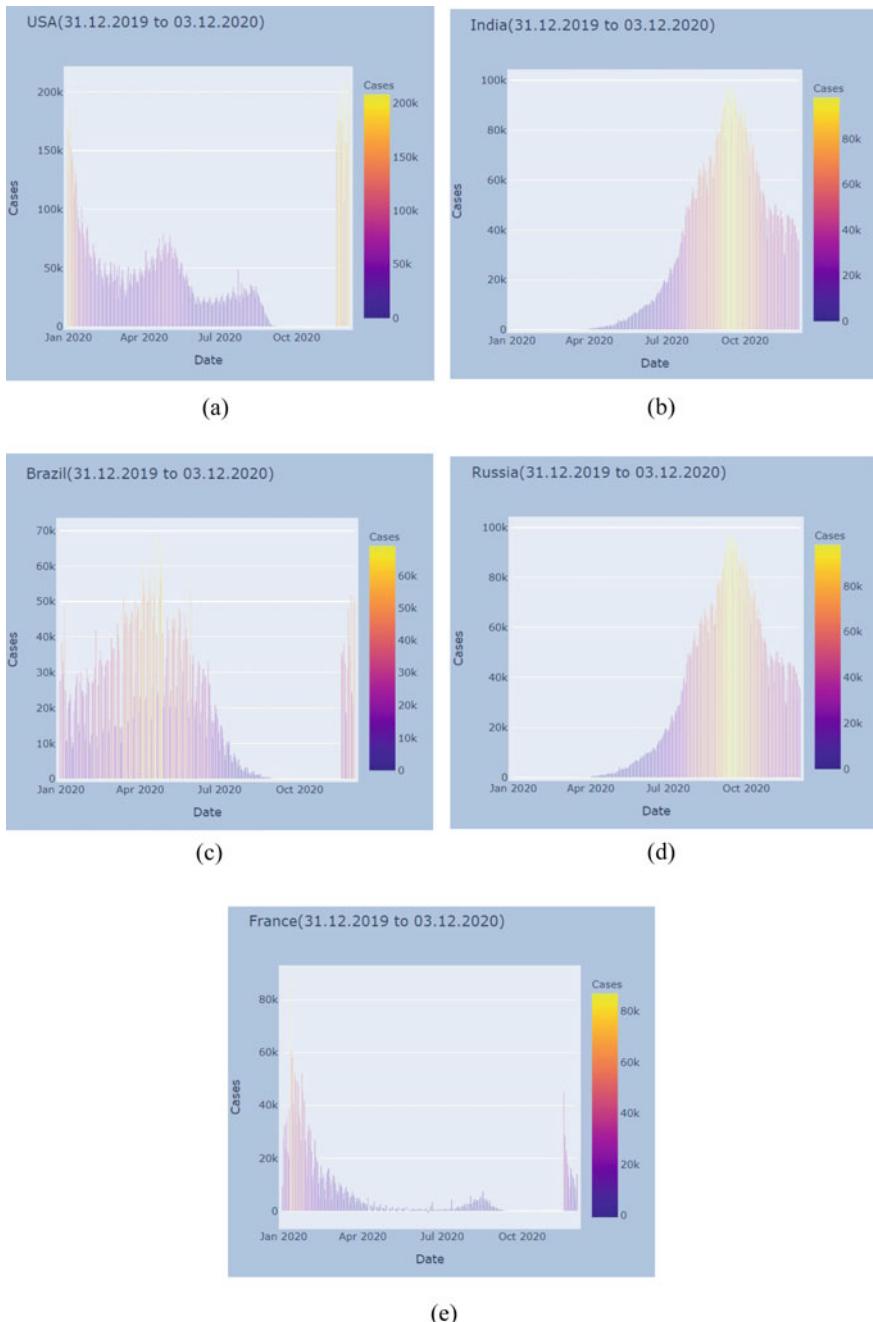


Fig. 1 Graphical representation of the trend of number of confirmed cases over 399 days

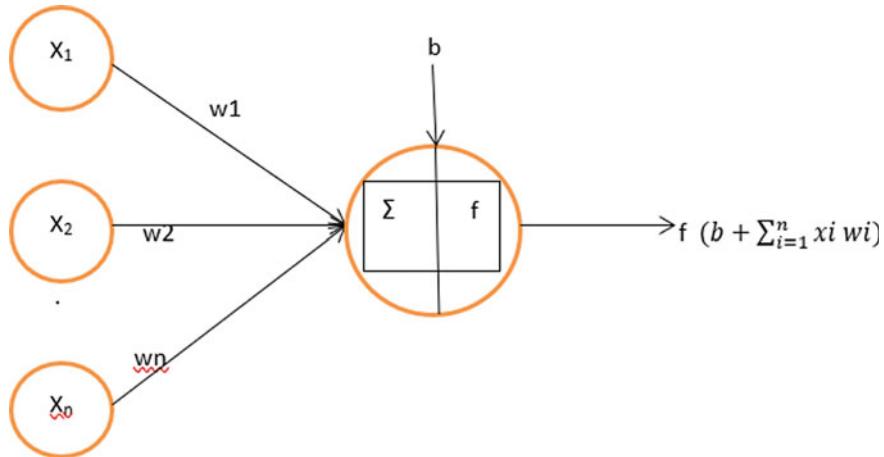


Fig. 2 Example of Artificial Neural Network Model

Few other activation functions are ReLu(Rectified Linear Units), Tanh, and Sigmoid activation functions (Fig. 2).

Activation function f is given by $f(b + \sum_{i=1}^n x_i w_i)$, where ‘ b ’ is the bias, ‘ x ’ is the input to the neuron, ‘ w ’ is the weight, ‘ n ’ is the input from the incoming layer and ‘ i ’ is the counter. Long Short Term Memory (LSTM).

LSTM (Long Short Term Memory) is a type of Recurrent Neural Network architecture which is unsupervised learning method used in the field of deep learning. Just like RNN it has feedback connections. It processes single data as well entire sequence of data. It is commonly applied in classifying, processing, and making predictions which are based on time series data. It has great application in complex problem domains like machine translation, speech recognition, and more (Fig. 3).

A single LSTM cell has 4 different components. They are forget gate, input gate, output gate, and the cell state. The inputs to the LSTM cell is X_t (current input). Then the output of hidden layer, namely, the current hidden state h_t is computes as-

$$f_t = \sigma(w_f [h_{t-1}, X_t] + b_f),$$

$$i_t = \sigma(w_i [h_{t-1}, X_t] + b_i),$$

$$\sigma = \sigma(w_0 [h_{t-1}, X_t] + b_0),$$

$$C_t = f_t * C_{t-1} * i_t * \tanh(w_c [h_{t-1}, X_t] + b_c), h_t = o_t * \tanh(C_t),$$

Here f_t , i_t , and o_t are the forget, input, and output gates, respectively, h_{t-1} is the previous hidden state, C_{t-1} and C_t are previous and current cell memories. The weight matrices w_f , w_0 , and w_c and the bias vectors b_f , b_i , b_o , and b_c are model parameters.

5 Proposed Models

ANN Model

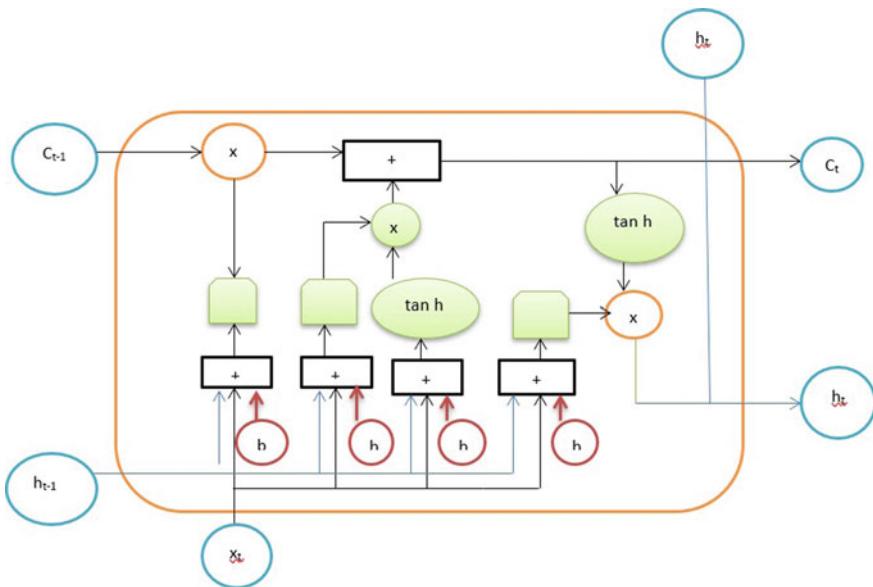


Fig. 3 Structure of LSTM

In this ANN model, a simple three-layered structure is used with 12, 8, and 1 neurons in hidden layers 1, 2, and 3, respectively [14]. The activation function used is ReLu and the best hyperparameters used are a verbose of 1 and a batch size of 32. MSE or Mean Squared Error is used as the metric for verification of predicted results (Fig. 4).

LSTM Model

In this model, a total of 4-layered LSTM structure followed by a Dense Layer is used as for verifying prediction result. The best hyperparameters used in this model are a batch size of 32 and a dropout rate of 0.2. The LSTM layers use a sequence of 50 nodes [15]. The metric used for identifying the accuracy of the model is MSE. In this model, the LSTM_1 layer is followed by Dropout_1 layer and LSTM_2 layer is followed by Dropout_2 layer. Similarly, LSTM_3 and LSTM_4 layers are followed by Dropout_3 and Dropout_4 layers, respectively (Fig. 5).

6 Results and Discussions

The models described in the previous section are used to predict the number of confirmed cases of COVID-19. Here, data for five countries have been considered and a comparative study has been made between them based on MSE metric. Fig. 6a–e shows the comparison between the LSTM model and the ANN model.

Fig. 4 Structure for ANN Model

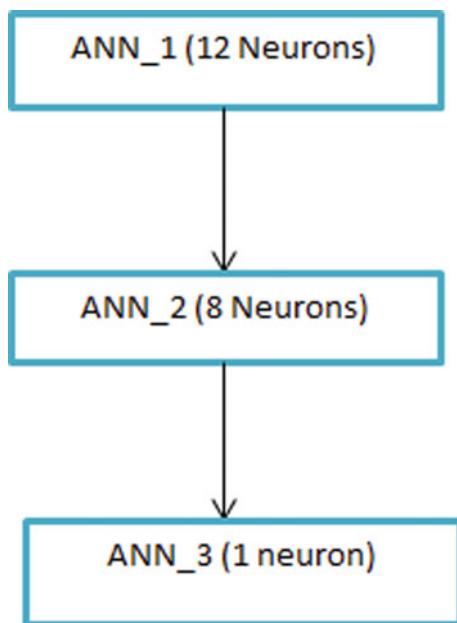
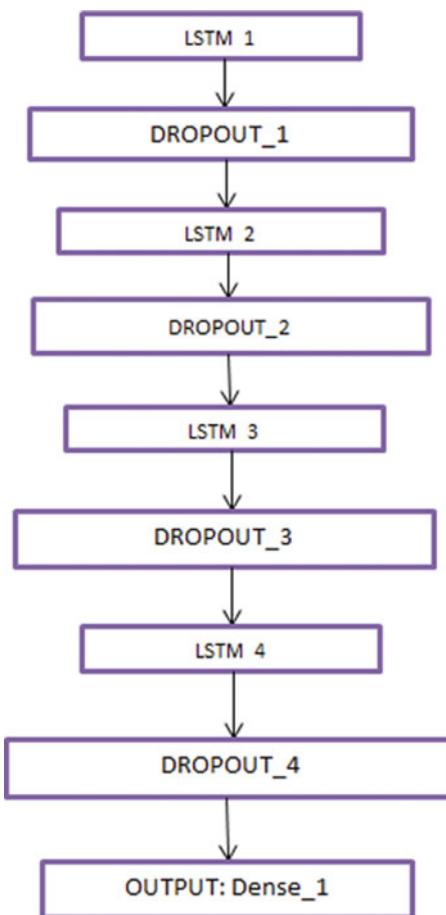


Figure 6a represents data of USA, where it can be seen that the real-time data and the predicted data have better matched in the representation using the LSTM model. In Fig. 6b, data of India is depicted. The graph clearly shows that the predicted data have better matched with the real-time data in LSTM algorithm. Figure 6c shows the predicted and real-time data of Brazil, and from the representation, it is clear that the data have better matched in case of LSTM model. Figure 6d represents data of Russia and after analysis with ANN and LSTM models, here also, it can be seen that LSTM model shows better match of real-time data and predicted data. From Fig. 6e, that depicts the data of France, again it is evident that the real-time data and predicted data have better matched in case of LSTM model rather than in ANN model.

The table consisting of MSE values of five countries for both the models used have been listed below. From the MSE values and the visual representation of the predictions, it can be seen that the LSTM model performed way better than the ANN model. The representation on the left shows the prediction results of the LSTM model and those on the right shows the prediction results of the ANN model.

The MSE values of the predictions by the LSTM model for all five countries are less than the MSE values of the predictions by the ANN model. The values for both the models for each and every different case can be found in Table 1.

Fig. 5 Structure for LSTM model



7 Conclusion

In this paper, LSTM and ANN models have been deployed to predict the COVID-19 confirmed cases of five top most affected countries. Based on the predictions made by these two models, a comparative study of the ANN model and the LSTM model has been done. From the graphical visualization, as well as the MSE values for both the models, it can be assuredly concluded that LSTM model gives better results than the ANN model in all five cases. It is observed that LSTM models are well fitted with the training data and it can be further used to make predictions in the near future. COVID-19 has slowed the pace with which the world was developing in terms of technology as well as economics and it will continue to do so if necessary precautions are not taken. We need to learn to live with the disease, but at the same time, find ways to combat it. Hence, predictions using Deep Learning models can be proved beneficial to curb the cases and alleviate the effect of the global pandemic.

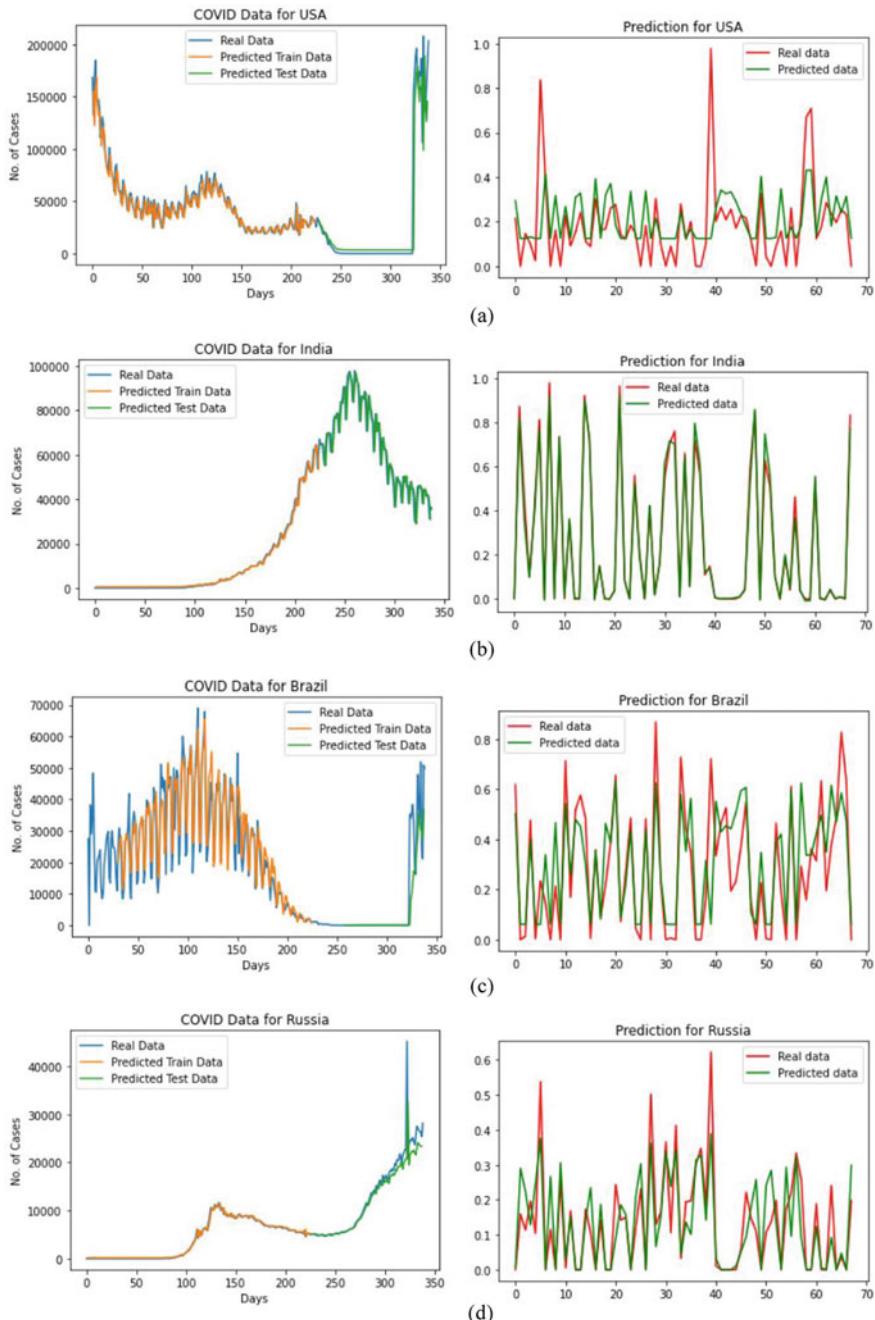
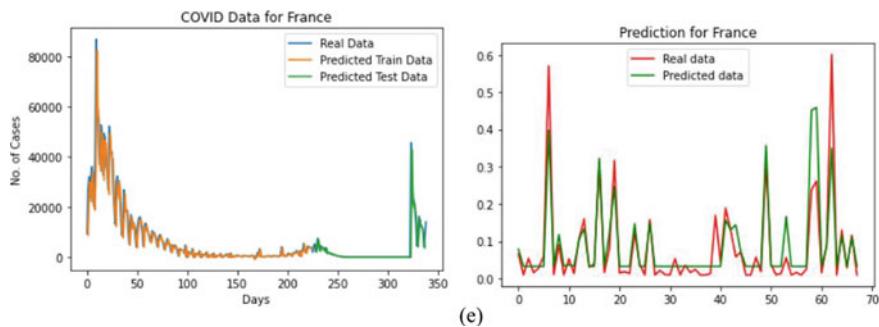


Fig. 6 Predictions by LSTM model Vs. Predictions by ANN model

**Fig. 6** (continued)**Table 1** MSE values for LSTM and ANN

Country	MSE for LSTM	MSE for ANN
India	2.3083e-04	0.0021
USA	0.0017	0.0333
Russia	7.5259e-05	0.0098
Brazil	0.0075	0.0247
France	0.0044	0.0140

The data driven model is capable of providing an automated tool for reassuring and confirming the present status and severity of the pandemic. It can also be used to estimate the future impact of the disease. Hence, this model can help the government and the health workers to make reliable decisions regarding the policy and precautions needed to ease the damages caused by COVID-19. Also, from this study, it can be recognized that, which countries should take more precautions and which of them should continue to maintain the similar protocols as before. In future, the LSTM model can be modified by adding more hyperparameters to improve the results. Even, other deep learning models can be deployed for the job of prediction to achieve more accuracy.

References

1. Sarkar, D.: COVID 19 pandemic: a real time forecast and prediction of confirmed cases, active cases using the ARIMA model and public health in West Bengal, India
2. World Health Organization (WHO) statement regarding the report of first case in Wuhan, China
3. Ahmad, I., Asad, S.M.: Predictions of coronavirus COVID 19 distinct cases in Pakistan through ANN
4. Shastri, S., Singh, K., Kumar, S., Kour, P., Mansotra, V.: Time series forecasting of COVID 19 using deep learning models: India-USA comparative case study
5. Dutta, S., Bandhopadhyay, S.K.: Machine learning approach for confirmation of COVID 19 cases: positive, negative, death and release

6. Mollalo, A., Riveria, K.M., Vahedi, B.: Artificial neural network modeling of novel coronavirus (COVID 19) incidence rates across the continental United States
7. Car, Z., Segota, S.B., Andelic, N., Lorecin, I., Mrzljak, V.: Modelling the spread of COVID 19 infection using a multilayer perceptron
8. EU Open Data Portal
9. Marquez, B.Y., Caldreon, E.A., Alanis, A., Luis, M.G.: Comorbidities in patients with COVID 19, case study: Baja California, using ANN
10. Crivellari, A., Euro BEinat: LSTM Based deep learning model for predicting individual mobility traces of short term foreign tourists
11. Mandal, S., Biswas, S., Balas, V.E., Shaw, R.N., Ghosh, A.: Motion prediction for autonomous vehicles from lyft dataset using deep learning. In: 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, pp. 768–773 (2020). <https://doi.org/10.1109/iccca49541.2020.9250790>
12. Mandal, S., Balas, V.E., Shaw, R.N., Ghosh, A.: Prediction analysis of idiopathic pulmonary fibrosis progression from OSIC dataset. In: 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, India, pp. 861–865 (2020). <https://doi.org/10.1109/gucon48875.2020.9231239>
13. Tamang, S.K., Singh, P.D., Datta, B.: Forecasting of COVID 19 cases based on prediction using artificial neural network curve fitting technology
14. Liu, F., Wang, J., Liu, J., Li, Y., Liu, D., Tong, J., Li, Z., Yu, D., Fan, Y., Bi, X., Zhang, X., Mo, S.: Predicting and analyzing the COVID 19 epidemic in China: based on SEIRD, LSTM and GWR models
15. Kumar, M., Shenbagaraman, V.M., Ghosh, A.: Predictive data analysis for energy management of a smart factory leading to sustainability. In: Favorskaya, M.N., Mekhilef, S., Pandey, R.K., Singh, N. (eds.) Innovations in Electrical and Electronic Engineering [ISBN 978-981-15-4691-4], pp. 765–773, Springer (2020)

Machine Learning Technique to Prognosis Diabetes Disease: Random Forest Classifier Approach



Prajyot Palimkar, Rabindra Nath Shaw, and Ankush Ghosh

Abstract Diabetes is one among many chronic diseases. It is the most common disease and lots of peoples are affected by this. There are many things that are liable for diabetes, mainly age, obesity, weakness, sudden weight loss, and many more. Diabetes patients have high risk of diseases like cardiopathy, renal disorder, stroke, nerve damage, eye damage, etc. Detection of the disease isn't very easy and prediction is additionally costlier. In today's situation, hospitals are extremely busy due to COVID-19 pandemic, and it might be revolutionary if one could know if they're at risk of being diabetic without visiting a doctor. But the rise in Artificial Intelligence techniques can be used for disease prognosis. The objective of this study is to develop a model with significant accuracy to diagnose diabetes in patients. Moreover, this paper also presents an effective diabetes prediction model for better classification of diabetes and to enhance the accuracy in diabetes prediction using several machine learning algorithms. Different machine learning algorithms are utilized for early stage diabetes prediction, namely, Logistic Regression, Random Forest Classifier, Support Vector Machine, Decision Trees, K-Nearest Neighbors, Gaussian Process Classifier, AdaBoost Classifier, and Gaussian Naïve Bayes. The performances of these models are measured on respective criteria like Accuracy, Precision, Recall, F-Measure, and Error. For this research work, latest available dataset dated 22nd July, 2020, is being utilized. Latest updated dataset will show comparatively better result.

Keywords Diabetes · Machine learning algorithm · Random forest · Decision tree · Predictive analysis technique · Classification

P. Palimkar · A. Ghosh (✉)

School of Engineering and Applied Sciences, The Neotia University, Kolkata, West Bengal, India

R. N. Shaw

Department of Electrical, Electronics & Communication Engineering, Galgotias University, Noida, India

e-mail: r.n.s@ieee.org

1 Introduction

Diagnosis is the most vital part of the medical science. They use different strategies for that. Under this strategy, classification of given data is done in different classes based upon some constraints. The existence of diabetes depends upon different types of factors. Diabetes is an illness due to inability of human body, to secrete harmon insulin. Imbalance insulation show symptoms like intensified thirst and hunger, high blood glucose, frequent urination. If diabetes is untreated, many complications could also be occurring, which end up in serious health problems. So, early prediction of diabetes is critical to cut back the consequences of it.

Different predictive analyses [1] include machine learning algorithms and statistical methods. Under this technique, the classification of past data is done to get knowledge to predict future events. Machine learning and regression technique can be effectively used for predictive analysis. Among different artificial intelligence techniques, machine learning is taken into account as the most vital feature. It supports computing system by acquiring knowledge from the past experience without any programming. So machine learning is an ultimate solution to reduce human effort because it supports automation with negligible error.

For proper prediction of diabetes, machine learning gives preferable good results. Various machine learning techniques are Logistic Regression, Random Forest Classifier, Support Vector Machine, Decision Trees, K-Nearest Neighbors, Gaussian Process Classifier, AdaBoost Classifier, and Gaussian Naïve Bayes. The aim of this paper presentation is to predict possibility of being diabetic by using machine learning model.

A brief introduction of this research paper is discussed as follows: Sect. 2 discusses the brief related work of other researchers. In Sect. 3, description of dataset is mentioned. Section 4 proposes applied methodology. Explanation of the proposed model is done in Sect. 5. Final results outcome is produced in Sect. 6. Section 7 is the concluding part of this research paper.

2 Related Works

Komi et al. [2] had evaluated five types of classifier techniques for diabetic candidate classification. The researcher had tested these methods, namely, Logistic Regression, Expectation Maximization, Support Vector Machine, Artificial Neural

Networks (Anns), and Gaussian Mixture Modelling. To improve the accuracy perfection of the model researcher had tuned some hyperparameters. ANNs outperforms comparatively to other tested models.

Perveen et al. [3] had discussed Bagging and AdaBoost Classifier machine technique models for classification [4]. After considering the various diabetes risk parameters for diabetes classification, the authors had used J48 decision tree. The conclusion of the research was that AdaBoost Classifier, an ensemble machine learning algorithm, proved to be better as compared to Bagging and J48 Decision Tree.

Orabi et al. [5] had developed a prediction model for diabetes patients. The objective of the study was to predict whether a patient is non-diabetic at a specific age . Decision Tree gave an optimum accuracy as it had properly predicted patient's diabetes risk factor at specific age [6, 7].

Pradhan et al. [8] had obtained the diabetes dataset from UCI Repository for classifying and analyzing through Genetic Programming (GP) [9, 10]. GP gives the prediction values with maximum accuracy in minimum cost.

Kumar et al. [11] had tested mainly two models for diabetes prediction. The researcher had used ANN as well as Fasting Blood Sugar (FBS). For detecting the existence of diabetes among patients, Decision Tree [12] was used.

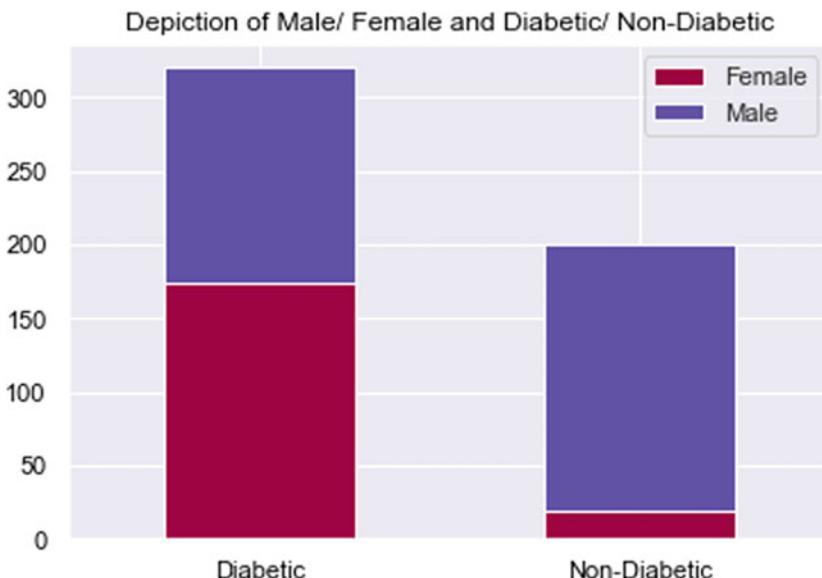
Mandal et al. [13] had described four different types of algorithm for diabetic patient classification, namely, Naive Bayes, Logistic Regression, ANNs, and Decision Tree. Some hyperparameters were tuned to enhance the accuracy of the model. Random Forest algorithm gave better results.

3 Dataset Description

The patient dataset is collected from Sylhet Diabetes Hospital, Bangladesh. The creation of dataset is based on a direct questionnaire to diabetic patients who have recently been diabetic, and few non-diabetic people who have some symptoms approved by a doctor. The dataset utilized in this paper is present in the UCI Repository [14]. This dataset contains patient's crucial features which are useful for the prediction of diabetes. This diabetes dataset contains 17 attributes of 520 patients out of which 328 are male patients, 192 are female patients as well as 320 are positive and 200 are negative which is represented in the following graph.

Table 1 Dataset description

	Attributes (nos.)	Instances (nos.)
Diabetes symptom dataset	17	520



Graphical Distribution of Male and Female Patients w.r.t. Diabetic/Non-Diabetic

The detailed description of dataset and attributes are as below, respectively (Tables 1 and 2).

Class variables are accustomed to find whether the patient is diabetic (Positive) or not (Negative).

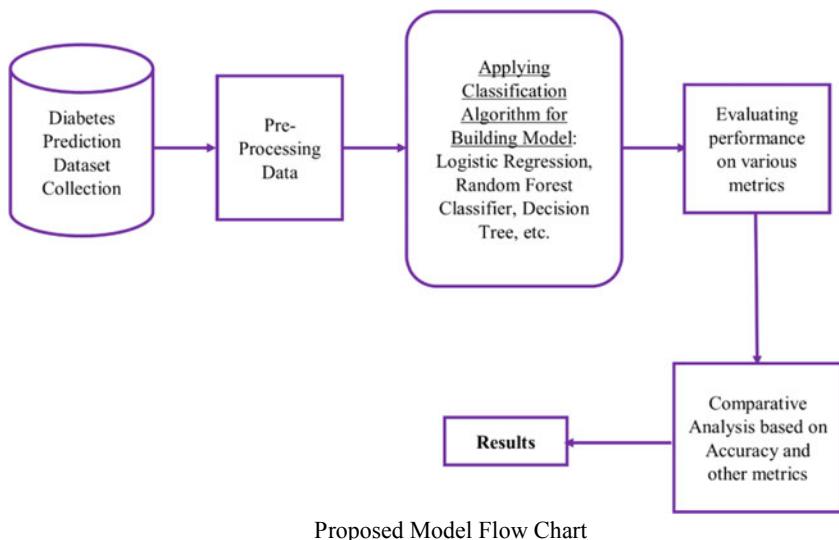
4 Methodology

This research focuses to minimize the complications of diabetes through early prediction so on to enhance the lives of the patients. The person has diabetes because of some considerable features, depending upon their age, gender, weight, and some other factors. The main purpose of research is to find whether the particular patient is diabetic or not by using the Classification technique.

For this, dataset is divided into training and testing sets. For model construction, training dataset is utilized. Testing dataset is employed just for accessing the performance of the model. Therefore, testing dataset is 30% of the entire dataset and rest 70% is utilized as training dataset.

Table 2 Attribute description

Attributes name	Description/Data dictionary
Age	16–90 (years)
Gender	0: Female; 1: Male
Polyuria	0: No; 1: Yes
Polydipsia	0: No; 1: Yes
Sudden weight loss	0: No; 1: Yes
Weakness	0: No; 1: Yes
Polyphagia	0: No; 1: Yes
Genital thrush	0: No; 1: Yes
Visual blurring	0: No; 1: Yes
Itching	0: No; 1: Yes
Irritability	0: No; 1: Yes
Delayed healing	0: No; 1: Yes
Partial paresis	0: No; 1: Yes
Muscle stiffness	0: No; 1: Yes
Alopecia	0: No; 1: Yes
Obesity	0: No; 1: Yes
Class	0: Negative; 1: Positive



This model has the following different modules:

- Data Collection
- Pre-Processing Data
- Building Model

- iv. Evaluation
- v. Comparative Analysis
- vi. Results

Let's have a glance at each model briefly.

4.1 Data Collection

For this research, latest available dataset dated 22 July, 2020, is being utilized. Diabetes dataset is available from the UCI Repository [14]. Latest updated dataset will show comparatively better results. Then information present in this is thoroughly understood by studying its pattern and trends. This diabetes dataset contains information about the symptoms of the patients.

4.2 Pre-processing Data

The data is pre-processed by one hot encoding as the data present is in kind of “Yes” and “No,” “Male” and “Female,” “Positive” and “Negative,” which are converted into 1 and 0, respectively.

4.3 Building Model

This dataset is going to be fed to the various classification algorithms like Naïve Bayes, Logistic Regression, AdaBoost Classifier, Random Forest Classifier, and many other for training and testing data using the classification algorithm.

Subsequent paragraphs, however, are indented.

4.4 Evaluation

The performance of the algorithms is going to be tested with an appropriate evaluation model. For evaluation, various metrics were considered like Error, Recall, F1-score, Support, Accuracy, Macro Average, Precision, Weighted, Average, Area Under Curve (AUC), and most significantly, Confusion Matrix.

Accuracy: Accuracy is the ratio between correct numbers of prediction to total number of predictions. It is given as follows:

$$\text{Accuracy} = \frac{\text{Correct Prediction (no.)}}{\text{Total Prediction (no.)}}$$

Confusion matrix: It is the matrix/table that is accustomed to depict the performance of the test data for which actual values are known.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	True Positive (TP)	False Positive (FP)
	Negative (0)	False Negative (FN)	True Negative (TN)

Now, the accuracy will be calculated from this confusion matrix as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

Precision: Precision is the ratio between the number of actual true positive results divided by total positive results predicted by the given model, i.e., false positive and true positive. It is expressed as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Mean Square Error: Mean Square Error (MSE) is the average of the squared error. MSE is the summation of square of the difference between the actual and predicted values of data points, divided by the total number. Root Mean Square Error (RMSE) is the square root of MSE.

$$\text{MSE} = \frac{\sum_{i=1}^N (\text{Predicted}_i - \text{Actual}_i)^2}{N}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^N (\text{Predicted}_i - \text{Actual}_i)^2}{N}}$$

Recall: It is the ratio of true positive results divided by the number of actual positive results predicted by the model, i.e., false negative and true positive. It is expressed as

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

F1-Score: F1-Score is used to measure the accuracy of the testing dataset. Harmonic mean of recall and precision is called F1-Score. The value of F1-Score lies between 0 and 1. Mathematically, it is given as

$$\begin{aligned}\text{F1 Score} &= 2 * \frac{1}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} \\ \text{F1 Score} &= \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}\end{aligned}$$

Support: Support is term used for the actual finding of specified category within given dataset. Imbalanced support within the training data may show structural weaknesses in the reported scores of the classifier and the necessity for proportional sampling or rebalancing.

Sensitivity: Sensitivity is the ratio of true positive results divided by the number of actual positive results predicted by the model, i.e., false negative and true positive. It is the metrics used to check whether model can predict true positive from the given dataset.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Specificity: Specificity is the ratio of true negative results divided by the number of actual negative results predicted by the model, i.e., true negative and false positive. It is the metrics used to check whether model can predict true negative from the given dataset.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

4.5 Comparative Analysis

So, considering the above metrics, comparative analysis has been done on this dataset. The main objective of this paper is to tune several hyperparameters of different machine learning methods for improving the accuracy of the proposed model.

4.6 Results

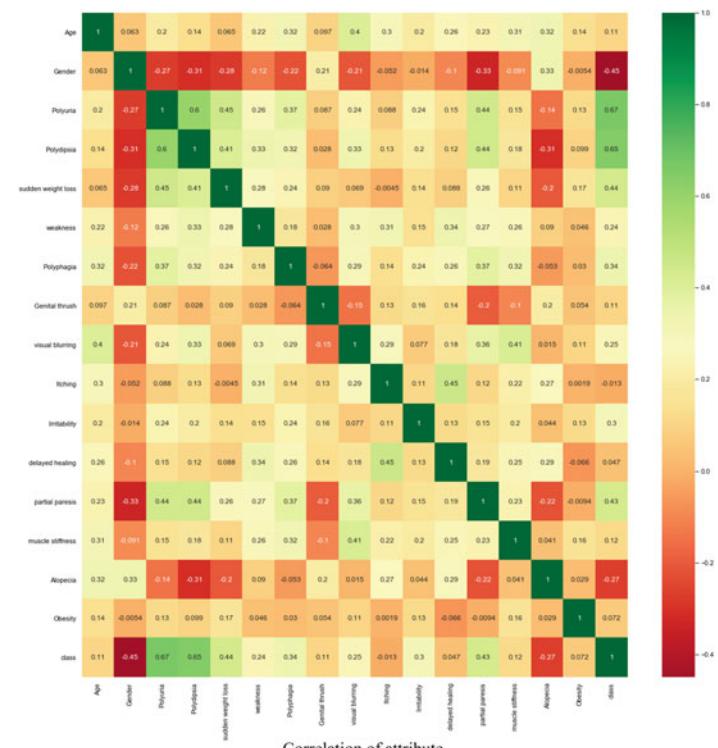
The results are obtained by doing the comparative study and finding the foremost effective algorithm which can be chosen to create the ultimate system for the end users using the dataset as database. So, the final model can be accustomed to predict whether the patient is diabetic or not by taking the symptoms from the user as input and providing it to model.

5 Proposed Models

Modeling

Collinearity, Correlation, and Covariance

When more than two variables are correlated to each other, that will be the case of collinearity. Correlation is a standardized value of strength and direction which shows the connectivity among the given attributes. Covariance denotes non-standardized value of direction and relation among them. The following figure shows the correlation of varied features with each other.



5.1 Logistic Regression

Statistical model which has binary variable quantity and uses logistic function is called Logistic Regression. Therefore, this can be an appropriate model to predict the category whether the patient is diabetic or not. The general form of logistic regression function is

$$\log \frac{\hat{\pi}_i}{1 - \hat{\pi}_i} = x_i^T \beta$$

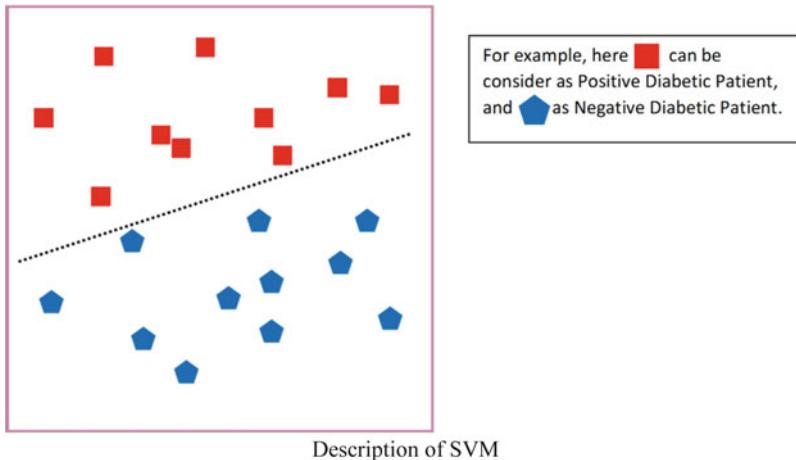
where $\hat{\pi}_i$ is the estimated probability that observation i is positive, x_i is the i th vector within the design matrix, and β is the vector of coefficients. In this case, the first element of x_i is 1 to activate the intercept in β , the second element of x_i is the age of observation i , and also the remaining elements are 1–0 dummy variables.

5.2 Random Forest

The extension of decision tree is called Random Forest. It builds multiple decision trees and merges them to induce more accurate and stable prediction. Random Forest is used for regression and classification. For an overall better performance, multiple machine learning models are combined which is called ensemble learning. Random forest is one of the ensemble learning models. Ho [15] had created the first algorithm for random forest by using the random subspace method [16].

5.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a very efficient, potent machine learning algorithm. It can even achieve far better results than neural networks in some areas. This is often used as a classification model with quite different methodology. Here, the aim is to work out the hyperplane that distinctly classifies the data points and has maximum margin to any or all the other points.



5.4 K- Nearest Neighbors (K-NN)

K-Nearest Neighbors (K-NN) is the simplest classification algorithm used. It also can be used for regression. K-NN is non-parametric, instance-based (that does not explicitly learn a model). Instead, it chooses to memorize the training instances. The output of K-NN is one of the class member which means result is one of the category from the available categories. There are three key elements in this approach: a collection of labeled object (set of stored record), distance between the objects, and the number of nearest neighbors.

5.5 Decision Tree

Decision Tree (D-Tree) is a supervised learning algorithm used for regression and classification models. D-tree represents the diagrammatic importance of each part. Availability of the number of choice is depicted by branches, Where each leaf shows final decision whether a candidate is diabetic patient or not. D-Tree is highly interpretable, easy to grasp, and visualize.

5.6 Gaussian Process Classifier (GPC)

It is a classification algorithm based on the Laplace approximation. GPC is a probabilistic classification method which uses Gaussian processes (GP), i.e., output of this is probability of having a particular class.

5.7 AdaBoost Classifier

The AdaBoost Classifier work as a meta-estimator, which applies training and testing methods on the original given dataset. It focuses on improving weights of subsequent identical data. AdaBoost Classifier comes under ensemble boosting classifier technique.

5.8 Gaussian Naïve Bayes (Gaussian NB)

Naive Bayes (NB) could be a powerful classification algorithm used for both multi-class and binary class classification problems. Just by assuming, a Gaussian Distribution Naïve Bayes was further extended to Gaussian Naïve Bayes (Gaussian NB).

6 Results and Discussion

See (Table 3).

Table 3 Model accuracy and error

Model name	Training accuracy	Test accuracy	MSE	AUC
Logistic regression	92.8571	93.5897	6.4102	92.7472
Random forest	100.0	99.3589	0.6410	99.2307
Support vector machine	93.6813	94.2307	5.7692	93.9560
K-nearest neighbors	100.0	94.2307	5.7692	94.6153
Decision tree	100.0	98.7179	1.2820	98.9010
Gaussian process classifier	99.7252	98.7179	1.2820	98.6813
AdaBoost classifier	93.1318	94.8717	5.1282	94.5054
Gaussian naïve bayes	89.0109	91.0256	8.9743	90.7692

6.1 Classification Report

Classification report is the report of performance of the machine learning model in the tabular format. It contains the results of metrics on which evaluation of the model had been done. In classification report, it has values of accuracy, precision, and F1-score which gives the detailed idea about the performance of the classification algorithm. Classification reports of various models are shown below:

Logistic Regression

	Precision	Recall	F1-Score	Support
Positive	0.97	0.88	0.92	65
Negative	0.92	0.98	0.95	91
Accuracy			0.94	156
Macro avg	0.94	0.93	0.93	156
Weighted avg	0.94	0.94	0.94	156

Random Forest

	Precision	Recall	F1-Score	Support
Positive	1.00	0.98	0.99	65
Negative	0.99	1.00	0.99	91
Accuracy			0.99	156
Macro avg	0.99	0.99	0.99	156
Weighted avg	0.99	0.99	0.99	156

Support Vector Machine (SVM)

	Precision	Recall	F1-Score	Support
Positive	0.94	0.92	0.93	65
Negative	0.95	0.96	0.95	91
Accuracy			0.94	156
Macro avg	0.94	0.94	0.94	156
Weighted avg	0.94	0.94	0.94	156

K-Nearest Neighbor (K-NN)

	Precision	Recall	F1-Score	Support
Positive	0.90	0.97	0.93	65
Negative	0.98	0.92	0.95	91
Accuracy			0.94	156
Macro avg	0.94	0.95	0.94	156
Weighted avg	0.94	0.94	0.94	156

Decision Tree

	Precision	Recall	F1-Score	Support
Positive	0.97	1.00	0.98	65
Negative	1.00	0.98	0.99	91
Accuracy			0.99	156
Macro avg	0.99	0.99	0.99	156
Weighted avg	0.99	0.99	0.99	156

Gaussian Process Classifier

	Precision	Recall	F1-Score	Support
Positive	0.98	0.98	0.98	65
Negative	0.99	0.99	0.99	91
Accuracy			0.99	156
Macro avg	0.99	0.99	0.99	156
Weighted avg	0.99	0.99	0.99	156

Adaboost Classifier

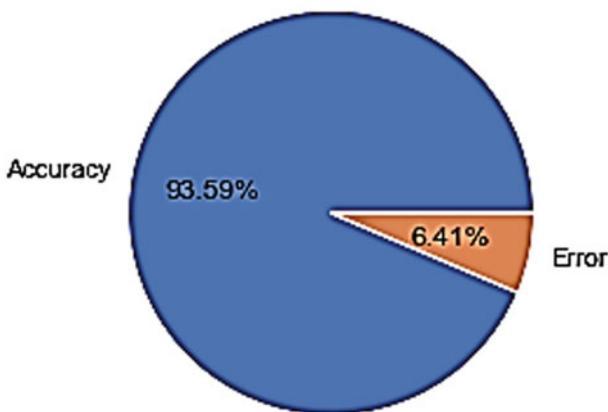
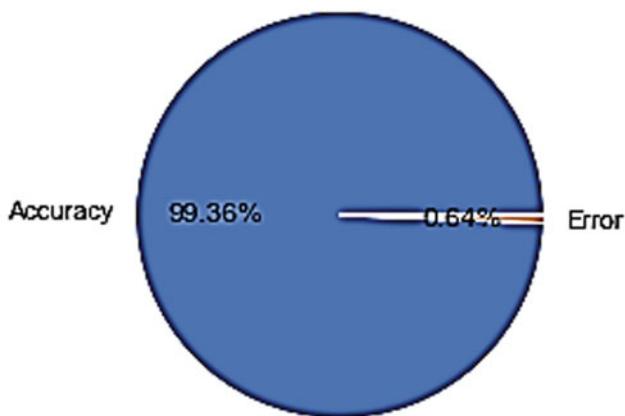
	Precision	Recall	F1-Score	Support
Positive	0.95	0.92	0.94	65
Negative	0.95	0.97	0.96	91
Accuracy			0.95	156
Macro avg	0.95	0.95	0.95	156
Weighted avg	0.95	0.95	0.95	156

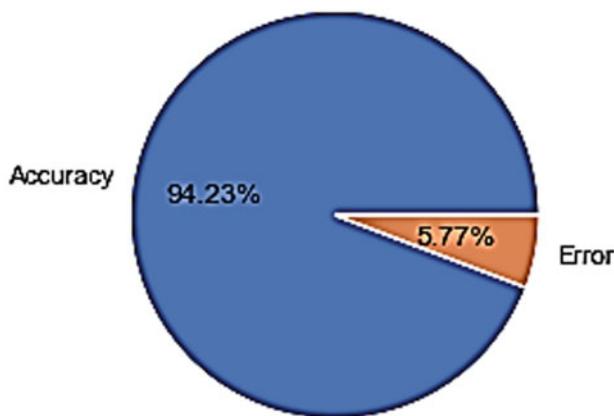
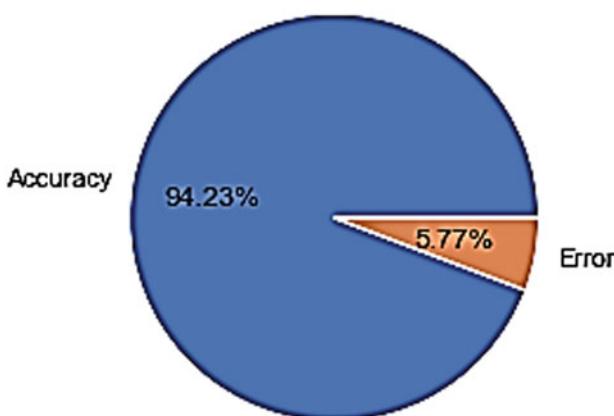
Gaussian Naïve Bayes

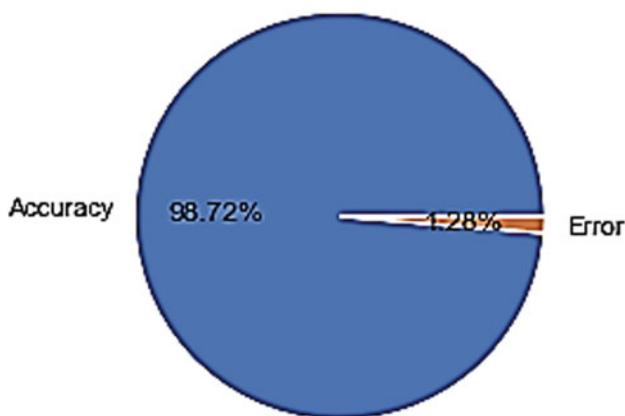
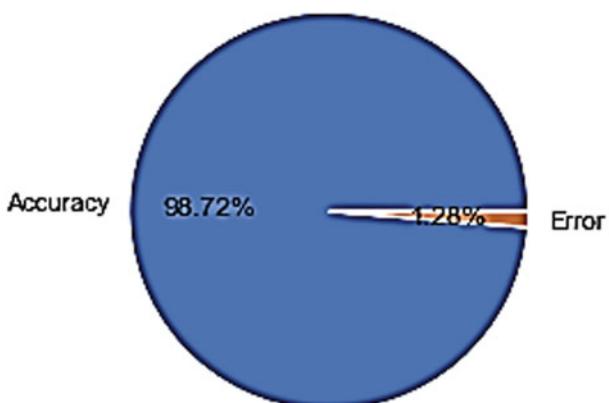
	Precision	Recall	F1-Score	Support
Positive	0.89	0.89	0.89	65
Negative	0.92	0.92	0.92	91
Accuracy			0.91	156
Macro avg	0.91	0.91	0.91	156
Weighted avg	0.91	0.91	0.91	156

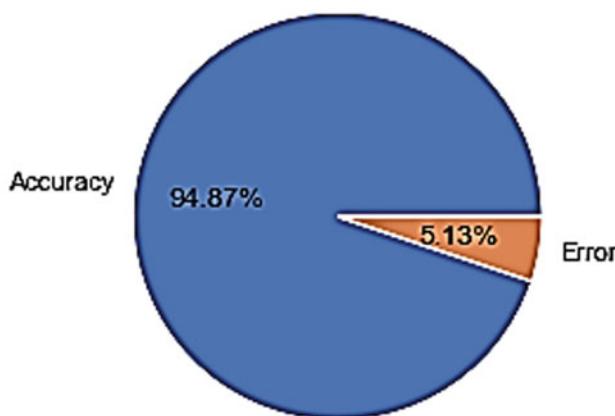
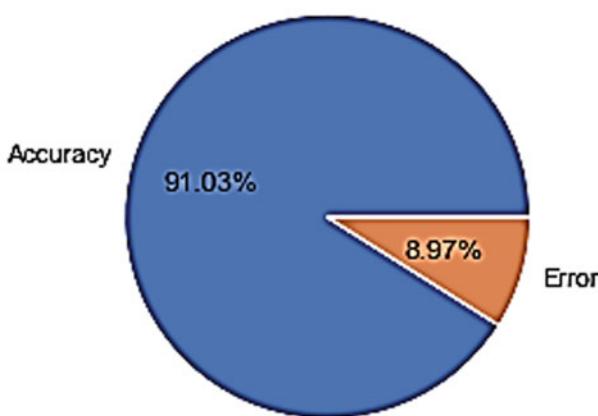
6.2 Graphical Presentation of Accuracy and Error

The following pie chart represents a graphical picture of accuracy and error found during the testing of different proposed models. Percentage of accuracy and error denotes a clear concept of productive working of each and every model.

LOGISTIC REGRESSION - Depiction of Accuracy and Error**RANDOM FOREST CLASSIFIER - Depiction of Accuracy and Error**

SUPPORT VECTOR MACHINE - Depiction of Accuracy and Error**K - NEAREST NEIGHBOR - Depiction of Accuracy and Error**

DECISION TREE - Depiction of Accuracy and Error**GAUSSIAN PROCESS CLASSIFIER - Depiction of Accuracy and Error**

ADABOOST CLASSIFIER - Depiction of Accuracy and Error**GAUSSIAN NAÏVE BAYES - Depiction of Accuracy and Error**

6.3 Confusion Matrix

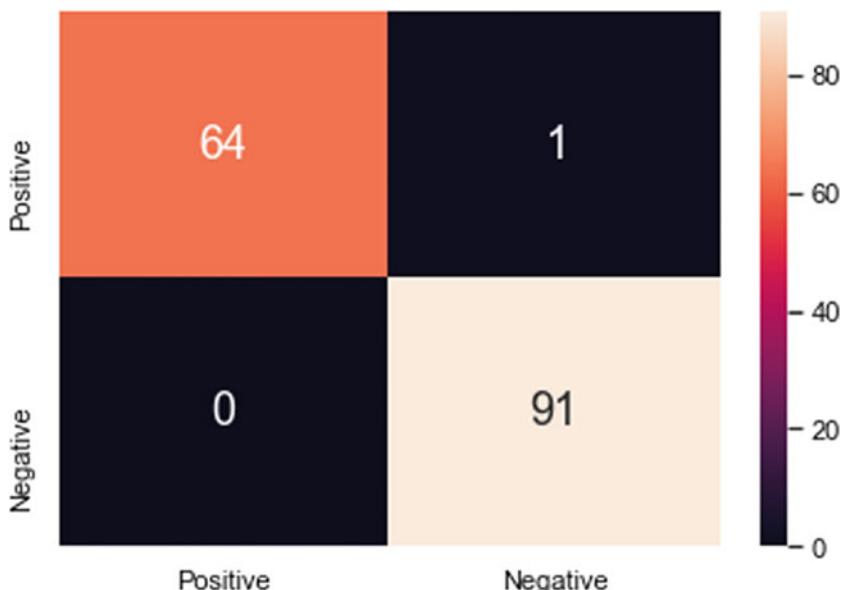
Confusion Matrix is a 2×2 matrix, which used to show the performance of the test dataset for which actual values are known in the schematic format. In this, two types of errors are shown in the diagram. The top right corner in the following confusion matrix depict “Type I” error which means patient was actually non-diabetic but model had predicted it as diabetic patient. And the bottom left corner of confusion matrix depicts “Type II” error which shows diabetic patient was predicted as non-diabetic

by the model. Following are the confusion matrix for different machine learning algorithms.

Logistic Regression

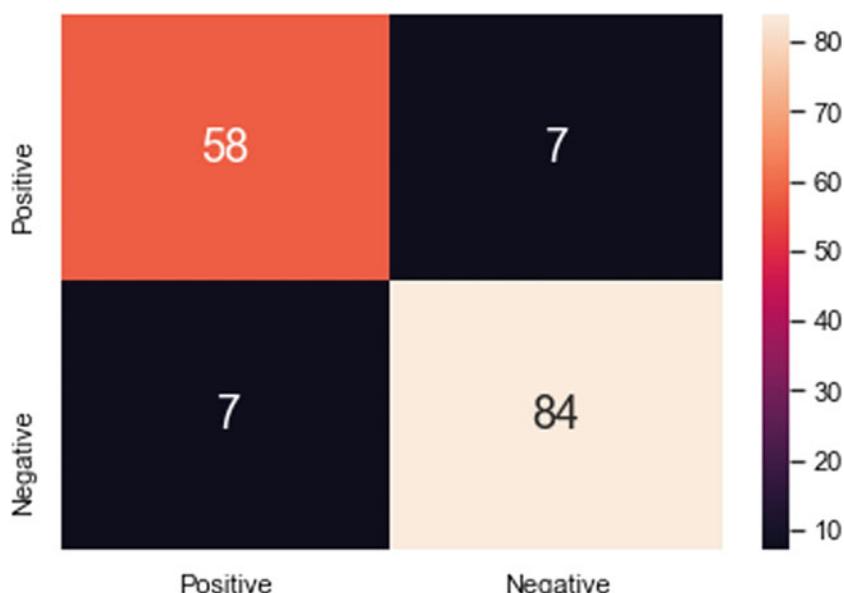


Random Forest



Support Vector Machine (SVM)**K-Nearest Neighbors (K-NN)**

Decision Tree**Gaussian Process Classifier**

Adaboost Classifier**Gaussian Naïve Bayes**

Various classification algorithms were used in the dataset of diabetes. Each and every algorithm was different, because the mathematical logic used for every algorithm was different. Evaluation of model was done based on the subsequent metrics:

- i. Accuracy
- ii. Precision
- iii. Error
- iv. Support
- v. Area Under Curve (AUC)
- vi. Recall
- vii. Macro Average
- viii. F1-score
- ix. Weighted Averages

So, considering these metrics final conclusion has been taken out. Many machine learning algorithms were used and plenty of them gave pretty good accuracy. Logistic Regression gave accuracy of almost 93.59% with error 6.41%. Decision Tree performed well with 98.71% accuracy with error 1.28%. Gaussian Naïve Bayes with accuracy 91.03% and error 8.97%. AdaBoost Classifier with accuracy of 94.87% and error 5.13%. Support Vector Machine (SVM) with 94.23% accuracy and 5.77% error. K-Nearest Neighbors also gave 94.23% with 5.77% error and Gaussian Process Classifier also performed well with 98.71% accuracy and 1.28% error.

But, among all of these machine learning algorithms, Random Forest Classifier gave the highest accuracy of 99.4% with precision of 99.4%, recall of 99.23%, and error of just 0.6%. Random Forest algorithm proved to be the best as compared to other because it gave accuracy of 99.4%. As random forest takes the random subset to build each tree, it had not suffered from high number of predictors. The confusion matrix also shows that random forest had achieved great accuracy, as it had accurately identified all the patients in the test dataset set with no Type II error and only one patient in Type I error. It means patient was non-diabetic but model had predicted it as a diabetic patient.

7 Conclusion

For this research, the patient dataset is collected from Sylhet Diabetes Hospital, Bangladesh. During this study, classification has been done by applying various machine learning algorithms, namely, Logistic Regression, Random Forest Classifier, Support Vector Machine, Decision Trees, K-Nearest Neighbors, Gaussian Process Classifier, AdaBoost Classifier, and Gaussian Naïve Bayes were used to create the model for carrying out the diagnosis of diabetes. Furthermore, the machine learning algorithm is tested by evaluating the performance in terms of accuracy. Our comparative analysis is done on the features in dataset and it also shows Random Forest Classifier is the best algorithm for the prediction of newly created datasets made for early stage diabetic risk prediction because it gives the most effective fit to the data

with respect to the diabetic and non-diabetic patients. So, conclusion of this paper is that by applying Random Forest Classifier algorithm with some hyper parameter optimization on diabetes dataset we are able to notice performance of the Random Forest Classifier had achieved higher accuracy. The preferred algorithm technique can be incredibly useful for diabetes prediction. Each prediction of diabetes would be of lot helpful for patient's health. Finally, this research will be extended further by optimizing hyperparameters and considering only important attributes of diabetes dataset to reinforce the performance of model by increasing its accuracy and also to predict in next few years possibility of diabetes to non-diabetic patients.

References

1. Kalyankar, G.D., Poojara, S.R., Dharwadkar, N.V.: Predictive analysis of diabetic patient data using machine learning and hadoop. In: International Conference On I-SMAC (2017). ISBN 978-1-5090-3243-3
2. Komi, M., Li, J., Zhai, Y., Zhang, X.: Application of data mining methods in diabetes prediction. In: Image, Vision and Computing (ICIVC), 2017 2nd International Conference on, pp. 1006–1010. IEEE (2017)
3. Perveen, S., Shahbaz, M., Guergachi, A., Keshavjee, K.: Performance analysis of data mining classification techniques to predict diabetes. *Proced. Comput. Sci.* **82**, 115–121 (2016). <https://doi.org/10.1016/j.procs.2016.04.016>
4. Nai-Arun, N., Sittidech, P.: Ensemble learning model for diabetes classification. *Adv. Mater. Res.* **931–932**, 1427–1431 (2014). <https://doi.org/10.4028/www.scientific.net/AMR.931-932.1427>
5. Orabi, K.M., Kamal, Y.M., Rabah, T.M.: Early predictive system for diabetes mellitus disease. In: Industrial Conference on Data Mining, pp. 420–427. Springer (2016)
6. Priyam, A., Gupta, R., Rathee, A., Srivastava, S.: Comparative analysis of decision tree classification algorithms. *Int. J. Current Eng. Technol.* **3**, 334–337, 2277–4106 (2013). [arXiv: ISSN](https://arxiv.org/abs/1308.2001)
7. Esposito, F., Malerba, D., Semeraro, G., Kay, J.: A comparative analysis of methods for pruning decision trees. *IEEE Trans. Pattern Anal. Mach. Intell.* **19**, 476–491 (1997). <https://doi.org/10.1109/34.589207>
8. Pradhan, M., Bamnote, G.R.: Design of classifier for detection of diabetes mellitus using genetic programming. *Adv. Intell. Syst. Comput.* **1**, 7630770 (2014). <https://doi.org/10.1007/978-3-319-11933-5>
9. Sharief, A.A., Sheta, A.: Developing a mathematical model to detect diabetes using multigene genetic programming. *Int. J. Adv. Res. Artif. Intell. (IJARAI)* **3**, 54–59 (2014). <https://doi.org/10.14569/IJARAI.2014.031007>
10. Mandal, S., Balas, V.E., Shaw, R.N., Ghosh, A.: Prediction analysis of idiopathic pulmonary fibrosis progression from OSIC dataset. In: 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, India, pp. 861–865 (2020). <https://doi.org/10.1109/gucon48875.2020.9231239>
11. Kumar, M., Shenbagaraman, V.M., Ghosh, A.: Predictive data analysis for energy management of a smart factory leading to sustainability Book Chapter, Springer. In: Favorskaya, M.N., Mekhilef, S., Pandey, R.K., Singh, N. (eds.) Innovations in Electrical and Electronic Engineering, pp. 765–773 (2020). ISBN 978-981-15-4691-4
12. Han, J., Rodriguez, J.C., Beheshti, M.: Discovering decision tree based diabetes prediction model. In: International Conference on Advanced Software Engineering and Its Applications, pp. 99–109. Springer (2008)

13. Mandal, S., Biswas, S., Balas, V.E., Shaw, R.N., Ghosh, A.: Motion prediction for autonomous vehicles from lyft dataset using deep learning. In: 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2020, pp. 768–773. <https://doi.org/10.1109/iccca49541.2020.9250790>
14. UCI—Machine Learning Repository, Early stage diabetes risk prediction dataset. Data Set
15. Ho, T.K.: Random decision forests (PDF). In: Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995, pp. 278–282 (1995). Archived from the original (PDF) on 17 April 2016. Retrieved 5 June 2016
16. Ho, T.K.: The random subspace method for constructing decision forests (PDF). IEEE Trans. Pattern Anal. Mach. Intell. **20**(8), 832–844 (1998). <https://doi.org/10.1109/34.709601>

A Low Cost and Enhanced Assistive Environment for People with Vision Loss



Tushar Vashisth, Ritika Khareta, Nishi Bhati, Venkata Chanakya Samsani, and Shantu Sharma

Abstract Usually, people with vision disabilities can be seen using white canes, which are static devices and are used for navigation around. The limitation of commonly used static mobility devices is their inability to perceive the surroundings around them. To provide a real-time visual experience to people with vision loss, a system with dynamic abilities is required. With this aim, in this paper, an advanced vision-enhancing architecture is proposed for visually impaired people to provide them a real-time experience of the surroundings along with navigation assistance through voice rendering. The framework of the proposed system is based on advanced technologies like IoT (Internet of Things), Machine learning, and Computer Vision. The proposed architecture is aimed to build a low-cost wearable device that will act as a guiding agent for a visually impaired person so that the individual can get to know about the things around with precision and ease both in an indoor as well as outdoor environment.

Keywords Assistive devices · Vision loss · IoT · Wearable devices · Machine learning · Object detection · Voice processing

1 Introduction

During our lifetime, we come across many disabled people, especially at bus stops, markets or even in parks. According to a report by WHO on blindness, around 15% of the world's population lives with a form of disability or other [1]. And this global estimate on people living with disabilities is on the rise despite better medical facilities and average higher standards of living. According to the recent Census of 2011 [2], there are 2.2% of individuals in India, who are suffering from a kind of disability or other. Vision disability stands tall at 48.5% among all the major disabilities that are present in India. With these statistics, it is easy to analyze that vision impairment indeed emerges as the top category [3].

T. Vashisth · R. Khareta · N. Bhati (✉) · V. C. Samsani · S. Sharma

Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Uttar Pradesh, Noida, India

Visual Impairment takes away the opportunity of seeing the world and people with this disability face challenges and in doing day-to-day chores and even in identifying common objects every day. They need proper navigation and training for even identifying basic things. They also face difficulty in knowing what's happening around them and they also are prevented from knowing basic social services persons like police, doctors etc. and this also applies for animals like dogs, cats etc. [3]. In the past decade, a boom in technologies like IoT, big data, high performance computing, Machine learning (ML), etc., can be seen. Researchers are exploring all possible options for incorporating these advanced technologies to provide better ease of living [4–8]. With the help of IoT and ML technology, various low-cost devices have already been developed in the field of home automation, health, agriculture, etc. People are now used to smart devices like smart watch, smart TV, smart traffic monitoring, etc., in their daily life. Researchers have also explored options for providing assistive solutions to visually impaired people such as smart canes, glasses, etc. [4, 9, 10]. But as the technology is continuously expanding, by the inclusion of new technologies like deep learning, edge computing, AIOT, etc., there is always a scope for the development of better and optimized solutions.

In this paper, the focus is on providing a low cost and more enhanced solution to visually impaired people. The paper is structured to provide knowledge about the various existing assistive devices along with the architecture of low cost and enhanced environment for people with vision loss. In Sect. 2, existing systems and work done in the related field are discussed and analyzed based on the feature of the mentioned systems. The features, architecture, required hardware components, and cost analysis of the proposed system are presented in Sect. 3. The paper is concluded in Sect. 4 by providing the advantages of the proposed architecture in comparison to some existing systems.

2 Assistive Technology for Visual Loss

Vision Disability refers to the partial or full loss of eyesight, thus decreasing the ability of a person to see that usually needs specialized devices to be fixable [2]. Till now a lot of work has been done in the field of assistive technology for providing ease of living to visually disabled people. Assistive Devices are technologies that usually help in maintaining and improving one's functionality and make one independent enough in order to participate easily and feasibly in daily life tasks and chores and are usually used by people with certain ailments, impairments, and secondary health conditions such as wheelchairs, prostheses, visual aids, hearings aids, etc. [11]. Assistive systems for visual loss are designed in such a way to provide help to the people with any kind of partial or full visual loss or any such disability. These systems include screen readers to magnifiers for people with low vision. Many eye trackers and visual aids are also available commercially for the assistance of visually impaired people. And many other systems are there that may help them to read and write without any difficulty and these systems are easy to handle. These aids serve different purposes

and work on different technologies. For example, in [12], an advanced real-time mobile robot system for obstacle avoidance is used to provide assistance for visually impaired people. A speech guidance-based system is presented in [13], for blind people to walk independently. A system BlindeDroid is also proposed by authors for real-time guiding of blind people [15]. Voice-based navigation systems using different technologies like bluetooth, GPS, and ultrasonic sensors can also be seen in literature [14, 16].

There is a lot of competition in the market to provide a more enhanced and optimized experience to their users. With the emergence of new technologies like deep learning, edge computing, AIOT, etc., researchers are also trying to incorporate new technology into existing systems for providing more optimized results. With the help of new technologies, the architecture of wearable devices is proposed in the next section, with the aim of providing a real-time assistance and guidance with a more enhanced experience to visually impaired people.

To analyze the proposed architecture in terms of cost, functionalities, and technologies, some of the recent work done in the field of assistive technology and some commercially available devices for visually disabled people are analyzed and summarized in Tables 1 and 2.

3 Proposed System

A study done in the previous section shows that various devices have been developed for assisting visually impaired people. Different devices are having different functionalities, but as new technology is getting included in every domain, there is always a scope for better and optimized solutions. In this section, an architecture of a low-cost device is proposed with the aim of providing a more enhanced and better experience to blind people. Various functionalities of the proposed system, its working, hardware-software requirements, and proposed cost for its development are discussed in further sub-sections.

3.1 *Functionalities of Proposed System*

A person with vision impairment can use the device as wearable headgear in form of spectacles with an Ultrasonic Sensor for Obstacle detection and a camera for Image detection. The device is proposed with two battery modes, one normal with all functionalities and the other is battery saver by preventing some features of the devices. Various functionalities of the proposed system are presented in Fig. 1 and described below.

- **Object Identification:** Everyday objects like Cup, Car, Street Lights, Chairs, etc., identification using existing datasets.

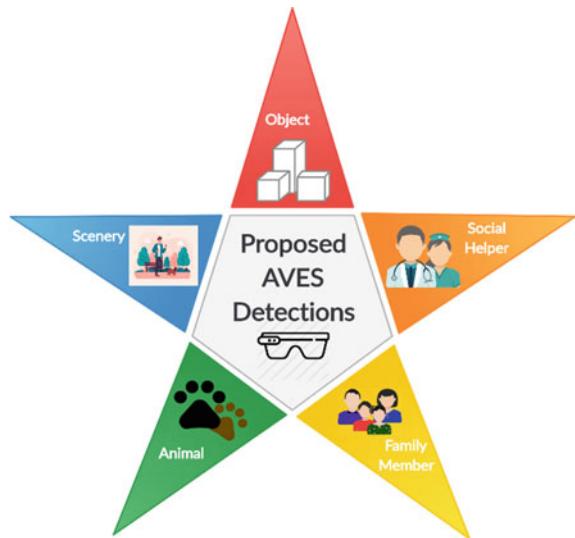
Table 1 Analysis of related work

Refs.	Description	Dataset/Methods/Hardware Used	Analysis
[17]	Authors proposed a system for determining the user's position, a detailed map of the building, and a mobile app with a voice-operated outdoor system for visually impaired people	Ultrasonic sensors, microcontroller, vibrator motor, and ISD2560 Chip Corder, speaker	The system is able to detect the objects by providing the vibration and voice-based feedback to users, but it don't provide any information about the object that what exactly the object is, and the distance of the object from the user
[18]	Authors proposed an Arduino based obstacle detector, which can detect objects over a wide range and give output as a buzzer sound	Arduino UNO, Ultrasonic sensor, Buzzer, Power bank	The mentioned system is only effective in detecting objects that are in a specific range. No provision for object identification is mentioned as the output is a buzzer sound which neither tells the distance nor the size of the obstacle
[19]	Here authors tried to transform the visual world into an audio world for blind potential	Portable camera device, 3D, sound generation application based on Unity game engine, YOLO	The prototype works well for an average distance of 3–4 m, giving the user both object type and distance accurately. But it is unable to detect objects too close or too far and also suffers from information overload in certain instances
[20]	Authors proposed a system for obstacle detection and distance measurement procedure	PIC16F877A microcontroller, Bluetooth module, headphone	This system uses three sensors, placed on the user's body. It can detect any obstacle in front and holes on the surface. The arrangement of the sensors is very complex
[21]	Authors proposed a device that captures an image and converts the text written on the image into speech	Optical Character Recognition Software, Text-to-Speech engines, Raspberry Pi camera, Raspberry Pi	The system works well while converting text to speech. But does not identify objects, so the functionality is limited

- **Animal Identification:** Street or Domestic animals like Cats, Dogs, Horses, Cows, etc., identification.
- **Specific People Identification:** Specific people like Police Officers, Army Officers, Doctors, Firefighters, Cleaners, etc., in India identification with the help of created datasets.

Table 2 Analysis of commercially available systems

Device	Functionality	Limitations
Arduino smart cane for the visually impaired [22]	This could help them walk without tripping and avoid serious injuries	Can only detect objects at a certain angle, so this product may not be able to detect objects some objects
TapTapsee application [23]	It utilizes the device's camera and voice over functions to take a photo or record a video of anything and identify it out loud	It works on captured images only, not in real time. So, this product is difficult for blind people to use
Smart shoes [24]	Shoe sync with a user's phone and an app that piggybacks on Google Maps allowing the shoes to keep track of where the user is going	Very expensive
OrCamMyEye [25]	It has facial recognition as well as object recognition, which makes this product very helpful for blind people to use	Not available in India and very expensive

Fig. 1 Functionalities of the proposed system

- **Family Member Identification:** To provide a personalized experience for the user, the device will have a mode to allow detection of family members with the help of personalized dataset creation.
- **Scenery Visualization:** Based on the user choice, the user can get a description of whole scenario in front of him with voice-based feedback using advanced technology like deep learning and video processing etc.

- **Text to Speech Conversion:** Text to speech conversion feature is provided to read aloud text written in front of the user.
- **Navigation System:** Sensing the obstacles in the surrounding environment for the visually impaired person and warning them about the obstacle.
- **Dual Battery Mode:** This feature can be used by the user as per their need. The user has two options to change the mode of the battery used in the device. The ‘Normal’ mode can work for all the detections mentioned whereas the ‘Battery saver’ mode works only for Object detection. With this, the user can save the battery life of the device in order to work for the whole day.

3.2 Working of Proposed System

The proposed device is a wearable device with a combination of a camera and ultrasonic sensor to provide the above-mentioned features. The device will have one activation mode to switch on/off the device and one battery mode to operate the device in normal and battery saver mode. The system software will be connected to a whole bunch of databases trained and tested for the detection of various real-level things. The basic functioning of the system is presented in Fig. 2 and is described below:

- The device starts with the help of a power button. After that, the system activates and it will be in the normal mode of the dual battery which is the default mode.
- The object detection and recognition by the system is done in two different ways, i.e., using a camera and ultrasonic sensor. Basic object detection will be done using ultrasonic sensors, whereas object recognition like animal detection, social helper detection, family member detection, and scenery detection will be performed with the help of a camera.
- The required detection is selected according to the object which is in front of the device. If there is an object, then the feature selected would be object detection. When there is an animal found in front, then the feature selected would be animal detection.
- In case of human beings, there are two possibilities, if the person is wearing any particular uniform like a doctor, policeman or a nurse, then the feature selected would be social helper detection, whereas a person who is in the family is detected by family member detection. And the last feature where the user is facing a particular scenario, then scenery detection takes place.

3.3 Hardware Software Requirements of Proposed System

This project focuses on being cost-efficient so that people can afford the device and also to make the device easily available in the market. Every device needs some hardware and software components in order to have the perfect functionality. Some

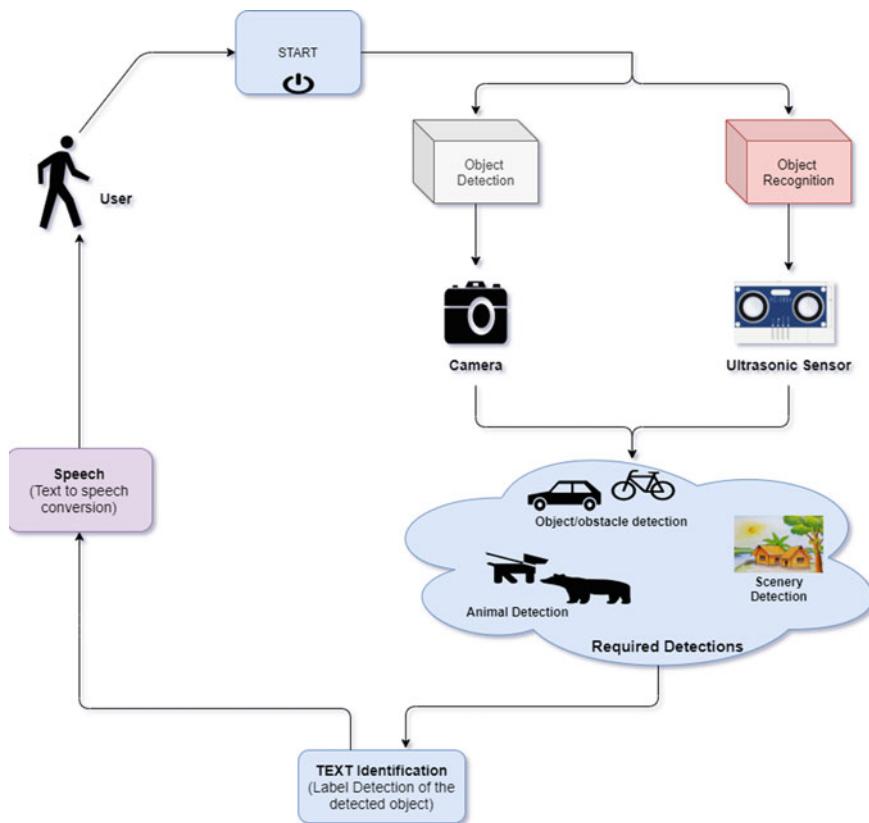


Fig. 2 Flow diagram of the proposed system

of the major hardware components which are required for the proposed systems are described in Table 3 along with its description and advantages

3.4 Cost Analysis of Proposed System

The estimated cost of various components used for the development of the proposed architecture is analyzed in Table 4. The total cost of the product would be estimated in between Rs. 5000–6000 in India.

Table 3 Hardware description of products used in proposed system

S. No	Device image	Description	Advantages/Use
1		ATmega328P based microcontroller board 'Arduino UNO' has a 16 MHz ceramic resonator, 14 digital input and output pins of which 6 can be considered as PWM outputs, ⁶ analogue inputs, USB connection, a power jack, a reset button and an ICSP header. Recommended input voltage is between 7 to 12 V, but the operating voltage is about 5 V. The length of the board is 68.6 mm, the width of the board is 53.4 mm. The total weight of the board is 25 g	<ul style="list-style-type: none"> Inexpensive Cross-platform Open source and extendable software/hardware Simple and clear programming Better RESET circuit
2		Among the Raspberry Pi range of computers the Raspberry Pi 4 Model B is the latest product which provides innovative increases in the speed of processor, memory, performance of multimedia, and connectivity. This model produces a performance similar to desktop. A 64 bit high performance Quad Core processor, dual display support at 4 K resolutions through a pair of micro-HDMI ports, choice of variable RAM up to 8 GB, hardware video decoder, along with Bluetooth 5.0, Gigabit Ethernet, dual band wireless LAN, USB 3.0 and a PoE capability with the separate PoE HAT add-on, are the major features of the product	<ul style="list-style-type: none"> The performance speed and the processing of the unit are much better Two 4 K monitors are supportable by the unit Inclusion of USB 'C' type power supply and USB 3.0 makes it more useful The availability of no of ports compared to previous versions is more

(continued)

Table 3 (continued)

S. No	Device image	Description	Advantages/Use
3		The ultrasonic sensor which is used in this architecture is HC-SR04 which gives a non-contact measuring functionality with a range of 2–400 cm with an accuracy of 3 mm. This ultrasonic sensor consists of an ultrasonic transmitter, a receiver and a control circuit. The sensor setup is easy and can be used in many projects. The sensor also has an ability to avoid the inconsistent bouncy data according to the application. The unit has an Operating voltage of 5 V DC, and the operating current is about 15 mA, the measuring angle of the unit is about 15°	<ul style="list-style-type: none"> The sensors are easy to use Interface with microcontroller is easy Can easily detect the shape, type and the orientation of the object
4		The camera model which is used in this proposed device is OV5647 sensor video webcam with a resolution of 5 MP 1080p. The camera module has a ribbon cable which is directly connectable to the Raspberry Pi 4 board. The still picture resolution of the camera module is 2592×1944 and the viewing angle is about 72°	<ul style="list-style-type: none"> Tiny size Easy to connect
5		In today's generation there are many varieties of batteries available. The combination of four rechargeable AA batteries at 1.25 V is used to provide the required voltage in proposed device. These batteries can be held together with the help of a battery case. The battery case also has an on and off switch	

Table 4 Cost analysis of proposed system

Component name	Qty.	Description	Approximate cost
Arduino UNO	1	Main board	Rs. 330–350
Raspberry Pi 4	1	Main board	Rs. 4000
Ultrasonic sensor	1	For object detection and distance	Rs. 60–100
Camera	1	For image recognition	Rs. 300
Batteries and battery case	4 + 1	Power supply of the device	Rs. 450–500
Wires	20–25 cm	Connection purposes	Rs. 20–40
USB cable	1	Connection purposes	Rs. 100–150
Glass frame	1	Main frame of the device	Rs. 100–150

4 Conclusion and Future Work

In this paper, the real-time problems faced by visually disabled people were discussed, followed by a discussion on various projects undertaken, devices, and related work done to assist visually impaired people. After the detailed discussion on the background along with the merits and demerits of various projects, an efficient architecture is proposed with the help of deep learning, IoT, and image processing. The main objective of this work was to propose an efficient device with the necessary functionalities so that more and more people can afford the end product. The idea focuses on providing the blind with a portable and real-time device. Different proposed and market-ready technologies were studied and a device best suited, especially for the disabled, who have total vision loss, is proposed in this paper. The device includes all the necessary functionalities. The budget of the proposed device would be low cost to increase the reach of the device.

References

- WHO Article on Blindness and Vision Impairment. <https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment>. Accessed 10 Nov 2020
- Census Data. www.censusindia.gov.in › 2011 census › Disability_Data. Accessed 12 Nov 2020
- Dandona, R., Pandey, A., George, S., Kumar, G.A., Dandona, L.: India's disability estimates: limitations and way forward. *PLoS ONE* **14**(9), e0222159 (2019). <https://doi.org/10.1371/journal.pone.0222159>
- Artificial Intelligence Helping the Visually Impaired. <https://www.cognixia.com/blog/artificial-intelligence-helping-the-visually-impaired>
- Nayyar, A., Puri, V., Nguyen, N.G.: BioSenHealth 1.0: a novel internet of medical things (IoMT)-based patient health monitoring system. In: Bhattacharyya, S., Hassanien, A., Gupta, D., Khanna, A., Pan, I. (eds.), International Conference on Innovative Computing and Communications. Lecture Notes in Networks and Systems, vol. 55. (Springer, Singapore, 2019). https://doi.org/10.1007/978-981-13-2324-9_16
- Alzubi, J.A., Manikandan, R., Alzubi, O.A., Gayathri, N., Patan, R.: A survey of specific IoT applications. *Int. J. Emerg. Technol.* (2019). <https://doi.org/10.14419/ijet.v7i2.7.11089>

7. Alzubi, J.A., Selvakumar, J., Alzubi, O.M., Manikandan, R.: Decentralized internet of things. Indian J. Public Health Res. Develop. (2019). <https://doi.org/10.5958/0976-5506.2019.00295.X>
8. Pandey, R.S., Upadhyay, R., Kumar, M., Singh, P., Shukla, S.: IoT-based help age sensor device for senior citizens. In: Khanna, A., Gupta, D., Bhattacharyya, S., Snasel, V., Platos, J., Hassani, A. (eds.), International Conference on Innovative Computing and Communications. Advances in Intelligent Systems and Computing, vol. 1059. (Springer, Singapore, 2020). https://doi.org/10.1007/978-981-15-0324-5_16
9. Ilag, B.N., Athave, Y.: A design review of smart stick for the blind equipped with obstacle detection and identification using artificial intelligence. Int. J. Comput. Appl. **182**, 55–60 (2019). <https://doi.org/10.5120/ijca2019918768>
10. Pruthvi, S., Pushyap, S.N., Ravin, R.M., Kumar, S.S., Tiwari, S.: Smart blind stick using artificial intelligence. Int. J. Eng. Adv. Technol. (IJEAT) ISSN **8**(5S), 2249–8958 (2019)
11. Teodorescu, H.N., Jain, L.: Intelligent systems and technologies in rehabilitation engineering (2001)
12. Shepherd, I.: Providing learning support for blind and visually impaired students undertaking fieldwork and related activities (2001)
13. Gharat, M., Patanwala, R., Ganaparthi, A.: Audio guidance system for blind Int. Conf. Electron. Commun. Aerosp. Technol. (ICECA) Coimbatore **2017**, 381–384. <https://doi.org/10.1109/ICECA.2017.8203710>
14. Rajan, K., Kalaiselvan, E.: Intelligent navigation system for blind people with real time tracking. Int. J. Eng. Res. Technol. **3**(22) (2015)
15. Cecilio, J., Duarte, K., Furtado, P.: BlindeDroid: an information tracking system for real-time guiding of blind people Proced. Comput. Sci. **52**, 113–120 (2015). <https://doi.org/10.1016/j.procs.2015.05.039>
16. Dabrowski, A., Kardys, P., Marciniak, T.: Bluetooth technology applications dedicated to supporting blind and hearing as well as speech handicapped people. In: ELMAR, 47th International Symposium, June 2005, pp. 295–298 (2005)
17. Mahmud, N., Saha, R.K., Zafar, R.B., Bhuiyan, M.B.H., Sarwar, S.S.: Vibration and voice operated navigation system for visually impaired person. In: 2014 International Conference on Informatics, Electronics & Vision (ICIEV), Dhaka, pp. 1–5 (2014). <https://doi.org/10.1109/ICIEV.2014.6850740>
18. Maragatharajan, M., Jegadeeswaran, G., Askash, R., Aniruth, K., Sarath, A.: Obstacle detector for blind peoples. Int. J. Eng. Adv. Technol. (IJEAT) ISSN **9**(1S4), 2249–8958 (2019)
19. Jiang, R., Lin, Q., Qu, S.: Let blind people see: real-time visual recognition with results converted to 3D audio (2016)
20. Islam, J.H., Sanjidul, H., Jamil, H.: Guidance system for visually impaired person (2016)
21. Hagargund, A.G., Thota, S.V., Bera, M., Shaik, F.E.: Image to speech conversion for visually impaired. Int. J. Latest Res. Eng. Technol. (IJLRET) **3**(6) (2017)
22. Dada, E., Shani, A., Adekunle, A.: Smart walking stick for visually impaired people using ultrasonic sensors and arduino. Int. J. Eng. Technol. **9**, 3435–3447 (2017). <https://doi.org/10.21817/ijet/2017/v9i5/170905302>.
23. Introduction of TapTapSee App. <https://www.taptapseeapp.com>. Accessed 20 Nov 2020
24. Project: Smart Shoes for Blind Person. <https://innovate.mygov.in/innovation/smart-shoes-for-blind-person/>. Accessed 9 Oct 2020
25. Orcam MyEye-Overview. www.orcam.com/en/myeye2. Accessed 5 Oct 2020

A Comparative Study of Myocardial Infarction Detection from ECG Data Using Machine Learning



Aritra Chakraborty, Santanu Chatterjee, Koushik Majumder,
Rabindra Nath Shaw, and Ankush Ghosh

Abstract Myocardial Infarction (MI) is a life-threatening heart disease, timely medical intervention of which can reduce the mortality rate. It can be detected from Electrocardiogram or ECG. Diagnostic methods of this disease by clinical approaches are typically invasive. They also do not fulfill the detection accuracy, and there is a chance of human error. In the medical field, machine learning techniques have great potential for disease diagnosis. We can achieve accurate detection from ECG by using deep learning methods. In this paper, we did a comprehensive comparative study of a few existing proposals with different approaches to detect and predict Myocardial infarction. And then we proposed a deep learning approach for future implementation in order to achieve superior performance.

Keywords Myocardial infarction · ECG · Machine learning · Deep learning · Autoencoder

1 Introduction

Cardiovascular diseases (CVD) are one of the predominant reasons responsible for mortality globally. Based on one of the reports of the World Health Organization, there are 30% of the total number of deaths, approximately 17 million per year (and growing) due to this reason [1]. In today's modern lifestyle, the causes of risk such as use of tobacco, unhealthy food habit, physical lethargy, and alcohol consumption increase the CVD [1] mortality rate. It is projected that the CVD mortality rate will

A. Chakraborty · S. Chatterjee · K. Majumder (✉)
Maulana Abul Kalam Azad University of Technology, Kolkata, India
e-mail: koushik@ieee.org

R. N. Shaw
Department of Electrical Electronics & Communication Engineering, Galgotias University,
Noida, India
e-mail: r.n.s@ieee.org

A. Ghosh
School of Engineering and Applied Sciences, The Neotia University, Kolkata, West Bengal, India

annually increase by more than 23 million by 2030 [2]. It is shown that machine learning has incredible potential for accurate disease detection and prediction. Many types of research have shown that machine learning can timely detect and predict CVDs thus reducing mortality [3–5]. This adds approaches like using measurements from electrocardiogram (ECG) to detect any exceptions in the signal by machine learning. Therefore, abnormality detection identifies the specific disease. ECG is widely used as a clinical investigation method. Also, it is a noninvasive and inexpensive [6] method. Recent works [7] have demonstrated the application of deep neural networks, especially convolution neural networks for the prediction of arrhythmia, atrial fibrillation, etc., thereby narrowing the chances of better work of the methods to detect the more fatal diseases like Myocardial Infarction.

In this article, we analyze the existing method for detecting and predicting the CVDs, more precisely Myocardial infarction from ECG signal data. A comparative study is done in the later section where we analyze the works based on the different approaches and accuracy of those models. Further, we proposed a method in the future scope where we focused on optimization of the model for less computational complexity and higher accuracy of detection.

2 Background

The Electrocardiogram is considered to be the most common clinical cardiac disease diagnosis tool because of its accessibility, noninvasiveness, and low cost. Differences in a cardiac cycle are visually represented by ECG. A signal is generated by these potential differences. These signals can be detected by using electrodes in various positions of the human body. These ECG signals are normally recorded and printed which is called an ECG report, shown in Fig. 1. Standard ECG detection is done with the 12 electrode method [8], but the 6 electrode method is also acceptable.

A typical ECG report has graphs with I, II, III, AVL, AVR, AVF, V1, V2, V3, V4, V5, and V6 parameters [9]. By analyzing these graphs, we can get the P wave, QRS complex, ST-segment, and T wave, shown in Fig. 2. Elevation, depression,

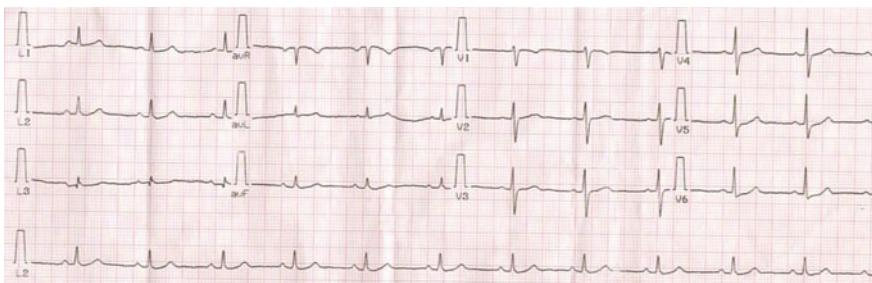


Fig. 1 Typical ECG report of a patient [9]

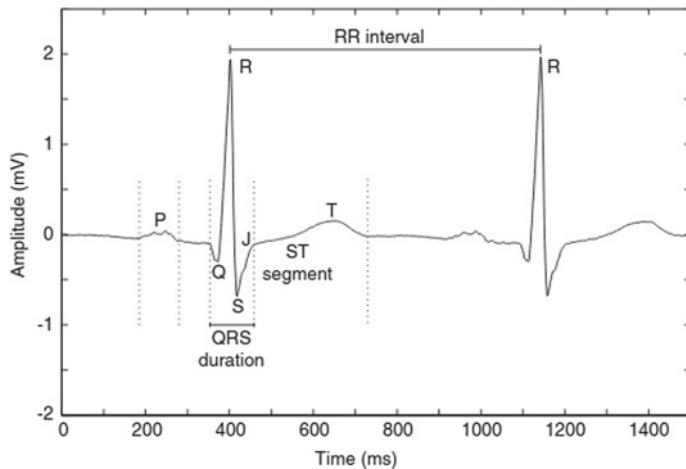


Fig. 2 Wave definitions of the cardiac cycle: the P wave, the QRS duration, the ST-segment, and the T wave, respectively, are depicted using dotted lines [6]

inversion, and abnormalities in these waveforms can indicate the actual Cardiovascular disease. Physicians use these waveforms as a reference point and detect heart abnormalities. For instance, ST-segment of a healthy heart should be isoelectric but in case of blockage in the coronary artery of the heart, this interval can be suppressed or elevated. The ST-segment is isoelectric by nature. It is located after the J point, the ending of the QRS complex, and ends at the beginning of the T wave. In clinical practice, the ST-segment is used to detect myocardial infarction. It happens in case the blood supply to the coronaries is obstructed. Therefore, the blood supply is interrupted. In this situation, the ST-segment will display elevation or depression [10].

In case of myocardial infarction, a coronary is obstructed. The T wave shows an increased amplitude. In a few moments, the ST-segment elevation leads and shows the affected area of the heart and ST-segment depression shows the other side of the affected region. If the myocardial infarction continues, the T wave will grow into negative amplitude, and the Q wave's amplitude will increase, as shown in Fig. 3.

3 Related Work

Several researchers have carried out their work, aiming the detection of Myocardial infarction (MI). Different types of approaches, as well as distinct types of the dataset, are used to detect and predict the MI. In the next two sections, we have reviewed different proposals and done a comparative study of these works.

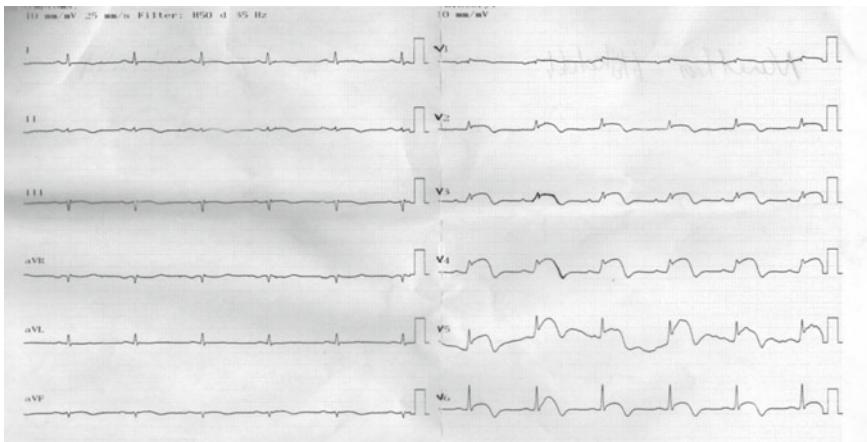


Fig. 3 ECG report with Myocardial infarction [11]

3.1 Survey of Related Works

There are different open source databases containing ECG data. PTB diagnostic ECG database from physiobank [12] is one of the most popular publicly open source available databases. There is MIT-BIH Arrhythmia, which is another popular database. ECG ViEW II [13] is a less explored database.

Dohare et al. [14] proposed a machine learning algorithm—support vector machine on PTB [12] dataset. To reduce the computational complexity and get the optimal result from the dataset, they used the principal component analysis. This method used 14 features from the total number of 220 features for detecting MI. This model has shown an accuracy of 96.66%.

Wang et al. [15] showed MI detection from the PTB dataset [12] using the R peak detection method. R-wave peak points were automatically detected on the basis of the ECG signals using the Pan-Tompkins algorithm [16]. They used the principal component analysis technique for extraction of features and reduction of dimensions. Then they used the dataset on different algorithms like random forests (RF) [17, 18], back propagation neural network (BPNN), support vector machine (SVM), and k-nearest neighbor (KNN). It is shown that the random forests algorithm gives the highest accuracy.

Baloglu et al. [19] showed MI detection from the PTB dataset [12] using deep CNN. They proposed a 10 layer deep CNN model where they used 4 one-dimensional convolutional layers, 2 one-dimensional Max Pooling layer with pool size 2, 1 dropout layer with a dropout rate of 20% and 1 flatten, 1 dense, and 1 softmax layer. All convolutional layers used a Rectified linear unit (ReLU) and the last layer used the softmax activation function. The proposed CNN model gives an accuracy of 99%.

Kora et al. [20] used the method of the hybrid firefly where they had used firefly algorithm and particle swarm optimization (FFPSO) together. For model optimization, they had used particle swarm optimization (PSO) to optimize the ECG data and implemented it on machine learning models like SVM, KNN, and artificial neural network (ANN) [21–23]. It is shown that FFPSO with ANN gives the best result.

Cho et al. [24] proposed a deep learning-based algorithm (DLA) and variational autoencoder (VAE). The deep learning was made with 7 layers where there were 2 one-dimensional convolutional layers, 2 batch normalizations, 1 max-pooling, 1 dropout layer and one flatten layer. The author developed a variational autoencoder (VAE) for performance improvement of DLA with 6-lead ECG data. This was used for image reconstruction and feature extraction. The VAE was reconstructed with an encoder, decoder composed of 6 CNN layers each. A one-dimensional dense layer connected the encoder and decoder.

Al-Zaiti et al. [25] presented a simple fusion model alongside the best low bias-low variance tradeoff. Machine learning algorithms like, Logistic Regression (LR), Gradient Boosting Machine (GBM), and ANN were used in this proposed work. They used three classifiers with 65 features from the ECG dataset. They have shown that linear prediction models like the LR model can perform equivalent to complex and computationally expensive models like ANN and GBM, assuming that linear prediction models take information from existing clinical knowledge from experts in classification decisions.

Sharma et al. [26] presented a filter bank with two-band optimal biorthogonal features for analyzing the ECG signals. ECG signals were decomposed further into six subbands with the use of a wavelet filter bank. Then they had extracted features based on entropy such as fuzzy entropy, signal-fractal-dimensions, and renyi entropy for differentiating MI and normal ECG signals. The features were used to train the KNN and the model offered an accuracy of 99.62%. This method involved the use of only a single-lead channel ECG signals with 18 features.

Al Rahhal et al. [27] proposed an approach that was based on a deep neural network for active classification of ECG signals. They used a deep neural network with Breaking-Ties (BT) and entropy. For feature extraction in an unsupervised manner, they used stacked denoising autoencoders (SDAEs) with sparsity constraints. In this work, they have shown the automatic learning using DAE for representing appropriate sparse features from raw ECG. Active Learning was used for important feature selection for model training in DNN. They had shown that selecting ECG beats for labeling directly impacted the classification accuracy, but increasing the number of hidden layers led to less accuracy. This was a drawback of this model.

Ibrahim et al. [28] presented a framework using CNN, a tree-based model—XGBoost, and recurrent neural network (RNN). Shapley values were used for identifying the most appropriate features. This reduced the computational overheads. The CNN model was designed with 12 layers in total. There were 4 one-dimensional convolutional layers using kernel sizes of 5, 3, 3, 3. In this model, there were 4 dropout layers and they had used a dropout rate of 10%, one global max-pooling layer, and 3 dense layers. For all the convolutional layers, ReLU activation functions were used, and finally a softmax activation function was used.

Kora et al. [29] presented a method of using ANN with optimization using the bat algorithm. They also proposed a method called Wavelet Coherence (WTC) for feature extraction from ECG signal. The WTC was used to calculate the similarity between two waveforms in a frequency domain. These parameters were considered as features. Bat algorithm was used to optimize these features. After that ANN was implemented on these features.

3.2 Comparative Study

In the previous section, we have inspected different works for the detection of MI. The main disadvantage of these models is computational complexity. If complexity decreases, the accuracy will decrease. Also, feature extraction is another important issue. Auxiliary data such as sex, age, and Age-adjusted Charlson Comorbidity Index (ACCI) had an impact on classification accuracy. Additionally, there are a few pieces of literature that describe the aspect of age and sex in the progress of CVD and Myocardial infarction [30–32].

Now, we have made a comparative study of these works concerning the used dataset, notable features, number of leads used at the time of data acquisition, which classifier had used, and the performance analysis of proposed works (Table 1).

4 Future Scope

In future advancement, we proposed a method based on the autoencoder-based deep learning approach, Sparse Stack Autoencoder (SSAE) [33]. Autoencoders are special types of a feed-forward neural network, where the input and output are identical. Multiple automatic encoders create the automatic encoder deep learning network (AEDLN) and a deep network, called Sparse Stack Autoencoder (SSAE) formed by multiple sparse autoencoders. Data is replicated from the input layer to the output layer in the SSAE neural network. During learning, if the output is approximately 1 or 0, then the neuron is activated or suppressed, respectively. Conventionally, the dimensions of image signal are increased after analyzing using deep learning, and many parameters optimization needed. So sparse constraints are added to the autoencoder by the sparse autoencoder [34, 35], which is a sigmoid function. It helps to control and reduce the computational complexity of deep learning classification. A constraint which is a sparse penalty like the cost function, added by the stack autoencoder. So, during training time, the number of hidden layer nodes can automatically adjust according to the features. This advantage of sparse autoencoder implemented on deep learning with more layers of sparse self encoding. Also, network learning is used for extraction expressions from features. Figure 4 shows a diagram of Sparse Stack Autoencoder.

Table 1 MI classification accuracies from different literature

Reference	Dataset	Notable feature	Number of leads	Classifier	Performance (%)
Dohare et al. [14]	PTB [12]	QRS complex, ST-T duration interval	12	SVM	Accuracy: 96.66
Wang et al. [15]	PTB [12]	R peak	12	Random Forest	Accuracy: 99.71 (inter) 85.82 (intra)
Baloglu et al. [19]	PTB [12]	R peak, ST-T segment	12	CNN	Accuracy: 99.78
Kora et al. [20]	MIT-BIH	R peak, Beat segmentation	12	FFPSO + ANN	Accuracy: 99.3
Cho et al. [24]	Hospital data	Not mentioned	6	Deep learning-based algorithm with a variational autoencoder	Accuracy: 85.4
Al-Zaiti et al. [25]	Hospital data	QRS-T angle	12	LR, GBM, ANN	Sensitivity:77 Specificity:76
Sharma et al. [26]	PTB [12]	Wavelet decomposition	12	KNN	Accuracy: 99.64 (A) 99.72 (B)
Al Rahhal et al. [27]	MIT-BIH, INCART, SVDB	Active learning	2, 12, 2	Deep neural network, denoising autoencoder	Sensitivity:95.1 Specificity:100 (Deep-Entropy) Sensitivity:98.1 Specificity:100 (Deep-BT)
Ibrahim et al. [28]	ECG View II [13]	Shapley Values analysis	12	CNN, RNN, XGBoost	Accuracy: 89.8 (CNN) 84.6 (RNN) 97.5 (XGBoost)
Kora et al. [29]	MIT-BIH	Continuous Wavelet Transform (CWT), Wavelet Coherence (WTC), Bat algorithm	12	ANN	Accuracy: 99.1

Multiple sparse autoencoders are superpositioned for constructing the SSAE. The basic construction of SSAE is shown in Fig. 5. At training time, to minimize the error of the model, each layer's output recognition signal is compared with the input signal. The SSAE model creates a hidden layer by adding sparse constraints.

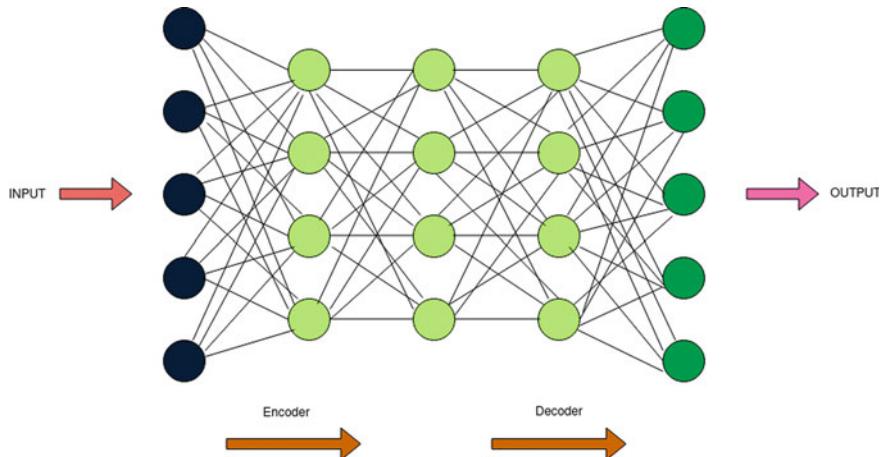


Fig. 4 Autoencoder diagram

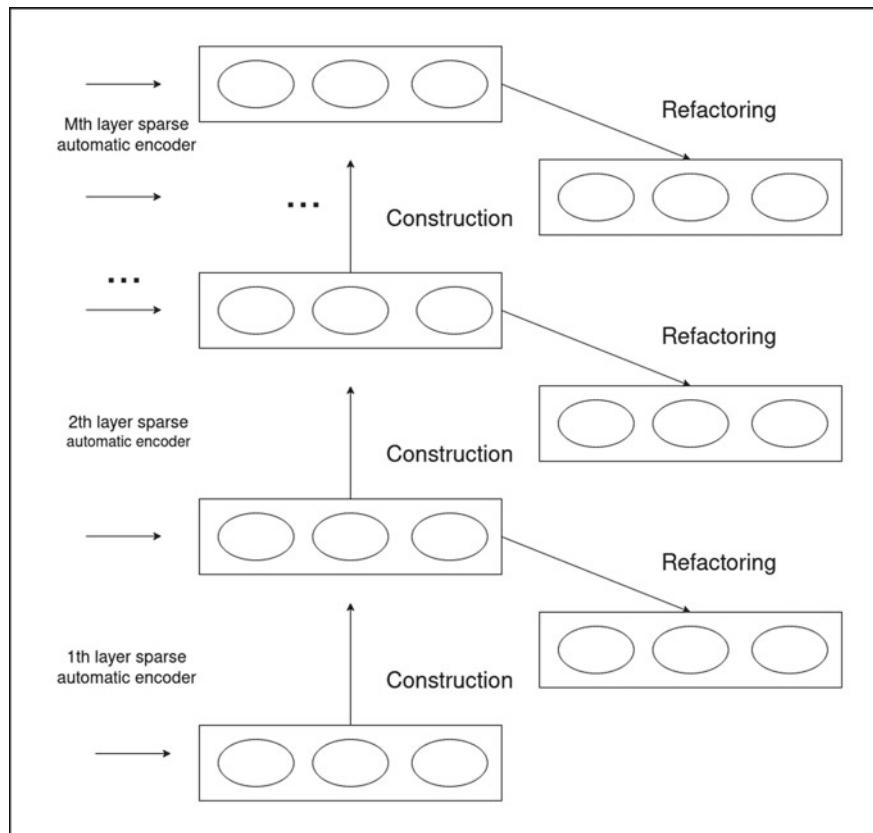


Fig. 5 SSAE diagram

The SSAE model will help to detect CVD. In conventional deep learning, if the number of features is increased, the training accuracy will be higher. But, the training speed will be slower. The intention of using SSAE is to increase training speed and extract the features to optimize the model. The algorithm can adequately extract the sparse informative factor of high dimensional image features. This idea will help us to detect Myocardial infarction efficiently in terms of computational complexity and accuracy. Later, we can implement this idea in a real-time scenario where CDVs can be detected from ECG reports directly.

5 Conclusion

This article focuses on reviewing the existing myocardial infarction detection using machine learning models. First, we introduced the background of CVDs and Myocardial infarction. Then we discussed the existing works for the detection of MI. Our goal is to drive research and development of an optimized and less computational complexity based deep learning-based model to predict and detect the disease with higher accuracy. With further improvement, we can design a model for real-time disease detection with lesser complexity.

References

1. Global Action Plan for the prevention and control of noncommunicable diseases 2013–2020. World Health Organization, Gendva, Switzerland (2013)
2. Laslett, L.J., Alagona, P., Clark, B.A., Drozda, J.P., Saldivar, F., Wilson, S.R., Hart, M.: The worldwide environment of cardiovascular disease: prevalence, diagnosis, therapy, and policy issues: a report from the American College of Cardiology. *J. Am. Coll. Cardiol.* **60**(25 Supplement), S1–S49 (2012)
3. Siddiqui, S.Y., Athar, A., Khan, M.A., Abbas, S., Saeed, Y., Khan, M.F., Hussain, M.: Modelling, simulation and optimization of diagnosis cardiovascular disease using computational intelligence approaches. *J. Med. Imag. Health Inf.* **10**(5), 1005–1022 (2020)
4. Muniasamy, A., Muniasamy, V., Bhatnagar, R.: Predictive analytics for cardiovascular disease diagnosis using machine learning techniques. In: International Conference on Advanced Machine Learning Technologies and Applications, pp. 493–502. Springer, Singapore (2020)
5. Wang, J., Liu, C., Li, L., Li, W., Yao, L., Li, H., Zhang, H.: A stacking-based model for non-invasive detection of coronary heart disease. *IEEE Access* **8**, 37124–37133 (2020)
6. Sörnmo, L., Laguna, P.: *Electrocardiogram (ECG) signal processing*. Wiley encyclopedia of biomedical engineering (2006)
7. Sadhukhan, D., Pal, S., Mitra, M.: Automated identification of myocardial infarction using harmonic phase distribution pattern of ECG data. *IEEE Trans Instrum. Measure.* **67**(10), 2303–2313 (2018)
8. Biel, L., Pettersson, O., Philipson, L., Wide, P.: ECG analysis: a new approach in human identification. *IEEE Trans Instrum. Measure.* **50**(3), 808–812 (2001)
9. Kalyakulina, A.I., Yusipov, I.I., Moskalenko, V.A., Nikolskiy, A.V., Kosonogov, K.A., Osipov, G.V., Ivanchenko, M.V.: LUDB: A new open-access validation tool for electrocardiogram delineation algorithms. *IEEE Access* **8**, 186181–186190 (2020)

10. Blanco, A.L.A., Grautoff, S., Hermann, T.: ECG sonification to support the diagnosis and monitoring of myocardial infarction. *J. Multimodal User Interf.*, 1–12
11. Demir, V., Turan, Y., Ede, H., Hidayet, S., Erkoç, M.F.: Electrocardiographic changes in right ventricular metastatic cardiac tumor mimicking acute ST elevation myocardial infarction: a case of misdiagnosis. *Turkish J. Emerg. Med.* **19**(1), 33–35 (2019)
12. Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101**(23), e215–e220
13. Kim, Y.G., Shin, D., Park, M.Y., Lee, S., Jeon, M.S., Yoon, D., Park, R.W.: ECG-ViEW II, a freely accessible electrocardiogram database. *PLoS ONE* **12**(4), (2017)
14. Dohare, A.K., Kumar, V., Kumar, R.: Detection of myocardial infarction in 12 lead ECG using support vector machine. *Appl. Soft Comput.* **64**, 138–147 (2018)
15. Wang, Z., Qian, L., Han, C., Shi, L.: Application of multi-feature fusion and random forests to the automated detection of myocardial infarction. *Cognit. Syst. Res.* **59**, 15–26 (2020)
16. Pan, J., Tompkins, W.J.: A real-time QRS detection algorithm. *IEEE Trans. Biomed. Eng.* **3**, 230–236 (1985)
17. Biau, G., Scornet, E.: A random forest guided tour. *Test* **25**(2), 197–227 (2016)
18. Amit, Y., Geman, D.: Shape quantization and recognition with randomized trees. *Neural Comput.* **9**(7), 1545–1588 (1997)
19. Baloglu, U.B., Talo, M., Yildirim, O., San Tan, R., Acharya, U.R.: Classification of myocardial infarction with multi-lead ECG signals and deep CNN. *Patt. Recognit. Lett.* **122**, 23–30 (2019)
20. Kora, P., Annavarapu, A., Borra, S.: ECG based myocardial infarction detection using different classification techniques. In: *Classification in BioApps*, pp. 57–77. Springer, Cham (2018)
21. Zupan, J.: Introduction to artificial neural network (ANN) methods: what they are and how to use them. *Acta Chim. Slov.* **41**, 327–327 (1994)
22. Johnson, J.M., Khoshgoftaar, T.M.: Survey on deep learning with class imbalance. *J. Big Data* **6**(1), 27 (2019)
23. Savalia, S., Acosta, E., Emamian, V.: Classification of cardiovascular disease using feature extraction and artificial neural networks. *J. Biosci. Med.* **5**(11), 64–79 (2017)
24. Mandal, S., Biswas, S., Balas, V.E., Shaw, R.N., Ghosh, A.: Motion prediction for autonomous vehicles from lyft dataset using deep learning. In: *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, Greater Noida, India, pp. 768–773 (2020). <https://doi.org/10.1109/iccca49541.2020.9250790>
25. Al-Zaiti, S., Besomi, L., Bouzid, Z., Faramand, Z., Frisch, S., Martin-Gill, C., Sejdić, E.: Machine learning-based prediction of acute coronary syndrome using only the pre-hospital 12-lead electrocardiogram. *Nat. Commun.* **11**(1), 1–10 (2020)
26. Sharma, M., San Tan, R., Acharya, U.R.: A novel automated diagnostic system for classification of myocardial infarction ECG signals using an optimal biorthogonal filter bank. *Comput. Biol. Med.* **102**, 341–356 (2018)
27. Kumar, M., Shenbagaraman, V.M., Shaw, R.N., Ghosh, A.: Predictive data analysis for energy management of a smart factory leading to sustainability. In: Favorskaya, M., Mekhilef, S., Pandey, R., Singh, N. (eds.) *Innovations in Electrical and Electronic Engineering. Lecture Notes in Electrical Engineering*, vol. 661. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-4692-1_58
28. Kora, P., Krishna, K.S.R.: ECG based heart arrhythmia detection using wavelet coherence and bat algorithm. *Sens. Imag.* **17**(1), 12 (2016)
29. Mandal, S., Balas, V.E., Shaw, R.N., Ghosh, A.: Prediction analysis of idiopathic pulmonary fibrosis progression from OSIC dataset. In: *2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON)*, Greater Noida, India, pp. 861–865 (2020). <https://doi.org/10.1109/gucon48875.2020.9231239>
30. Canto, J.G., Rogers, W.J., Goldberg, R.J., Peterson, E.D., Wenger, N.K., Vaccarino, V., NRMI Investigators: Association of age and sex with myocardial infarction symptom presentation and in-hospital mortality. *Jama* **307**(8), 813–822 (2012)

31. Ćulić, V., Eterović, D., Mirić, D., Silić, N.: Symptom presentation of acute myocardial infarction: influence of sex, age, and risk factors. *Am. Heart J.* **144**(6), 1012–1017 (2002)
32. Shih, J.Y., Chen, Z.C., Chang, H.Y., Liu, Y.W., Ho, C.H., Chang, W.T.: Risks of age and sex on clinical outcomes post myocardial infarction. *IJC Heart Vascul.* **23**, (2019)
33. Liu, J.E., An, F.P.: Image classification algorithm based on deep learning-kernel function. *Sci. Program.* (2020)
34. Sun, W., Shao, S., Zhao, R., Yan, R., Zhang, X., Chen, X.: A sparse auto-encoder-based deep neural network approach for induction motor faults classification. *Measurement* **89**, 171–178 (2016)
35. Han, X., Zhong, Y., Zhao, B., Zhang, L.: Scene classification based on a hierarchical convolutional sparse auto-encoder for high spatial resolution imagery. *Int. J. Remote Sens.* **38**(2), 514–536 (2017)

Construction of Effective Wireless Sensor Network for Smart Communication Using Modified Ant Colony Optimization Technique



Avishek Banerjee, Sudip Kumar De, Koushik Majumder, Victor Das, Debasis Giri, Rabindra Nath Shaw, and Ankush Ghosh

Abstract Wireless Sensor Network (WSN) is generally used for constructing an efficient network with minimum infrastructure. With this note, the WSN can be considered to construct smart communication in the existing city to facilitate the inhabitants. Nowadays in India, so many cities are going to be transformed into smart cities. In those cities, the maximum electronic gadgets will be IoT enabled. These IoT enabled gadgets should be connected through a robust network. The WSN can be an alternative for constructing a robust network between the electronic devices and the local server (Sink Node). The WSN can also sense many external environmental factors to facilitate the user, such as traffic status, rainfall, heat or smoke, vibration, and pollution detection. One of the major important challenges for the construction of an effective Wireless Sensor Network is to use the existing infrastructure of a city. In this paper, the effective Wireless Sensor Network has been constructed using the existing infrastructure and this is the major reason to choose the existing roads to deploy the WSN nodes. One can easily use the existing lamp posts of those roads to deploy the WSN nodes. In this paper, the modified Ant Colony Optimization (ACO) technique has been used to construct efficient WSN. The ACO is the probabilistic technique which is used to solve computational problems of choosing a minimized

A. Banerjee · S. K. De · V. Das

Department of Information Technology, Asansol Engineering College, Asansol, India

K. Majumder (✉)

Department of Computer Science and Engineering, MAKAUT, Kolkata, India

e-mail: koushik@ieee.org

D. Giri

Department of Information Technology, MAKAUT, Kolkata, India

R. N. Shaw

Department of Electrical Electronics and Communication Engineering, Galgotias University, Noida, India

e-mail: r.n.s@ieee.org

A. Ghosh

School of Engineering and Applied Sciences, The Neotia University, Kolkata, West Bengal, India

path in a graph. Therefore, this technique can be useful in the construction of cost-effective WSN. After the construction of the WSN network, the total number of nodes is calculated to cover the area efficiently. In order to find the energy used for the network, the distance between each node as well as the distance between sink node and WSN nodes are determined and, essentially, everyone can easily measure the lifespan of the WSN. The research work has been compared with some existing research works and remarkable improvement in lifetime enhancement (20.77% and 84.1% with respect to two existing literature) has been observed.

Keywords Wireless sensor network (WSN) · Effective construction of WSN in smart city (*ECwSC*) · Ant colony optimization (ACO) · Global pheromone update (GPU) rule · Local pheromone update (LPU) rule

1 Introduction

The wireless sensor node is a tiny device fully equipped with limited resources like processing unit, communicating unit, data storing unit, and power back-up unit. It is mainly used for sensing environmental information and transmitting it to the local server (Sink Node). To collect the necessary data, the wireless sensor nodes sense the external environment, and that data is processed through the built-in small processor. The processed data is transmitted through the network. The limited data storage is used for storing processed data, before and after the transmission. In India, many cities are being transformed into smart cities to make city life more comfortable and smarter. Nowadays in modern smart cities, communication should also be smart. There are many smarter applications of WSN in a smart communication system. These applications include sensing the external environment of the smart city through the development of smart communication. One example is going to be described here. To know about the outside weather, one can easily use a heat or moisture sensor to measure the external temperature, which will help the system as well as the inhabitants to inform about the external environment and suggest accordingly. For example, if it is raining outside, the system can advise the user to carry an umbrella.

Contribution(s):

1. The paper has described the effective construction of Wireless Sensor Network using the Algorithm for Effective Construction of WSN in Smart City (*ECwSC*) and the modified ACO algorithm.
2. The novelty of the modified ACO algorithm is that the attractiveness function value has been determined by comparing the “Local Pheromone Update” (LPU) rule and “Global Pheromone Update” (GPU) rule and the better one has been chosen. On the contrary, in case of the traditional approach, the local search is done through LPU rule and after completion of every ant, the pheromone is modified through the GPU rule.

2 Literature Survey

Many researchers as well as engineers have worked as well as working on the development [1, 2] of Wireless Sensor Networks. The development of WSN has generally led to different applications of WSN in different aspects of day-to-day life. The application of WSN in an urban area has been well described in the article of Rashid and Rahmani [3] in the year 2016. The application of WSN is not limited to urban areas only but it is also being popular in rural areas. In 2015, Ojha et al. [4] have published a paper on the application of WSN in agriculture. The researchers are not going to concentrate only on the current applications of WSN but they are also worried about the future trends of WSN in different arenas of modern life. In 2016, Khan et al. [5] published a book on “current status and future trends of Wireless Sensor Network”, and in this book, the authors have discussed the recent developments and future applications of WSN.

Evolutionary algorithms [6] are such kind of algorithms under computational intelligence where the repetition of different operators upon competent population increases the chance of evolving feasible solutions. The Ant Colony Optimization is such a kind of evolutionary algorithm that is highly effective to establish the connection between Wireless Sensor Nodes and the Sink node. ACO is the probabilistic technique that is used to solve computational problems. The fundamental of ACO is to choose the minimized path through graph theory to get the minimized and connected network system. The WSN can also be well described as a connected network of Sink nodes and WSN nodes. Therefore, the ACO technique can be very useful in the construction of WSN. Very few works have been noted where the ACO algorithm has been used to construct the Wireless Sensor Network. In the year 2016, Banerjee et al. [7] worked on the utility of the ACO algorithm to establish an efficient Wireless Sensor Network. In that work, there were many challenges like finding the position of the Sink node, Deployment policy of Wireless Sensor Nodes, Connection building between the Sink node and different sensor nodes, etc. In this paper, the above-mentioned challenges have been taken care of to construct the effective Wireless Sensor Network using modified Ant Colony Optimization technique. The paper contains five sections namely introduction, literature survey, solution methodology, problem formulation and numerical solutions, and lastly, the conclusion section.

3 Methodology

In this paper, a modified Ant Colony Optimization algorithm has been chosen as the optimization technique to minimize the coverage path to construct the WSN. Ant Colony Optimization technique is inspired by the bio-inspired phenomenon of the natural ants. In this technique, the solution is built through some artificial ants [7]. Those artificial ants are considered as a solution to the optimization problem. The ant solutions are refined by the phenomena, i.e., pheromone trail, which is carried

out by updating two rules, i.e., the LPU rule and the GPU rule. In this paper, a good balance between the LPU rule and the GPU rule has been maintained. In traditional ACO, every visited edge is modified by the LPU rule. After the refinement of all ant solutions using the LPU rule, the pheromone is modified by the GPU rule. But in this way, the solution may converge to local optima rather than global optima. Therefore, in our methodology, we have proposed to compare the mod (absolute magnitude) of the GPU rule and the LPU rule and update every ant solution by the suitable GPU rule or LPU rule. This methodology can solve the problem of convergence to a local minimum. This is measured by the “attractive function” as described below.

The attractiveness function for the GPU rule.

$$\tau_{ij} = \begin{cases} (1 - \rho) \cdot \tau_{ij} + \rho \cdot \Delta \tau_{ij}, & \text{if } (i,j) \in \text{best solution} \\ \tau_{ij} & \text{otherwise} \end{cases} \quad (1)$$

The attractiveness function for the LPU rule

$$\tau_{ij} = \{\tau_{ij} \cdot (1 - \varphi) + \varphi \cdot \tau_0 \quad (2)$$

The steps of the proposed meta-heuristic ACO algorithm [7] are given below (See also Fig. 1).

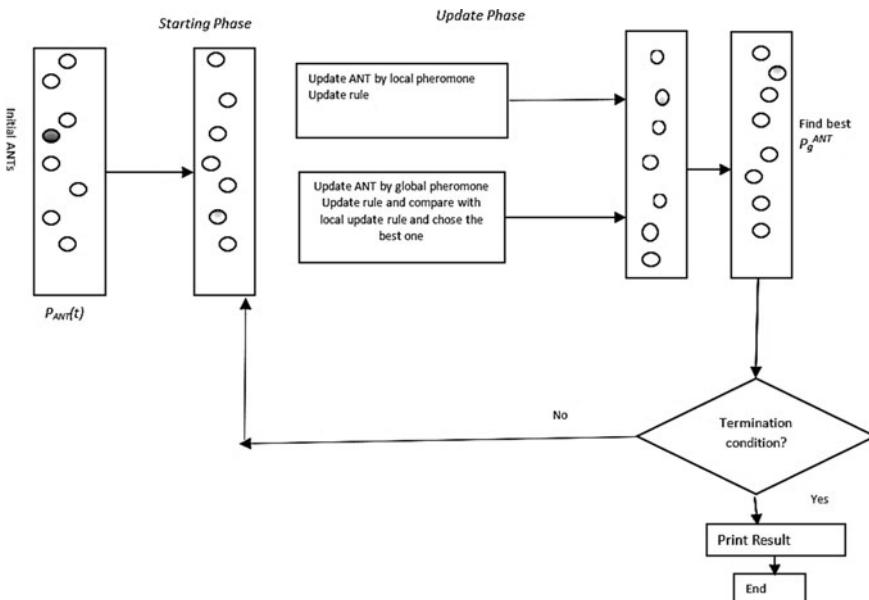


Fig. 1 Block diagram of modified ACO

Algorithm $m\text{-ACO}()$ **Input:** Graph (G) [node: Ant's position, edge: distance], velocity of Ants**Output:** Efficient Network**Steps:**

1. Generate the initial ANT solution, i.e., $P_{\text{ANT}}(\text{time})$.
2. Set the iteration as time = 0.
3. Compute the next ANT solution according to the Attractiveness function (τ_{ij}) as described above. The attractiveness function is dependent upon the GPU rule (Eq. no. 1) and LPU rule (Eq. no. 2).
4. Update the pheromone trail by comparing between the absolute value of attractiveness function, obtained by GPU rule and the LPU rule
5. Find the best ANT solution (P_{ANT}^g) with the best fitness function value.
6. Increment iteration by: time = time + 1.
7. If the termination criterion is not satisfied, go to Step 3, otherwise, go to Step 8.
8. Display the value of the fitness function of the best solution.
9. End.

4 Problem Formulation and Numerical Solutions

The motivation of this paper is to make a smart communication system for the smart city using the already existing resources by constructing an effective Wireless Sensor Network [8–10]. To make a smart city, any city can be chosen. In this paper, a portion of Durgapur B-Zone has been taken for experimental purposes. Here, in this paper, it is called as an experimental zone. For getting the various geographical information of the experimental zone, the API of “<http://overpass-api.de/api>” has been used. The experimental zone is bounded by Latitudes and Longitudes. This boundary is termed as Bounding Box. The bounding box of the experimental zone is defined by quadruplet (south, west, north, east), which is (23.55596, 87.29517, 23.57854, 87.33277). Figure 2 shows the experimental zone. The total area of the experiment zone is 3.62×3.62 sq. km.

After fetching all the street information, the position of WSN nodes for all streets and the position of Sink Node have been calculated and it is depicted in Fig. 3. The symbol * denotes the WSN nodes. In the figure, different colors have been used to denote the WSN nodes, but the colors of WSN nodes do not have any special meaning. The symbol of Tower (Tower icon) denotes the Sink Node. For the establishment of the WSN network, the ACO optimization technique has been used.

The Algorithm for Effective Construction of WSN in Smart City (**ECwSC**) is given below.

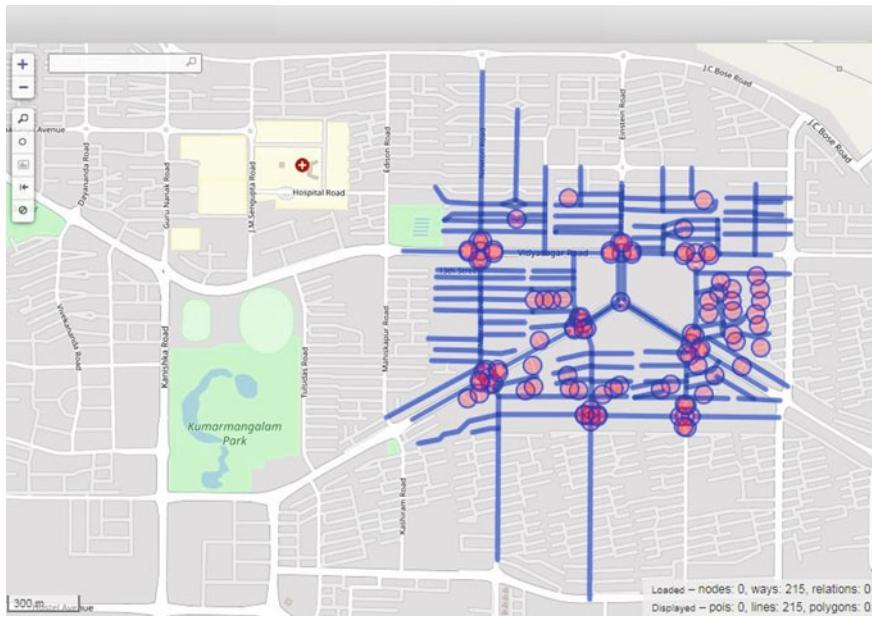


Fig. 2 Experiment zone, depicted by blue lines

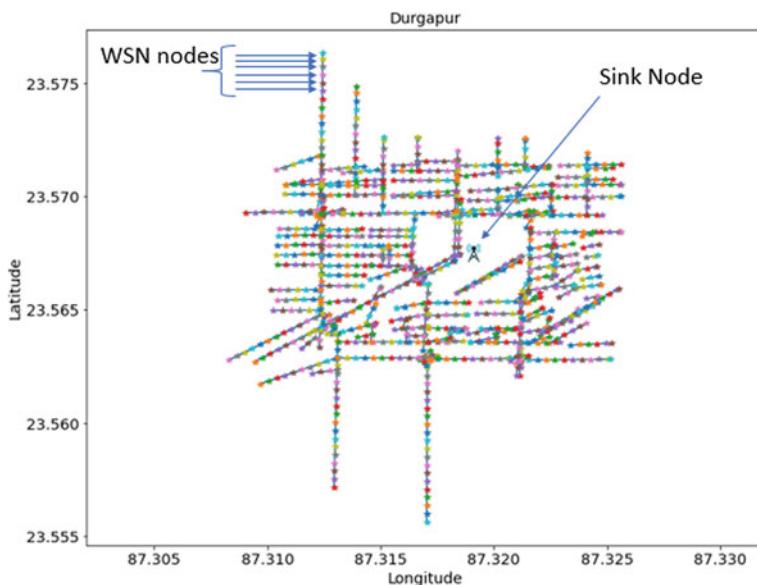


Fig. 3 Street and the positions of all WSN nodes with Sink Node's position

Algorithm ECwSC (Bbox bbox):

Input: Accept the boundary of Experimental Zone in bbox

Output: An Effective WSN network

Steps:

1. Fetch all the geographical information (in terms of *Latitudes and Longitudes*) within the bbox and store it in “*GeoData*” variable.
2. Fetch all the Streets from *GeoData* and assign unique *Street-ID* to every street and store it in *GeoStreetData*.
3. Apply the *K-Mean* algorithm on *GeoData* to choose the position of the *Sink Node (Psn)*
4. Fetch the information from *GeoStreetData* and for each *Street-ID* perform the following operations:
 - 4.1 Find the Locations of WSN nodes on a particular *Street-ID*
 - 4.2 Store all the Locations in the triplet format [*Street-ID*, *WSN-ID*, (*Latitude*, *Longitude*)] in “*GeoWsnData*”
5. Use *GeoWsnData* and *Psn* and apply the modified ACO optimization technique to establish an effective WSN network between WSN nodes and Sink Node.
6. Calculate the total energy required for the whole network.
7. End.

In this paper, the minimization of energy consumption in WSN is formulated through Eqs. 3–8. We have used this concept to minimize the consumed energy for WSN using the modified ACO algorithm.

$$f(k, d) = \text{Minimize}(\text{Energy}_{\text{communication}}(k, d)) \quad (3)$$

Subject to $d > d_o$ for “two-ray ground propagation model”, where d_o is the “threshold transmission distance”.

$$\text{Energy}_{\text{communication}}(k, d) = \text{Energy}_{\text{tx}}(k, d) + \text{Energy}_{\text{rx}}(k) \quad (4)$$

$$\text{Energy}_{\text{tx}}(k, d) = k * (\text{Energy}_{\text{Electronic-energy}}(k) + \text{Energy}_{\text{Amplifier}}(k, d)) \quad (5)$$

$$\text{Energy}_{\text{rx}}(k) = k * (\text{E}_{\text{Electronic-energy}}(k)) \quad (6)$$

$$\text{Energy}_{\text{tx}}(k, d) = k * \text{Energy}_{\text{Electronic-energy}}(k) + k * \text{Energy}_{\text{Amplifier}}(k, d) \quad (7)$$

$$\text{Energy}_{\text{rx}}(k) = k * \text{E}_{\text{Electronic-energy}}(k) \quad (8)$$

In this research, the distance between two nodes has been considered in between allowable range, i.e., 50–100 m [8] in the case of the suggested propagation model.

$\text{Energy}_{\text{Amplifier}}$ = energy required for the transmitting “data packets” between two head-to-head nodes for the amplification to preserve an acceptable “signal-to-noise ratio (SNR)”.

Using the proposed modified ACO algorithm at first, the WSN has been constructed and then the “energy consumption to stabilize the network for 1 s” has been calculated using Eqs. 3–8.

The obtained result from the proposed algorithms has been compared with some existing literature [2, 8, 11, 12], and it has been found that the proposed algorithm provided better results than that of the existing literature in Table 1. From the above table, it is quite clear that the proposed algorithm is better than the existing literature in many decisive parameters (Coverage area, Energy consumption to stabilize the network, and Total lifetime of the WSN).

5 Conclusions

The novelty of this research work is the modification of the ACO algorithm. In the case of the traditional approach, the local search is done through the LPU rule and after applying the LPU rule for every ant, the pheromone is modified through the GPU rule. This strategy may lead to local minima convergence. In this modified ACO algorithm, there is less chance of local minima convergence because in the proposed ACO algorithm the ANT solution is obtained by comparing both the LPU and GPU rules to choose the better one. The ultimate goal of this paper is to construct an efficient WSN for the Smart Communication System for the smart city projects. The paper has a bigger impact because the result obtained by the modified ACO algorithm is much better than the existing literature. The research work has been compared with the existing research work and the Improvement in lifetime enhancement has been recorded 20.77% and 84.1%, respectively [2, 8].

There are plenty of scopes present for the effective construction of WSN using other biologically inspired optimization algorithms. In the future, we are planning to achieve further improvements in terms of performance with the design and implementation of modified ACO algorithms.

Table 1 Comparison between different parameters for exiting literature [2, 8] and the proposed algorithm

Required parameters	Exiting literature [2]	Exiting literature [8]	Proposed algorithm	Improvement in lifetime
Coverage area	1Sq KM	500 Sq Mt	3.62Sq KM	20.77% improvement in lifetime with respect to [2]
Initial energy	2.17E + 14	1E + 14	6.29E + 14	and 84.1% improvement in lifetime with respect to [8]
Energy consumption to stabilize the network (for 1 s)	1756090368 Pico-joules	3512180736.00Pico-Joule	4214843486.72Pic-o-joules	
No of nodes involved in one duty cycle	49	100	629	
The total lifetime of the WSN (in hours)	34.32499	6.590465	41.45402	
The total lifetime of the WSN (in days)	1.430208	0.274603	1.727251	

References

1. Varshney, S., Kumar, C., Swaroop, A., Khanna, A., Gupta, D., Rodrigues, J.J., Pinheiro, P.R., De Albuquerque, V.H.C.: Energy efficient management of pipelines in buildings using linear wireless sensor networks. *Sensors* **18**(8), 2618 (2018)
2. Banerjee, A., Das, V., Biswas, A., Chattopadhyay, S., Biswas, U.: Development of energy-efficient and optimized coverage area network configuration to achieve reliable WSN network using meta-heuristic approaches. *Int. J. Appl. Metaheuristic Comput. (IJAMC)* **12**(3), Article 1 (2019)
3. Bushra, R., Husain, M.R.: Applications of wireless sensor networks for urban areas: a survey. *J. Netw. Comput. Appl.* 192–219 (2016)
4. Kumar, M., Shenbagaraman, V.M., Shaw, R.N., Ghosh, A.: Predictive data analysis for energy management of a smart factory leading to sustainability. In: Favorskaya, M., Mekhilef, S., Pandey, R., Singh, N. (eds.) *Innovations in Electrical and Electronic Engineering. Lecture Notes in Electrical Engineering*, vol 661. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-4692-1_58
5. Das, I., Shaw, R.N., Das, S.: Location-based and multipath routing performance analysis for energy consumption in wireless sensor networks. In: Favorskaya, M., Mekhilef, S., Pandey, R., Singh, N. (eds.) *Innovations in Electrical and Electronic Engineering. Lecture Notes in Electrical Engineering*, vol 661. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-4692-1_59
6. Das, I., Shaw, R.N., Shaw, Das, S.: Analysis of energy consumption in dynamic mobile Ad Hoc networks. In: *Advances in Intelligent Systems and Computing*, vol. 1049, pp. 235–243. Springer (2020). https://doi.org/10.1007/978-981-15-0132-6_15
7. Ojha, T., Misra, S., Raghuwanshi, N.S.: Wireless sensor networks for agriculture: the state-of-the-art in practice and future challenges. *Comput. Electron. Agricul.* **118**, 66–84 (2015)
8. Khan, S., Pathan, A.K., Alrajeh, N.A.: Wireless sensor networks: current status and future trends. CRC Press (2016)
9. Pétrowski, A., Sana, B.: Evolutionary algorithms. ISTE (2017)
10. Mandal, S., Biswas, S., Balas, V.E., Shaw, R.N., Ghosh, A.: Motion prediction for autonomous vehicles from lyft dataset using deep learning. In: *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, Greater Noida, India, pp. 768–773 (2020). <https://doi.org/10.1109/iccca49541.2020.9250790>
11. Banerjee, A., Chattopadhyay, S., Mukhopadhyay, A.K., Gheorghe, G.: A fuzzy-ACO algorithm to enhance reliability optimization through energy harvesting in WSN. In: *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pp. 584–589. IEEE. 2016/3/3
12. Lande, S.B., Kawale, S.Z.: Energy efficient routing protocol for wireless sensor networks. In: *8th International Conference on Computational Intelligence and Communication Networks (CICN)*, pp. 77–81. IEEE, 2016/12/23

A Framework for Personalizing Atypical Web Search Sessions with Concept-Based User Profiles Using Selective Machine Learning Techniques



Pradeep Bedi, S. B. Goyal, Anand Singh Rajawat, Rabindra Nath Shaw, and Ankush Ghosh

Abstract Human beings have been using all kinds of devices to perform diverse things since their evolution. The human brain's imagination contributed to the creation of numerous devices. This devises human existence simply by empowering individuals to fulfill diverse needs for existence, including transport, industries, houses, and computers. Current custom online search mechanisms do not take into consideration specific sites that are unvisited by the customer and could be direct responses to the information needs of the consumer. Furthermore, pages in the outcome package, while not explicitly applicable to the need for user knowledge, may include a connection to relevant pages. Only through doing semantic analysis will certain connections be established. This paper aims to classify certain related sites by semantic review of the quest route and offers an efficient customized site search. This facilitates online search by offering content and the relationship between the search question and the related web pages dependent on individuals.

Keywords Machine learning · Concept-based user profiles · Personalizing atypical web search sessions etc.

P. Bedi (✉)

Department of Computer Science Engineering, Lingaya's Vidyapeeth, Faridabad, Haryana, India

S. B. Goyal

City University, Petaling Jaya, Malaysia

A. S. Rajawat

Department of Computer Science Engineering, Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore, India

R. N. Shaw

Department of Electrical Electronics and Communication Engineering, Galgotias University, Noida, India

e-mail: r.n.s@ieee.org

A. Ghosh

School of Engineering and Applied Sciences, The Neotia University, Kolkata, West Bengal, India

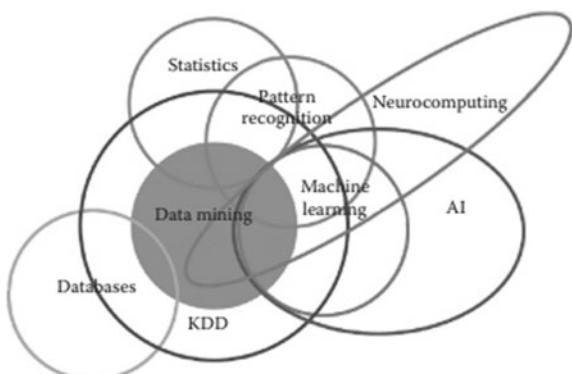
1 Introduction

Machine learning is a subset of artificial intelligence that seeks to enable computers to use smart software to fulfill their tasks effectively. The mathematical methods of learning form the foundation of smart applications used to build machine intelligence. Because machine learning calculations expect knowledge to be learned, the control must be correlated with the database order [1]. Similarly, concepts like Information Exploration from Data (KDD), data mining, and pattern recognition are familiar. One asks how to perceive the big picture in which it shows such a relation. SAS Institute Inc., North Carolina, is the maker of the popular Statistical Analysis System (SAS) analytical tools. We will use the example from SAS to demonstrate the relation of the discipline of machine learning with various relevant disciplines. In a data mining course given by SAS in 1998, this example was actually used (See Fig. 1).

1.1 Web Personalized Search Personalization

Web Personalized Search Personalization is the way of delivering services to customers tailoring to their needs. Web Personalization has been an extremely dynamic research field over the most recent quite a long while and client profile development is a vital part of any personalization framework. So as to find out about a client, frameworks should gather data about them, break down the data, and store the after-effects of the investigation in a client profile. Data can be gathered from clients in two different ways: The first is requesting criticism, for example, inclinations or appraisals; and the second is watching client practices, for example, the time spent to scrutinize an online report. The development of client profiles has a few downsides. The client may give conflicting or mistaken data, the profile fabricated is static though the client's advantages may change after some time, and the development of the profile puts weight on the client that they may not wish to acknowledge.

Fig. 1 Different disciplines of knowledge and the discipline of machine learning [1]



Along these lines, many research endeavors are in progress to make precise client profiles. Client browsing patterns are the most much of the time utilized wellspring of data about client interests. The client profiles are made by ordering the gathered Web pages concerning reference metaphysics [2]. Client profiles are from a similar source, and anyway they use grouping to make a client intrigue chain of command. The gathered Web pages are then allocated to the fitting group. The way that a client has visited a page shows that client's interest in that page's content. Broadening this thought, Chan et al. depict a measurement to appraise the dimension of User's Interest; for instance, the level of connections visited on a page or URL exhibited in bookmarks. The client's browsing history and communications on a person's framework could be the information hotspot for client profiles [3]. This focus on repetitive terms restricts the dimensionality of the archive collection, which further gives an unmistakable comprehension of individuals. This module permits the web crawler for better comprehension of a client's session and possibly tailors that client's requirement understanding as per their necessities. When a bunch of queries is documented, that the web indexes could have a decent representation of the pursuit content behind the recent question utilizing inquiries and instances in the relating question gathering [4].

- (a) Data gathering: In this stage, information is gathered which can be utilization information, content information, client profile information, and structure information [5].
- (b) Data preprocessing: Data is then preprocessed which is an essential assignment to discover intriguing use designs. The web utilization information is put away as web signs in web servers, intermediary servers, or customer programs. Intermediary server web utilization logs are utilized for the experimentation.
- (c) User profiling: It is the way toward social event client explicit data, either verifiably or expressly. The client profile can incorporate his/her own data, intrigue, and navigational conduct while surfing on the net. For the most part, there are two fundamental sorts of client profiles: intrigue-based client profile and conduct-based client profile. Additionally, the client profile could be either static or dynamic. The static client profile never or once in a while changes, for example, client's close to home data such as name and sex. The information in the unique client profile changes every now and again [6].

2 User Profile-Based Search Web

Crawlers have turned into a basic opening to the huge amount of learning accessible in the net and clients normally take a gander at an essential couple of pages of query items, the positioning will acquaint a major with their sight of the web and their information picked up [7]. Most customary internet searchers have vocabulary issues like polysemy and synonymy, and they produce unessential data to the client. It defeats such issues utilizing personalization of the web seeking process result dependent on the area and client profile. Any personalization framework whose goal

is to alter administrations for a client needs to manufacture a client profile. The client profile has data that can recognize one client from a large number of different clients. Profiles regularly incorporate points of interests yet may likewise incorporate subjects of lack of engagement by considering significant and non-applicable reports. Client profiles are commonly worked by giving proper loads to catchphrases that are regarded to speak to a client's advantages or by utilizing weighted ideas from current metaphysics [8]. Then again, a Weighted Concepts profile contains vocabulary which is sufficiently huge to speak to a client's present and future interests [9]. This is on the grounds that first hubs from a current idea chain of command are distinguished as client's advantage dependent on some criticism which might be express or verifiable. A legitimate subset of pages from those hubs (which have been physically characterized before) is then gotten which results in a vast vocabulary. In any case, the philosophy ought to be manufactured legitimately as it ought to speak to address relations between different ideas [10].

3 Literature Survey

Gauch et al. [10]. This paper is probably about the most mainstream method for gathering data about clients, speaking to, and building client profiles. Specifically, unequivocal data strategies are stood out from verifiably gathered client data utilizing program reserves, intermediary workers, program operators, work area specialists, and search logs. We examine in detail client profiles spoke to as weighted catch-phrases, semantic organizations, and weighted ideas. We survey how every one of these profiles is developed and give instances of tasks that utilize every one of these methods. At long last, a short conversation of the significance of security assurance in profiling is presented [10].

Viloriaa et al. [11]. The proposed information portrayal framework permits us to imitate the area models utilized in a wide scope of hypermedia frameworks. The dynamic age of the free introduction update of the application's status makes the device viable with other help applications for versatile courses. The methodology permits the determination of the introduction separate from the development of substance, preferring the reuse and consistency of the introduction, subsequently diminishing the expense of advancement [11].

Silvaa et al. [12]. The ACVPR calculation and engineering of a metasearch framework based on the IMSS-P for planning an arrangement of reconnaissance technology and reputable insight for the SME sector. The framework has two principle modules. The first assists with controlling the client in the assortment of prerequisites for the inquiry cycle producing more exhaustive pursuit keys than the ones utilized in the web indexes. The following module is responsible for Web Mining, investigating connections from URLs returned by the most used web crawlers' public interfaces. One part of the importance of the proposed model is that it doesn't do a customary query-reaction search; however, that the outcomes acquired are refined consistently and sorted by client necessities. This means that once the investigation cycle is

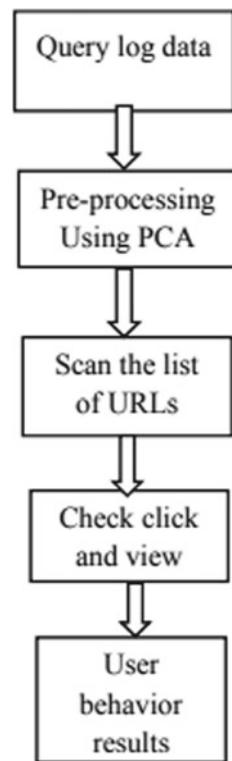
started by the client, a similar cycle will continue and collect new results until the client decides to end it. Despite the fact that the investigation carried out in the work is a qualitative method, due to the way the main encounters gave the framework, the tests show that the framework discovers data according to the type of response expected. In this sense, the three tests had various goals, and the assets acquired by the framework harmonize with said targets [12].

Wang et al. [13], considering this, propose an inert class generative model, named Multi-modular Mention Topic User Behavior-Based Web Search Personalization Framework Using Selective Machine Learning Techniques 35 Model (MMTM), to mimic the producing cycle of clients' referencing exercises by artificially misusing the literary and visual substance. In particular, MMTM embraces the pack of word method for the two messages and pictures to lighten the data sparsity while portraying multi-modular archives. By thinking about the effect of the two modalities on clients' referencing inclinations in a bound together manner, MMTM mutually learns clients' semantic examples and the relationships between substance in various modalities of clients' multi-modular referencing records. At that point, in the wake of learning an information model containing the vital experiences about clients' referencing propensities based on MMTM, they recover the top-k mentionees for a multi-modular referencing post. They led broad trials on a huge dataset slithered from true web-based media administration to assess the presentation of our proposed arrangement. The test results showed the predominance of our methodology over the best-in-class methods [13].

Rimutha et al. [14] have created a customized work suggestion framework based on the client profile. The customized suggestion intends to give results that are probably going to bear some significance with a specific client. Customized proposal is valuable in the space of quest for new employment so as to give people more customized suggestions of occupation postings based on their inclinations. Client profiles are subsequently built based on the individual client's inclinations. Then again, client profiles are useful in improving the proposals. The customized work suggestion framework use ontologies in the area of customized work proposal to show client profiles. The significant goal of this paper is to give an ontology-based client profile for the area of occupation suggestion. Specifically, they distinguished appropriate classes, qualities, and relations that are explicit to the work proposal framework. In introduced OWL representation of the proposed ontological model with the aim that it can be reused by space specialists [14].

Kumar et al. [15] have proposed a Profile-Based Semantic Method utilizing Heuristics for Web Search Personalization. The list items don't meet client inclinations, because of the way that the inquiry is catchphrase-based instead of semantic-based. Abusing client profiles with the use of semantic web technology into personalization may deliver a stage forward in future recovery frameworks. By embracing profiling approach and utilizing ontology-based qualities, a semantic-based method utilizing heuristics and KNN calculation is proposed. It connects with looking through User Behavior-Based Web Search Personalization Framework Using Selective Machine Learning Techniques [16] 34D ontology base spaces on a level plane and vertically to find and concentrate the nearest idea to the significance of the query

Fig. 2 Architecture for GWSP framework based on UBA using query log



watchword. The extricated idea is utilized to grow the client query to customize the output and present the tweaked data for individuals [15] (Fig. 2).

4 Proposed Methodology

Generally, there is a contact between a user and GWSP Framework website. At first, there is the user's requirement to have a subject/topic or reason for searching, we said these are search intention. As per the intention of the user, they create the queries with basic quality of data and experiences. Generally, it needs small phrases and one or two words, and presents the words to traditional search engines, such as GYBA. There is a certain group of ranked URLs which is transferred with explanations of these ranked pages and then users submit their queries. The GWSP framework is organized to assign rank to the URLs and return the top N mostly related webpages for a user based on the algorithm it accepts. The user makes a comparison of the descriptions, like abstracts and titles, in order to make a decision about which URL should be clicked. If the data of the webpage is in the state to meet his demands, he can rest this search process.

If the data can't complete their requests, the user can use another link or continue the procedure of varied query search. If the decision is made to end searching, the user can stop and close the browser or can also have the facility to change alternative [17] search engine to continue when a user didn't get the information that was requested. In Fig. 3, we show the interaction between GWSP Framework and the user. In order to analyze user's behavior for searching, we composed the search log with the help of the GWSP Framework throughout a period of 122 days, from August 1, 2018 to November 30, 2018. Here almost 7500 entries of the user's search query consisting of 2350 sessions. The complete untried result analysis was made on the basis of web log data employed in the GWSP Framework Server. Using the PCA algorithm, we feature selection from the query log, and the size of the query log dataset was reduced. Then the length of the query was found with the help of TF-IDF method. The length of the query is found by categorizing the number of terms or words in a query.

Fig. 3 Interaction between GWSP framework and user

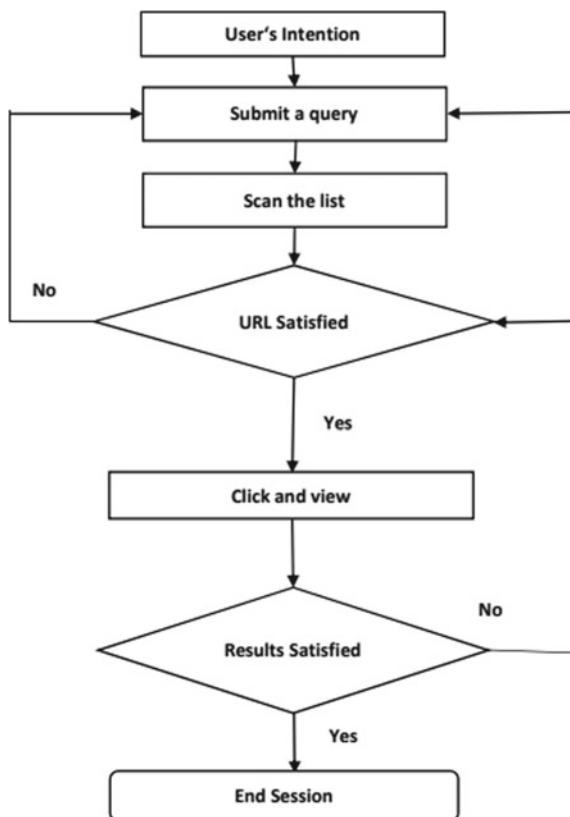


Table 1 QT number distribution of GAWK search engine

LOQ	Ratio (%)
1 term	20.00
2 terms	20.00
3 terms	20.00
More than 3 terms	40.00

5 Result Analysis

5.1 Query Analysis

Queries that are provided to the browsers strengthen as the bond between the web users and search engines increases. The terms which are used in a query determines the distance of the query and also notices the number of times the same query is submitted to the search engine. These measures should be known as they are essential factors of presenting metrics for a search engine to examine and adopt, the inquiry of what search report should be retrieved to the user.

5.2 Length of Queries (LOQ)

The measurement of the query is designed on basis of the number of words or terms that are extended in the query. Here, the word length is described as the connecting parts that are provided in a query. The supply of LOQ is reliable on the TF-IDF method. It also includes the maximum number of terms presented in a query after division. Normally, a single query consists of a term before separation, and if there are more than 3 terms, it means that web users are not familiar with the process of dividing and arranging words by themselves. They ought to give a query along with a sentence or may be a one of a sentence. Inquiry done in this research work discovered that minimum average LOQ is denoted as 1 term, which is reflected in Table 1. This inconsistency proves that the involvement of using GWSP Framework makes the users recognize the significance of using short-term languages substitute of using prolonged and NL sentences.

5.3 Refining of Queries

After reporting the query to the Gawk search engine with the help of the users, they may be required to change some part of the query, only if they didn't get any related results or if the user is not satisfied by the outcomes. We associated with the

Table 2 Refining ratio of query over query types

Query types	Ratio (%)
All	100
Search keyword	40.00
Search keyword and domain	60.00

proportion of refining queries on the basis of the three types of queries and can view the outcomes in Table 2.

From Table 2, we can view that almost 60% to 100% of the queries changed after they were reported. This is a pointer that is provided to users who are normally not satisfied with the usual query function. This too influenced us that rare simple terms are not sufficient to characterize User intention, which in turn restricts the significance of search engines. Hence, the mixture of keywords and field of knowledge provides lower query refining.

5.4 Analysis of Result Clicks

User clicks are the main function after the search engine produces results and it is an essential indicator used for the presentation of search engines. The click behavior maybe varied due to different search purposes. For example, users are required only one URL if they use a search engine as a mean to the process to navigate. Hence, for information searchers, they can click for more options as more than one page in order to collect all-around information.

5.5 Ratio of Result Click

After results are produced by a search engine, then users will click to the subjects they think and wait for the relevant result. There are situations like the users click nothing. In Table 3, we provide the statistics.

We can view that not all users select the URLs after they receive the results. Only 56–100% of them could be judged as clicked after the results were reverted. Also, users looking for domain-based information consist of a lesser ratio of clicks. There remain many probable reasons why people disregard the results. For example, they

Table 3 Ratio of result click

Type of query	The ratio of query refining (%)
All	65.34
Search keyword	45.45
Search keyword and domain	54.55

Fig. 4 Comparison of the position of the click with visitors



cannot receive the proper results or they approach to advertisements or there presents some interesting topics on the page.

5.6 Position of Click

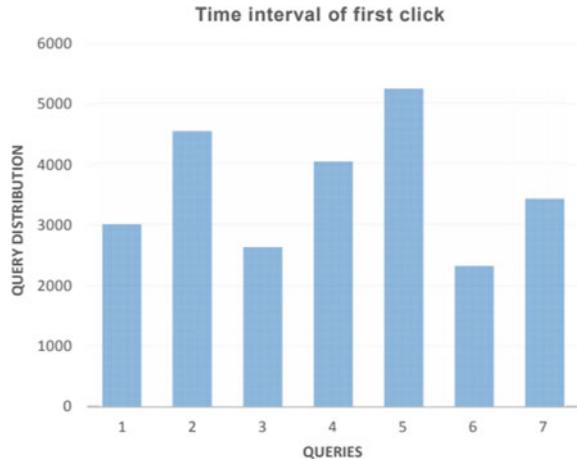
Users can briefly examine the images of returned results which include titles, abstracts, and also URLs [17, 18] before they choose which part to click. There are positions in the user's clock that form a significant indicator of the fulfillment of the search engine, or if they have to click on "next page" or always get the related results at the bottom of the pages, concludes that the results are not good enough. In Fig. 4, we demonstrated the supplies of positions of the clicks when the users report their queries. We adopt this power-law distribution that was created to handle the user's behaviors from search engines as they do not want to consider or to check each result in search engines which are returned. As viewed in Fig. 4, the distribution is reliable on five positions.

This singularity proves the highest five positions of user clicks and merely the satisfaction of the users. Hence, 90% of the users were admired with the Gawk results, but very rare users change on to the next search results. This proves that certain User's behavior can be determined: rare users usually have the patience to search for each URL one by one in order to get relevant resources. They depend on the search engine to offer them the most related ones, and they adapted to believe in the information they are searching for will seemed to be in the first five results.

5.7 Time Interval

The access duration between a user click. The access duration between a user click and the submission of the request mimics the performance provided in the GWSP

Fig. 5 Time interval—the submitting query among user click



framework in generating the corresponding results. If the interlude exceeds a certain edge, it states that the explanations of the results are indirectly prescribed by the users whereas users required time to think whatever to click. Figure 5 shows the supply of the time period between the five groups of users.

It is evident from the graph that most of the time intervals are short and the maximum of the users will click a URL lesser than 2 s. But it is specified that intervals of one user's query exceed 4 s. It is due to the non-relevant results or change in user's interest. Hence, most of the users were satisfied with the results received through the Gawk search engine.

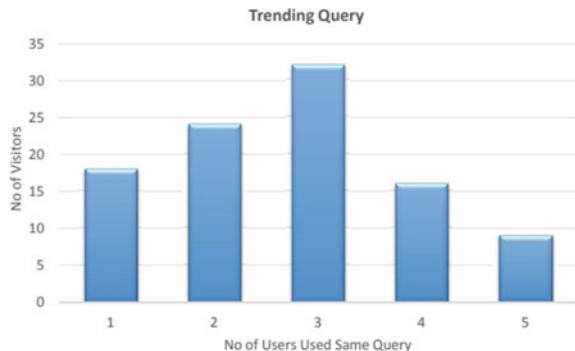
5.8 Distribution Between Trending Queries and the Users

Users would select incessantly by clicking or take alternative actions after the results are reverted. Instead of typing various mixtures of the query, the trending queries admired by the users were used by the GWSP framework to create satisfying results. We demonstrate the supply of the trending query admired by the users in Fig. 6.

6 Conclusion

The techniques for personalization of the web search process are discussed in this research work. The major issues in the machine learning techniques lie in Binary Classification for the existing methods with learning the user's interest, implicit, and explicit feedback. The recommendation system using the collaborative filtering

Fig. 6 Click number distribution of the trending query



method mainly suffers when the information is given as explicit feedback. The foremost challenge with the Content-based recommendation system is to investigate and scrutinize the web page content and match the similarities between them with the given user query. The problem in rule-based recommendation system exists within the exact rules construction as the web user's profile is explicitly built with the user's involvement. The personalized system with a hyperlink may not give clear search results to satisfy the user's need. In link-based personalization system, the web user's involvement is compulsory in rating the query items whereas, in content-based personalization system, the user can explicitly provide his or her personal details at the time of registration itself. In this research work, a new methodology for Web Personalized search engine, named, GWSP framework, was proposed for search results generation based on the User's Behavior. This GWSP Framework model creates the user's profile by using the profile creator technique and modifies the user's profile by profile updater technique dynamically to maintain up-to-date information with reference to the user's interest in searching the Web. The proposed personalized search engine is an innovative approach for personalizing search results. The results generated by the GWSP framework are retrieved based on the phrase-based search and it is the personalized search based on the user preference and the domain of the subject. It also overcomes both polysemy and synonymy problems by using WordNet. The results show that WordNet has helped in improving the effectiveness of queries. Mining the content, user profile, domain, and user preferences is done to personalize search results for a user. The proposed GWSP model overcomes the existing traditional search engine based on the performance metrics mentioned in the experimental results done in this research work.

References

1. Mohammed, M., Khan, M.B., Bashier, E.B.M.: Machine Learning Algorithms and Applications (2016). <https://doi.org/10.1201/9781315371658>
2. Dwivedi, S.: User's activity analysis from weblog: web log expert. Int. J. Eng. Sci. Res. Technol. ISSN: 2277-9655, p. 10. <https://doi.org/10.5281/zenodo.1130885>

3. Marchionini, G.: From information retrieval to information interaction. In: ECIR 2004: Advances in Information Retrieval, Lecture Notes in Computer Science, Springer, pp. 1–11 (2004)
4. Arampatzis, A.T., Tsoris, T., Koster, C.H.A., Van Der Weide, Th.P.: Phase based information retrieval. *Inf. Process. Manage.* **34**(6), 693–707 (1998)
5. Mandal, S., Balas, V.E., Shaw, R.N., Ghosh, A.: Prediction analysis of idiopathic pulmonary fibrosis progression from OSIC dataset. In: 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, India, pp. 861–865 (2020). <https://doi.org/10.1109/gucon48875.2020.9231239>
6. Masłowska, I.: Phrase-based hierarchical clustering of web search results. In: European Conference on Information Retrieval, ECIR 2003: Advances in Information Retrieval, pp. 555–562 (2003)
7. Laura, L., Me, G.: Searching the web for illegal content: the anatomy of a semantic search engine, methodologies and application. *Soft Comput.* 1245–1252 (2017)
8. Adriani, M., van Rijsbergen, C.J.: Term similarity-based query expansion for cross-language information retrieval. In: International Conference on Theory and Practice of Digital Libraries ECDL 1999: Research and Advanced Technology for Digital Libraries, pp. 311–322 (1999)
9. Mandal, S., Biswas, S., Balas, V.E., Shaw, R.N., Ghosh, A.: Motion prediction for autonomous vehicles from lyft dataset using deep learning. In: 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, pp. 768–773 (2020). <https://doi.org/10.1109/iccca49541.2020.9250790>
10. Gauch, S., Speretta, M., Chandramouli, A., Micarelli, A.: User profiles for personalized information access. *Comput. Sci. Inf. Telecommun. Technol.* Center 2335 Irving Hill Road, Lawrence Kansas 66045-7612 (2019)
11. Viloriaa, A., Bonerge, O., Lezama, P.: An intelligent approach for the design and development of a personalized system of knowledge representation. International Workshop on Web Search and Data Mining (WSDM), April 29–May 2, Leuven, Belgium, Proc. *Comput. Sci.* **151**, 1225–1230 (2019)
12. Shaw, R.N., Walde, P., Ghosh, A.: IOT based MPPT for performance improvement of solar PV arrays operating under partial shade dispersion. In: 2020 IEEE 9th Power India International Conference (PIICON), SONEPAT, India, pp. 1–4 (2020). <https://doi.org/10.1109/PIICON49524.2020.9112952>
13. Wang, K., Meng, W., Li, S., Yang, S.: Multi-modal mention topic model for mentionee recommendation. *Neurocomputing* 190–199 (2019). <https://doi.org/10.1016/j.neucom.2018.10.024>
14. Rimitha, S.R., Abburu, V., Chandrasekaran, K.: Ontologies to model user profiles in personalized job recommendation. *IEEE Distrib. Comput. VLSI, Electrical Circuits and Robotics (DISCOVER)* (2018). [10.1109/discover.2018.8674084](https://doi.org/10.1109/discover.2018.8674084)
15. Kumar, M., Shenbagaraman, V.M., Shaw, R.N., Ghosh, A.: Predictive data analysis for energy management of a smart factory leading to sustainability. In: Favorskaya, M., Mekhilef, S., Pandey, R., Singh, N. (eds.) Innovations in Electrical and Electronic Engineering. Lecture Notes in Electrical Engineering, vol. 661. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-4692-1_58
16. Singh Rajawat, A., Jain, S.: Fusion deep learning based on back propagation neural network for personalization. In: 2nd International Conference on Data, Engineering and Applications (IDEA), Bhopal, India, pp. 1–7 (2020). <https://doi.org/10.1109/idea49133.2020.9170693>
17. Rajawat, A.S., Mohammed, O., Bedi, P.: FDLM: fusion deep learning model for classifying obstructive sleep apnea and type 2 diabetes. In: 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, pp. 835–839 (2020). <https://doi.org/10.1109/i-smac49090.2020.9243553>
18. Rajawat, A.S., Upadhyay, A.R.: Web personalization model using modified s3vm algorithm for developing recommendation process. In: 2nd International Conference on Data, Engineering and Applications (IDEA), Bhopal, India, pp. 1–6 (2020). <https://doi.org/10.1109/idea49133.2020.9170701>

Smart Luminaires for Commercial Building by Application of Daylight Harvesting Systems



S. B. Goyal, Pradeep Bedi, Anand Singh Rajawat, Rabindra Nath Shaw, and Ankush Ghosh

Abstract Nowadays, in commercial buildings, the major source of light is artificial lights (LED) which consume a major portion of conventional electricity. Commercial buildings such as schools, colleges, hotels, and malls are designed with quite old lighting systems with fluorescent lamps. Due to the crisis of non-renewable resources of power, the adoption of a smart renewable grid to supply electricity is increasing day by day. This paper is mainly focused to upgrade the existing commercial building infrastructure with application daylight harvesting systems with the integration of Artificial Intelligence. This system is composed of intelligent light collectors, optic fiber guided medium, intelligent luminaires, and control systems. Different controllers are integrated into the framework that is co-ordinated through IoT. The hybrid system will show energy conservation for both old and new commercial buildings.

Keywords Solar energy · Daylight harvesting · Optic fiber · Energy savings · LED luminaires · Artificial intelligence · IoT

S. B. Goyal (✉)
City University, Petaling Jaya, Malaysia

P. Bedi
Department of Computer Science Engineering, Lingaya's Vidyapeeth, Faridabad, Haryana, India

A. S. Rajawat
Department of Computer Science Engineering, Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore, India

R. N. Shaw
Department of Electrical, Electronics and Communication Engineering, Galgotias University, Noida, India
e-mail: r.n.s@ieee.org

A. Ghosh
School of Engineering and Applied Sciences, The Neotia University, Kolkata, India

1 Introduction

The construction sector, especially building construction, consumes about one-third of the total energy generated worldwide. This energy consumption depends on the design and architecture of the building [1]. Different measures are adopted to design and promote better energy utilization in residential or commercial buildings. The improper design and architecture of the buildings had led to an increased usage of electricity from conventional power generation systems. Apart from the conventional system, another alternative to reduce the consumption of electricity is renewable and eco-friendly solar energy. As sunlight is a more safe, comfortable, pollution-free, renewable source of light and energy. So, solar energy is being used in buildings to reduce the power consumption requirement for lightning. Such building is considered to be zero energy buildings (ZEB) or net-zero energy buildings (NZEB). The performance of such buildings is showed a reduction in CO₂ emissions [2, 3]. The ZEB concept had reduced the energy requirement inside the buildings and reduced the current energy challenges worldwide such as rising prices and climate change. The traditional or conventional sources of lightning in buildings are windows and doors but they are no longer useful in the current scenario. Wang [3] surveyed the school buildings for energy utilization and observed that 93% of energy is fulfilled by conventional lighting system and approx. 7% was fulfilled by the fuel and gas system. Therefore, electricity is considered as the main concern area for any commercial buildings. The technology of luminaries used in commercial buildings for energy conservation affects the overall performance of the system. But adoption of the alternative light source doesn't reduce overall energy requirements unless the design of installation of luminaries is not properly done. These optimized selections and installation of luminaries can result in the establishment of wellness conditions concerning thermal comfort in buildings [4, 5].

In this paper, the daylight harvesting technique is proposed using fiber optic technologies for illumination of buildings. A fiber optic lightning system was developed to harvest sunlight for dim lumenated spaces under different scenarios. Gorthala et al. [6] used fiber optics for designing a lighting system for military shelters. This lightning condition was hybridized with conventional sources. Similarly, Han et al. [7] evaluated and showed the effectiveness of fiber optics for harvesting sunlight. The system was designed such that to control the dimming condition of conventional electric lighting by merging it with sunlight-integrated fiber optics luminaries. Kandilli et al. [8] performed an experimental analysis to evaluate the efficiency of optical fiber luminaries. The author concluded these fiber optic luminaries as the most effective alternative to the off-grid electricity supply. These are integrated with solar energy where photovoltaic (PV) collectors are provided to collect light and illuminate the building which results in reduced consumption of conventional energy sources and increases work efficiency of employees working in such buildings [9]. So, the following problems are faced in existing conventional lighting systems:

- Most of the buildings are designed with an old lightening system. So, the modernization of such buildings results in increased cost.

- Power consumption and wastage are high.
- Dependency on conventional methods.

Therefore, according to the above study, fiber optics are considered to be an interesting area for the development of hybrid lighting systems for commercial buildings. As the daylight is transferred inside the building through fiber optic whereas the PV system stores sunlight that was eventually converted into light under night condition or cloudy weather conditions. But most important concern that arise here is how to automatically control such dim-light condition. In this paper, a model or framework is designed for the control strategy of daylight harvesting for illumination in commercial buildings by application of artificial intelligence (AI) and the Internet of Things (IoT).

1.1 Scope of the Research

In the recent scenario, many research efforts are being applied to minimize energy consumption in the construction sector, especially commercial buildings. However, daylight is sufficient throughout the year which leads to the emergence of scope for energy conservation with compensation with solar energy. The scope of this paper is to design a hybrid energy conservation framework by optimizing the lighting system of any commercial building. The daylight harvesting is done with the usage of AI integrated and IoT-enabled optical fibers and PV collectors.

For this, the following objectives are needed to be achieved:

- To select an optimal location to set up an optical collector.
- Usage of old lighting systems. There is no need to change the structure of the building to make it a smart building.
- Application of AI to the lightning system to control automatically.
- To minimize the power consumption of the entire building.

2 Zero Energy Building (ZEB)

Based on the boundary and the metric, zero energy building can be defined in several ways. Considering the values obtained by the owner of the building and the design team, and the objective of the analysis, various definitions may be appropriate. For instance, the cost is the primary cause of concern for most of the building owners. Primary sources of energy or national energy numbers are something that premiere organizations such as DOE or the Department of Energy are concerned of. For the needs of the energy codes, a designer of the building will primarily be interested in the actual use of energy at the site. Apart from this, reducing emissions is the primary objective of those who work to mitigate fossil fuel burning and power plant pollution. The most general definitions are:

- Net-zero site energy.
- Net-zero energy emissions.
- Net-zero source energy.

Each definition deals with a certain renewable energy source and employs a grid for net use accounting. For structures that are independent of the grid, the above definitions are the most suitable. If there is the availability of a certain resource for the entire life of a building, a supply-side option could be employed for each definition [8]. To achieve off-site ZEBs, renewable energy could be purchased from off-site sources. Emission credits could also be purchased for off-site zero-emission buildings. ‘Off-site ZEBs’ are those where only a part of the renewable energy generation is supplied through off-site sources. To fulfill the research requirements of DOE’s ZEB, the definitions about ZEBs employing supply-side options on the site are as follows:

- Net-zero Site Energy: A site ZEB produces only that amount of energy that gets consumed in a year at the site.
- Net-zero Source Energy: A source ZEB produces only that amount of energy that gets consumed in a year at the source. The primary energy employed for the generation and delivery of energy to the site is source energy. Suitable site-to-source multipliers are effectively utilized to figure out the total source energy related to a building. For this, the multipliers are applied to the imported and the exported energy.
- Net-zero Energy Costs: In cost ZEBs, money paid by the utility to the owner of the building for the export of energy to the grid is equal to or more than the money paid by the owner for the energy services and energy consumption throughout the year.
- Net-zero Energy Emission: The renewable energy free of emissions produced by a net-zero emission building is equal to or more than that utilized from energy sources that produce emissions.

2.1 Net-Zero Energy Building (NZEB)

The design of a building may satisfy several NZEB definitions, but the net-zero position may not be achieved year after year. Depending upon the state of operations within a building and the weather conditions, any NZEB may get close to the NZEB category for a certain year. An NZEB performing very well might change to a near NZEB during years of abnormal weather conditions. The abnormal conditions may be like wind and solar resources that are below the average level, and cooling and heating loads that are above the average level. Utility bills or sub-metering should be effectively utilized to measure, review, and track the NZEB status every year.

3 Smart Buildings: Key Features

To promote renewable energy production, flexibility, and interaction among the users, the Smart Buildings concept was put forward by the Energy Performance Building Directive. Several definitions related to Smart Buildings (SBs) have been proposed by various researchers in the literature across the world. Still, the concept and features of smart buildings aren't completely clear. To bring down the consumption of energy and achieve the Zero Energy Building target, the concept of building energy retrofit has been put forward [10–12]. There is an urgent need to convert the retrofitting strategies that exist at the present to smart retrofitting strategies so that the target of nearly zero energy building is easily achieved and external conditions such as weather and grid are easily handled [13]. Smart buildings constitute the features mentioned below:

- Automation: Automatic functions can easily be performed in smart buildings. Also, automatic devices can be accommodated.
- Multifunctionality: Several functions can be performed simultaneously.
- Adaptability: Smart buildings can easily adapt to the external environment. Also, they can easily understand and fulfill the needs of the users.
- Interactivity: Users can easily interact among themselves within a smart building.
- Efficiency: The energy in smart buildings is provided efficiently with optimum time and costs.

To measure SB's performance, it aids in measuring the adaptability of the operation within the buildings to the requirements of the occupant and the grid [14]. The primary functions related to smart readiness indicator within the buildings are [15]:

- Adaptability to an occupant's needs and enable them to monitor their energy consumption directly.
- Adaptability to the requirements of the grid.
- Aids in automated and controlled operation and maintenance of the building.

4 Daylight Harvesting

For the design of sustainable lightning conditions for buildings, daylight harvesting is considered to be one of the prominent techniques. In such buildings, the dim light is adjusted with natural light such as sunlight coming from windows or skylights that can minimize the energy requirement from artificial lights. The daylight harvesting system is employed with light sensors that sense the intensity of light from natural sources and send data to the control system. The control system in turn adjusts the electric lights according to measurement level. Generally, the daylight harvesting system control is divided into two types, open-loop and closed-loop system [16]. In an open-loop system, the sensors sense the intensity of light and balance the brightness inside the building with artificial lights. They are generally installed outside the

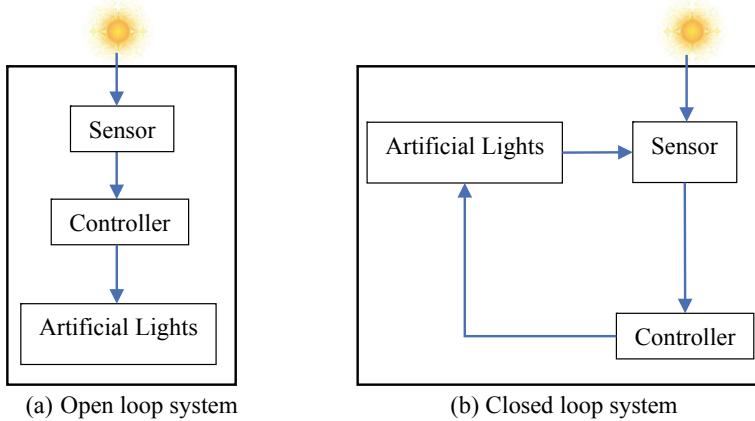


Fig. 1 Daylight harvesting control system

building (Fig. 1a). In such a system, feedback is not required from the artificial system. Whereas in a closed-loop system, the amount of light intensity was measured from both natural and artificial sources. A pre-set ambient light intensity is taken for reference, according to which light intensity level is adjusted (Fig. 1b). These sensors are installed inside the building away from sunlight. The artificial lighting system is measured using sensors and provides feedback to the controller resulting in a closed loop.

As energy consumption is affected by daylighting, thus more artificial light can be saved by the installation of daylight harvesting. Today, the modern building architecture is designed with the potential to calculate lighting energy before the installation of the controller for light. But in some constructions, there are photosensors installed to reduce artificial light consumptions and thus achieving the requirements of an nZEB [17]. Tullio De Rubeis et al. [18] proposed a framework for daylight harvesting which controls the occupancy and achieved energy savage up to 69.6%. Delvaeye et al. [19] studied the energy-saving control systems which saved energy utilization ranging from 18% to 46% by the implementation of daylight harvesting control systems in a school building. Similarly, Yu et al. [20] used RELUX software for calculation of annual saving of energy by the implementation of daylight harvesting and recorded that 40%–46% of energy is conserved [21].

5 Methodology

In modern architecture, buildings are being designed to tackle the energy consumption problem by using the daylight via sensors which are called daylight harvesting that results in a reduction in utility bills. In a commercial building, the main aim is to minimize the entire consumption of power. So, dependency on artificial lights will

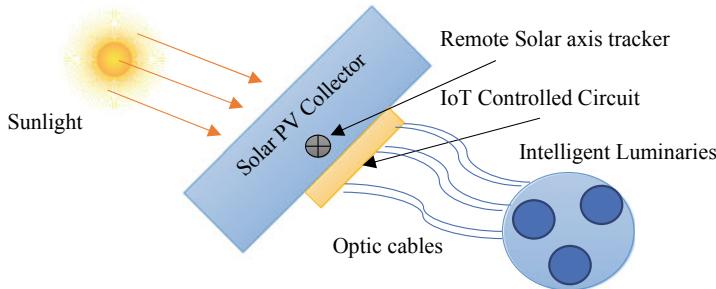


Fig. 2 The intelligent daylight harvesting system

cause an increase in power consumption. In this work, an AI-enabled IoT system is proposed to make commercial buildings zero energy building by minimizing the usage of artificial lights and illuminate the building with sunlight-integrated optic fiber.

5.1 Design of Smart Daylight Harvesting System

The smart daylight harvesting framework consists of an AI-enabled PV solar collector, fiber optic guiding system, luminaries, and IoT-enabled control circuits. All these components are controlled through IoT as illustrated in Fig. 2. This framework shows the fabrication prototype for commercial building lighting applications.

5.1.1 Intelligent Collectors

Solar PV collector is considered to be a light-collecting panel which is the main component of this framework whose function is to collect light during the daytime and store the light energy for the night. Generally, it is made up of metals such as aluminum as they are light in weight as well as show good strength. Along with sheet lens are mounted on the light-collecting panel which acts as solar concentrator which focuses the light to a collecting area where receiver of guiding system, fiber optics receiver, is connected which receives concentrated light and distributes it to cables. To receive maximum sunlight on lenses, throughout the day, the rotating motor is implemented with artificial intelligence that sets rotating angle without any human interactions (Fig. 3). The generated power in PV cells are utilized during the daytime and excess power gets stored in the accumulator which acts as a reservoir to supply energy demand in dim-lighting condition.

Fig. 3 AI-enabled solar PV collector

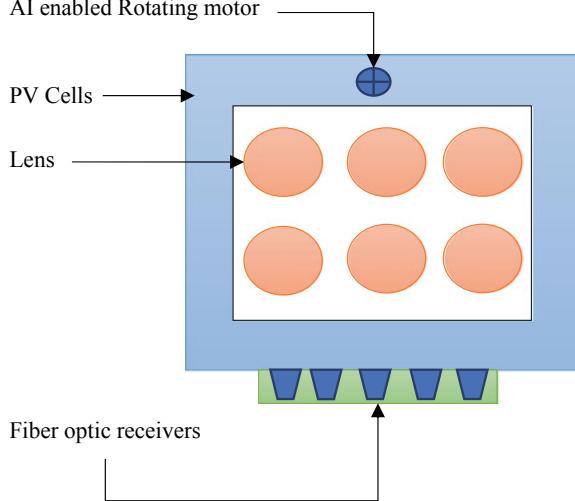
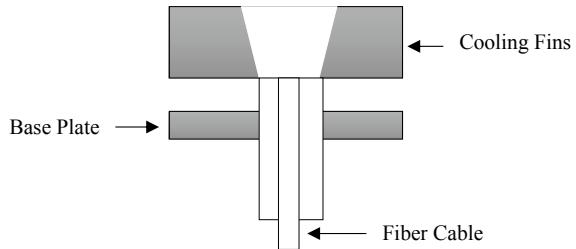


Fig. 4 Optic fiber receiver

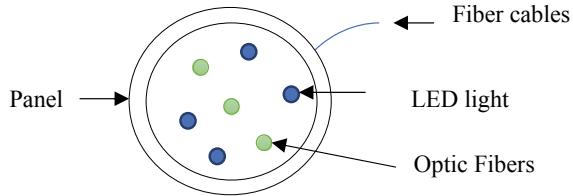


5.1.2 Intelligent Guiding System

The fiber optic cables are used to distribute the collected sunlight to the luminaries throughout the buildings. In this paper, fiber optic cables are used as it possesses high transmittance of light with low dispersion. Before the cables, there is a receiver which minimizes the temperature of the inlet as light is concentrated here which may damage the cables (Fig. 4).

5.1.3 Intelligent Luminaires

The intelligent luminaires are integrated with fiber optics, diffuser lens, panel, and light-emitting diode (LED) light. The diffuser is used to ensure wide-angle emittance uniformly. These panels are used to attach LED lights over them that provide the light to rooms or buildings in the condition of insufficient sunlight (Fig. 5). There is a control system that sets the ambient light intensity needed according to the density of the room. The controller present here will sense the density of the room and send

Fig. 5 Intelligent luminaries

a signal to the light controller that regulates the amount of energy required. If the requirement is fulfilled by sunlight, then artificial light would not be used.

5.1.4 Control System

The AI-enabled IoT integrated control system for the smart building consists of three control units, i.e., solar collector controller, room density controller, and power generation controller. All these data are co-ordinated remotely from the IoT cloud server (Fig. 6). The solar collector controller is implemented over the solar collector which senses the sun orientation and rotates the collector accordingly. The density controller is implemented in luminaries which measures the number of persons entered into the room or hall of the building. The sensors are integrated which evaluate the density of room by evaluating the person count. If someone enters the room, then the count increases and if someone exits the room, then the count decreases. The power generation controller is the master controller which co-ordinates with the other two controllers and accumulator. The stored power in the accumulator is used to control the power of LED lights. Whether, in case of insufficient light, these accumulators will supplement the need.

6 Results and Discussion

In this section, the results of many research works are summarized. After surveying different expertise work in the field of daylight harvesting, many quantitative parameters are identified such as (Table 1):

- Luminous flux.
- Area of luminance or diameter of optic fibers.
- Cost.
- Energy Saving.
- Integration of AI/ML/IoT.

In [22], the author studied the daylighting impact under a government office building. The author observed that internal lightning is not that much sufficient to meet the requirement. So, they performed experimental daylight harvesting on the

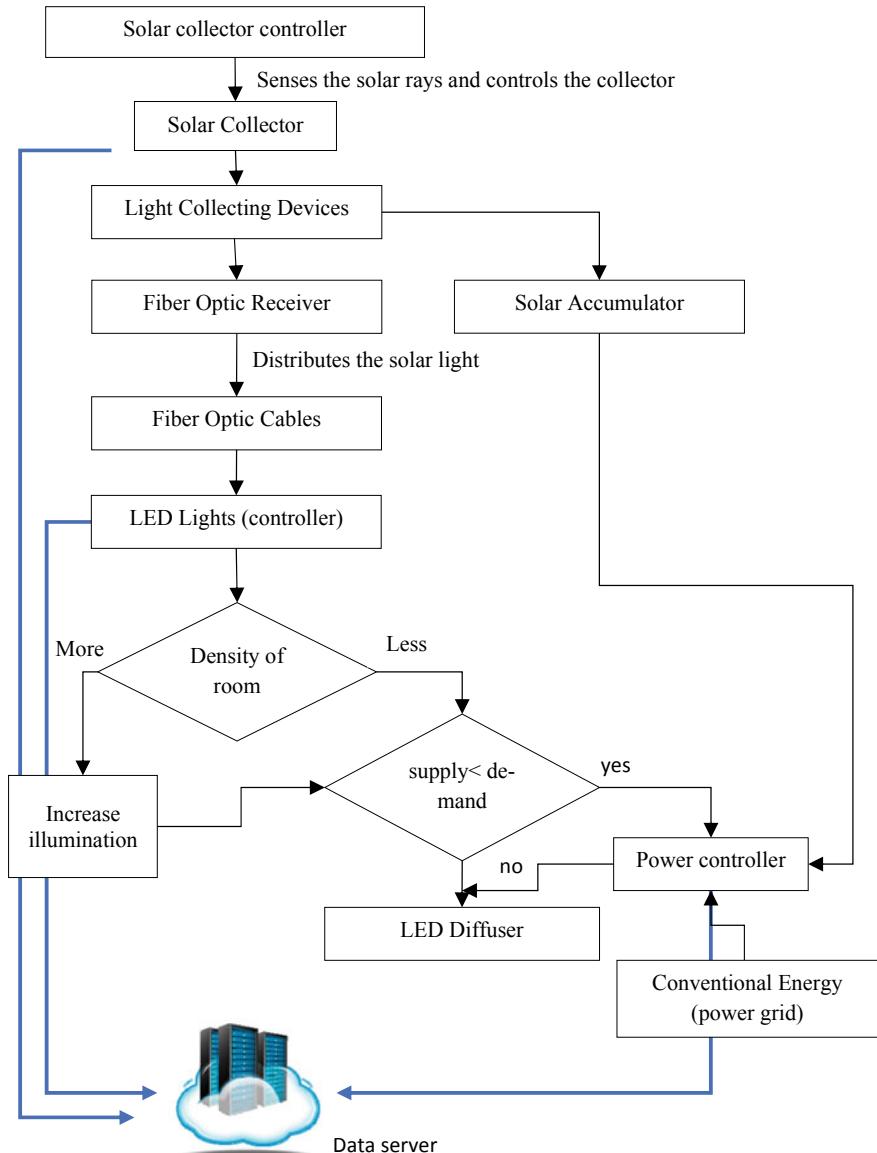


Fig. 6 Design of control system

basis of area and luminance received. Similarly, [23, 24] focuses their investigation on the percentage of the energy savings under condition for visual comfort when artificial lights are integrated with daylight. But, in these research works, daylight harvesting was performed by using traditional techniques. Then in [25, 28], the author explored the application of artificial intelligence (ANN) techniques for reducing

Table 1 Performance comparison of different daylight harvesting methods

Ref.	AI/ML/IoT	Max. flux (lm)	Area/diameter	Transmission range (m)	Cost (lm/\$)	Energy saving (%)
[19]	–	–	–	–	–	18–46
[22]	–	200–4000	1–10	~4	N/A	–
[23]	–	–	~0.01–0.2	~7	–	~15%
[24]	–	–	–	~3	–	51.0–59.3%
[25]	ANN	2000–14000	–	~3.5	–	–
[26]	–	300–1000	0.1–0.5	~6	2–4	–
[27]	–	1–10	1×10^{-6}	~8	0.4–3	–
[28]	ANN	1000–3000	–	–	–	~40–50
[29]	–	4000–30000	0.1–0.5	~8	~40	–

energy consumption. Table 1 illustrates the different contributions of researchers for daylight harvesting.

Table 1 shows that most of the researchers focused their work on the design, location, range, and area of the collector for harvesting daylight. But the change in these parameters will not be cost-effective. So, there is a requirement for the integration of some solutions with existing designs to improve their efficiency. The best solution that comes to mind is artificial intelligence. However, very few contributions are observed here that applied the application of artificial intelligence and machine learning. So, this paper explores an application of AI/ML and IoT to enhance the efficiency of the existing design of daylight harvesting systems.

7 Conclusion

The construction of smart zero energy buildings leads to an environmentally friendly building. As with the increase in the global population, there is an increase in demand for energy requirements. Nowadays, building structures are designed that consume a major portion of energy worldwide. This situation will lead to the depletion of all-natural resources within a few years. So, there is a need to develop technologies that meet the current energy demand and also reduces the wastage of conventional sources of energy. In this paper, an energy-efficient framework is presented that will be beneficial for energy conservation in both new and existing buildings. Using fiber optics and solar energy for a hybrid energy conservation system integrated with AI controllers will lead to a reduction of power consumption inside the building. Along with power consumption, controllers will also prevent downtime of building equipments, improve sustainability as well as reduce the costs, and increase profit.

References

1. Marszal, J., Heiselberg, P., Bourrelle, J.S., et al.: Zero Energy Building—a review of definitions and calculation methodologies. *Energy Build.* **43**, 971–979 (2011)
2. Torecellini, P., Pless, S., Deru, M.: Zero energy buildings: a critical look at the definition. In: ACEEE Summer Study on Energy Efficiency in Buildings. Pacific Grove, CA, USA (2006)
3. Wang, J.C.: A study on the energy performance of school buildings in Taiwan. *Energy Build.* **133**, 810–822 (2016)
4. Kim, S.-K., Le, S.-J.: Zero-energy home development in Korea: energy-efficient and environmentally friendly design features and future directions. *J. Housing Soc.* **42** (2015)
5. Doulos, L.T., Kontadakis, A.: Minimizing energy consumption for artificial lighting in a typical classroom of a Hellenic public school aiming for near Zero Energy Building using LED DC luminaires and daylight harvesting systems. *Energy Build.* **194**, 201–217 (2019)
6. Gorthala, R., Tidd, M., Lawless, S.: Design and development of a faceted secondary concentrator for a fiber-optic hybrid solar lighting system. *Sol. Energy* **157**, 629–40 (2017)
7. Han, H.J., Riffat, S.B., Lim, S.H., Oh, S.J.: Fiber optic solar lighting: functional competitiveness and potential. *Sol. Energy* **94**, 86–101 (2013)
8. Kandilli, C., Turkoglu, A.K., Ulgen, K.: Transmission performance of fiber-optic bundle for solar lighting. *Int. J. Energy Res.* **33**, 194–204 (2009)
9. Barman, M., Mahapatra, S., Palit, D., Chaudhury, M.K.: Performance and impact evaluation of solar home lighting systems on the rural livelihood in Assam, India. *Energy Sustain Dev.* **38**, 10–20 (2017)
10. Torecellini, P., Pless, S.: Zero energy buildings: a critical look at the definition. In: ACEEE Summer Study Energy Eff. Build. 3, 417–428 (2006)
11. Danny, H., Li, W., Yang, L.: Zero energy buildings and sustainable development implications e a review. *Energy* 1–10 (2013)
12. Musall, E., Weiss, T., Lenoir, A., Voss, K., Garde, F., Donn, M.: Net zero energy solar buildings: an overview and analysis on worldwide building projects. In: EuroSun Conference. Graz, Austria (2010)
13. Lê, Q., Nguyen, H.B., Barnett, T.: Smart homes for older people: positive aging in a digital world. *Future Internet* **4**, 607–617 (2012)
14. Rochefort, H.C.: EU smart readiness indicator for buildings (2019)
15. Verbeke, S., Ma, Y., Van Tichelen, P., Bogaert Waide Strategic Efficiency, S., Waide OFFIS, P., Uslar, M., & Schulte, J.: Support for setting up a smart readiness indicator for buildings and related impact assessment Interim report (2017)
16. Yu, X., Su, Y.: Daylight availability assessment and its potential energy saving estimation—a literature review. *Renew. Sustain. Energy Rev.* **52**, 494–503 (2015)
17. Shaw, R.N., Walde, P., Ghosh, A.: IOT based MPPT for performance improvement of solar PV arrays operating under partial shade dispersion. In: 2020 IEEE 9th Power India International Conference (PIICON), SONEPAT, India, pp. 1–4. 10.1109/PIICON49524.2020.9112952 (2020)
18. de Rubeis, T., Nardi, I., Muttillo, M., Ranieri, S., Ambrosini, D.: Room and window geometry influence for daylight harvesting maximization—Effects on energy savings in an academic classroom. *Energy Procedia* (2018)
19. Delvaeye, R., Ryckaert, W., Stroobant, L., Hanselaer, P., Klein, R., Breesch, H.: Analysis of energy savings of three daylight control systems in a school building by means of monitoring. *Energy Build.* **127**, 969–979 (2016)
20. Yu, X., Su, Y., Chen, X.: Application of RELUX simulation to investigate energy saving potential from daylighting in a new educational building in UK. *Energy Build.* **74**, 191–202 (2014)
21. Dascalaki, E.G., Balaras, C.A., Gaglia, A.G., Droutsa, K.G., Kontoyiannidis, S.: Energy performance of buildings—EPBD in Greece. *Energy Policy* **45**, 469–477 (2012)

22. Mandal, S., Balas, V.E., Shaw, R.N., Ghosh, A.: Prediction analysis of idiopathic pulmonary fibrosis progression from OSIC dataset. In: 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, India, pp. 861–865 (2020) <https://doi.org/10.1109/gucon48875.2020.9231239>
23. Fasi, M.A., Budaiwi, I.M.: Energy performance of windows in office buildings considering daylight integration and visual comfort in hot climates. *Energy Build.* **108**, 307–316 (2015)
24. Qiu, C., Yang, H.: Daylighting and overall energy performance of a novel semi-transparent photovoltaic vacuum glazing in different climate zones. *Appl. Energy* **276**, 115414 (2020)
25. Mandal, S., Biswas, S., Balas, V.E., Shaw, R.N., Ghosh, A.: Motion prediction for autonomous vehicles from lyft dataset using deep learning. In: 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, pp. 768–773 (2020) <https://doi.org/10.1109/iccca49541.2020.9250790>
26. Kennedy, D.M., Rourke, F.O.: Experimental analysis of a scaled, multi-aperture, light-pipe, daylighting system. *Sol. Energy* **122**, 181–190 (2015)
27. Vu, N.H., Shin, S.: Cost-effective optical fiber daylighting system using modified compound parabolic concentrators. *Sol. Energy* **136**, 145–152 (2016)
28. Kumar, M., Shenbagaraman, V.M., Shaw, R.N., Ghosh, A.: Predictive data analysis for energy management of a smart factory leading to sustainability. In: Favorskaya, M., Mekhilef, S., Pandey, R., Singh, N. (eds.) *Innovations in Electrical and Electronic Engineering. Lecture Notes in Electrical Engineering*, vol. 661. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-4692-1_58
29. Li, X., Wei, Y., Zhang, J., Jin, P.: Design and analysis of an active daylight harvesting system for building. *Renew. Energy* **139**, 670–678 (2019)

Multi-objective Fuzzy-Swarm Optimizer for Data Partitioning



S. B. Goyal, Pradeep Bedi, Anand Singh Rajawat, Rabindra Nath Shaw, and Ankush Ghosh

Abstract To boost the performance level of big data, data partitioning is considered to be as the backbone of big data applications. In recent years, many researchers are focusing their work toward data science and analysis for real-time applications with the integration of big data. Human interaction with data partitioning of big data is quite time-consuming. So, it is needed to make the data partition elastic as well as scalable while handling a high workload under the distributed system. In this paper, a multi-objective fuzzy-swarm optimization algorithm is proposed for cluster-based data partitioning. This paper also provided an analytical result analysis of different optimization algorithms for data partitioning, i.e., reduction or clustering along with their limitations. This paper provides an approach to enhance the efficiency level for clustering large complex data.

Keywords Big data · Data partition · Clustering · Optimization · Data mining

S. B. Goyal (✉)
City University, Petaling Jaya, Malaysia

P. Bedi
Department of Computer Science Engineering, Lingaya's Vidyapeeth, Faridabad, Haryana, India

A. S. Rajawat
Department of Computer Science Engineering, Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore, India

R. N. Shaw
Department of Electrical Electronics & Communication Engineering, Galgotias University, Noida, India
e-mail: r.n.s@ieee.org

A. Ghosh
School of Engineering and Applied Sciences, The Neotia University, Kolkata, West Bengal, India

1 Introduction

In recent years with the increasing development of technology, there is a drastic increase in application over the internet that ultimately increases the volume of data every day. This leads to an issue of how to use such a large volume of data for decision-making strategies. Data mining approaches use such a large volume of data for association mining, clustering, classification, etc. Data mining includes processes such as data collection, pre-processing, analysis, and result evaluation. Data collection is the process to collect raw data for further process. Data pre-processing include cleaning of data, normalization, feature extraction and reduction, etc. All these tasks require processing over large data to get the desired output. The high-dimensional data leads to large mathematical calculation complexity as well as consumes more time for processing and does not give efficient output. So, there is a requirement for data reduction or data partitioning techniques to reduce such complexities [1, 2]. Nowadays researchers are focusing their interest area to handle such a large volume data, high-velocity data, and large variety of data. These 3 V introduces the concept of big data. So, to handle big data, parallel distributed processing is performed to reduce processing time.

1.1 *Data Reduction*

Data reduction is specified as in three terms:

- Feature Reduction
- Data Compression
- Instance Reduction

There are many existing techniques developed to handle big data by feature reduction techniques. This is performed by the application of feature selection and extraction whereas instance reduction has been handled with instance selection and construction. But data compression is applied to reduce the size of the data that makes the analysis process difficult. For example, a large volume of data means a large number of samples in data whereas high-dimensional data represents a large number of features in data. Both, large volume and large dimension data needed to be reduced for analysis in a big data environment [3].

1.2 *Data Partition*

Parallel processing of distributed data has many policies. Storage of data over distributed nodes results in the partitioning of the entire dataset into subsets. In this paper, an analytical study is presented for data reduction or clustering methods

that can result into effective methods for data partitioning along with their limitations. Storage of data over distributed nodes will increase the processing efficiency of the data mining techniques. Scalability, cost efficiency, and processing speed will be increased. But for replication, the data consistency problem will arise and each node has to communicate and regularly update. So, the data partitioning solution will result as one of the effective methods to handle such limitations and issues [4].

In this paper, different data partition, data reduction, or clustering policies based on swarm optimization are discussed and proposed a multi-objective swarm optimization algorithm for data clustering or partitioning of big data for further processing. The proposed multi-objective method by improving the global searching ability for the best solution will achieve high accuracy.

2 Data Partitioning

The superabundance data, that is being produced from various sources, like from social media, from mobile devices, and from business transaction, etc., can be utilized in the production of high value information which is a must for business intelligence, to perform a deep study on the science of data-intensive, for forecasting and various sources of applications. In order to administrate this huge and complex data, it becomes essential to implement Traditional technologies, like Structured Query Language (SQL)-based Relational DataBase Management Systems (RDBMSs) and data warehousing [6].

For the analysis of big data, the frameworks like scalable distributed computing architectures are demanded because it is impossible to analyze terabyte-scale datasets with the use of a single machine. In this paper, through a survey, a brief summary is demonstrated on the methods of sampling and partitioning, which are mainly responsible for ascending and accelerating the big data analysis algorithms with the analysis of big data available on the Hadoop clusters. “Big data” is the name given to the data when its size itself becomes one of the problems and the techniques of divide-and-conquer are implemented to do the analysis of these big data analysis on computing clusters and for the implementation of this technique or strategy, the MapReduce computing model [7] can be employed in the mainstream big data analysis frameworks [8], for example, Apache Hadoop and Spark. In these types of frameworks, shared-nothing architecture strategy is used because in these frameworks, in the context of both data and resources, each node is free and not dependent on any other frameworks.

The Hadoop Distributed File System (HDFS) [9] on Hadoop clusters works in order to organize and also create a copy of big data file in the form of small distributed data blocks. If we go through the past studies, it has been found that in the cluster parallelization which depends on distributed data blocks may result in a

linear speed-up as the resources which are responsible for computing increase when the data size is big [10]. Actually, when we surplus extra machines in the computing cluster then we can easily increase and enhance and even the growth rate of available resources can exceed quickly. In the process of scaling out a computing cluster, some extra investment is needed and in practice, it may not be always available [11]. So it is the requisition of this approach to minimize the expense of cluster computing and at the same time, it should enhance the workability of big data analysis in computing [12], and for the production approximate results, like acquiring low latency and for the complete utility of resource, only a subset of the input data is employed [13].

2.1 Overview of Data Partitioning

Data partition was used in the central database for the first time ever and its objective was the processing of query in databases systems, and in big data analysis frameworks, the objective is data-intensive computing. The solution or we can say the remedies for distributed database systems was proposed by Das et al. [14] and Baker et al. [15] and for big data applications on NoSQL, again partitioning solution was the solution that was proposed. By Kamal et al. a workload-aware partition was proposed [16, 17]. As shown in Fig. 1, the three main parts of data partitioning methods are: functional partitioning, horizontal, and vertical.

Functional Partitioning: In this type of partitioning, data are associated with each other according to their usage and linkage. For example, for any commercial store, invoice data are stored in one partition and stock summary is stored in another partition and all partitions are associated with each other (Fig. 2).

Horizontal partitioning: In such type of partition, subsets from main data are partitioned row-wise. Each sub-section stores a number of instances.

Vertical partitioning. In such type of partition, subsets from main data are partitioned column-wise. Each sub-section stores samples feature-wise.

2.2 Horizontal Partitioning

The idea of horizontal partitioning emerged from parallel distributed computing systems which was termed as sharing. In this partition of the records, each data partition has the same number of vertical features or columns as the whole dataset has and the dataset is dissected into disjoint subsets. Some noteworthy contributions for horizontal data partition are such as Apache, MongoDB, MySQL Cluster, and Oracle NoSQL database. If we consider the schemes of horizontal partitioning, so this includes various schemes like random schemes, range, and hash schemes. In vertical

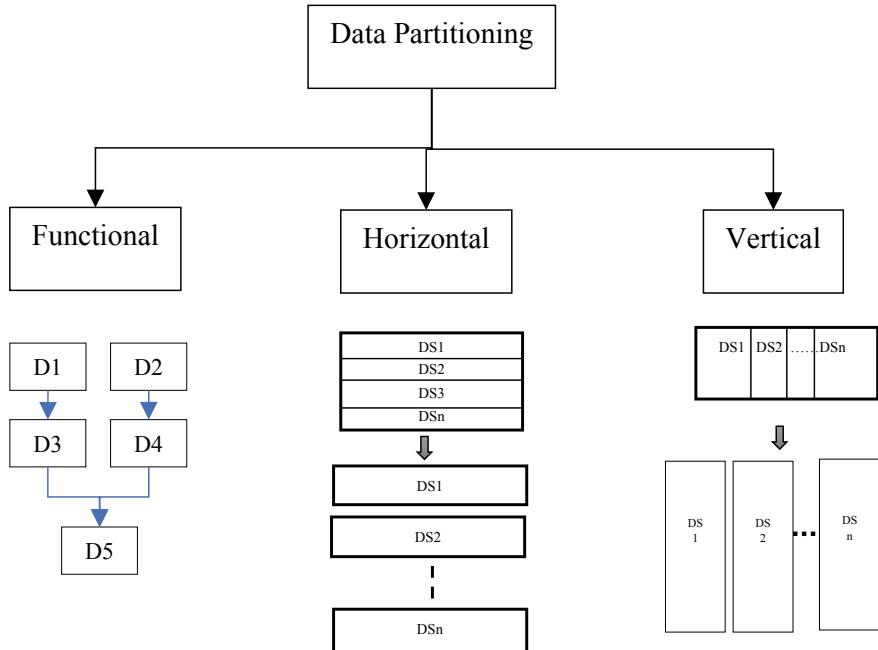


Fig. 1 Data partitioning types

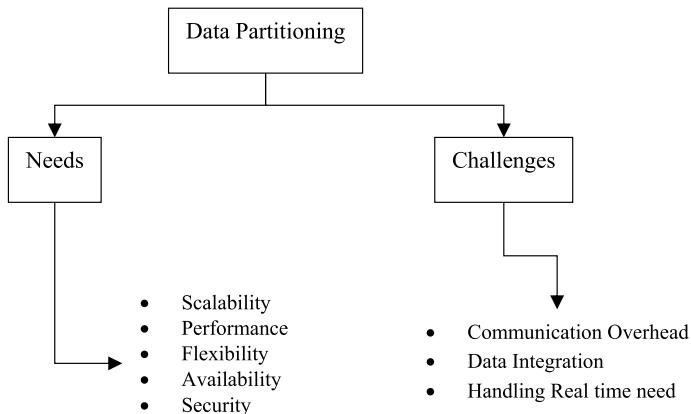


Fig. 2 Data partitioning needs and challenges

partitioning, the column of the dataset is fractioned according to their pattern of use into subsets which give and use a key column.

In this scheme, the column which is frequently accessed are kept in vertical partition data and fields which are least accessed might be placed in some other partition. In two different subcategories, we can divide vertical partitioning methods, i.e.,

restrictive partition and heuristic approach. A Hybrid partition can be defined as the combination of horizontal and vertical partitioning in order to partition or reduce the dataset in accordance with the target application or we can term it as workload. Many a times, it can be called functional partitioning which works for the data access performance and also improve isolation, like from read-only data extracting the read-write data and in the same way, the system of e-commerce divides the data into two separate subparts: one for the storage of invoice data and the other for the storage of product inventory data.

2.2.1 Hash Partitioning

For the hash partition, record key or hash key is required for mapping data samples taken for partition along with their key or hash value to be stored at different locations or nodes. The partition records are divided into subsets and this is done through various mapping methods and among that a round robin is one of them. In this partition, the modulus operation is performed over hash value. The requisition of this partition is that it offers a guarantee on the basis of a key-wise independent, and the main reason for this is the fact that they must have the same hash value if records carry the same key value and for this reason, they would be mapped to the same divisions. In a hash-based partition, there is no assurance of proper sequence among partitions but in the context of the round-robin method, because the size of subsets is equivalent, it offers a balanced partitioning [17].

2.2.2 Range Partitioning

For global order, the best partition method is range partition because it divides data in accordance to a specified range for data partition because it offers a key-wise independent assurance as well as partition-wise sequencing. It works on the pattern and style of hash-based partition. In such type of partition, hash partition is performed within a range or division of data sample points. In the context of range partitioning, a set of key ranges is required that has to be predefined. In the beginning of the partition, there is non-availability of information about data distribution or partition. This results in difficulty to partition data because in a distributed environment, it is really very challenging to choose partition boundaries. So, at this point, range partitioning demands a cost-effective as well as correct and effective pathway to resolve partition limitations and to improve accuracy and make them cost-effective also [18, 19].

2.2.3 Random Partitioning

In the context of random partition, with the use of random number generator for the purpose of determining the place to put the record, the records are parted into subsets. In this random division or partition, approximately equal-sized subsets are produced which is something similar to round robin. In this partition, extra processing is required to find the output of a random value for each record. In Table 1, we can see the similarities of different horizontal partitioning schemes. In the context of output or result of workload in responsiveness, processing, and cost storage, each scheme has its pros and cons. However, different mixed partitioning approaches are also developed that attracted the researcher's interest. A guarantee of balanced partitions is provided by the methods of random and round-robin partitioning.

3 Cluster for Big Data Partitioning

In companies with small computing clusters, it becomes extremely challenging to do big data analysis for the data scientists because the amount of data in different industrial areas goes beyond the petabyte scale. Although on computing clusters, the paradigm of divide-and-win is used in order to do mining algorithms to big data scale and iterative data analysis [1–5], and also the extensible of these algorithms is restricted to the accessible resources. For the curative of these types of problems, approximate computing is used in which, to get an output at lower costs, samples of data are employed. Nevertheless, with the enhancing number of distributed data sampling on computing clusters transform inadequately. With the increasing volume of distributed data, if various samples are needed in statistical analysis and diagnostics, then it is not granted anyway. In big data analysis frameworks, data partitioning plays an essential role. It works in order to achieve scalability to large computing clusters by controlling parallelism. When working with big data, it can prove to be a computationally expensive operation [18, 19].

4 Methodology

Good clusters are formed when the data samples in the same group or population are very much similar as well as different from other populations. Therefore, for cluster analysis, there is a requirement of continuous optimization of the objective function. Single objective optimization function gives efficient result in condition of less complex data but if complete datasets are considered, multi-objective optimization gives efficient result. A multi-objective optimized clustering analysis

satisfies multiple constraints to reach a best fit feasible solution set. Many research works on multi-objective optimization clustering were discussed nowadays such as multi-objective particle swarm optimization and multi-objective genetic algorithm, etc.

4.1 Multi-objective Selection Function

In order to search the global optimal data partition using clustering techniques for any optimization technique, the problems have to be identified and represented. In the proposed method, if the dataset is represented as ‘D’ with ‘N’ dimension having constraint to create maximum of ‘C’ clusters, then each population of the optimization represents a sub-data vector of dimension ‘CxN’ and represented as the example given below:

Suppose the $D = 2$ and $C = 3$ and the population, $D_i = [61, 70, 13, 17, 18, 30]$. Then three clusters will be formed such as (61, 70), (13,17), and (18,30).

But the selection of sub-data is based on some criteria or constraints, and this is termed as an objective function. The performance of the cluster depends on clustering the objective function efficiently. The selection of objective function depends on the application and sometimes different objective functions are contradictory in nature. But sometimes, it results in finding the optimal global solution.

4.2 Multi-objective Optimization Clustering Algorithm

In this work, a multi-objective optimization problem is proposed with an application of fuzzy logic and swarm optimization. As multi-objective algorithm optimizes two or more objective functions, these objectives may be sometimes conflicting in nature and don't reach to the best solution by single objective algorithms, which are often conflicting, in a feasible solution set. In this proposed multi-objective fuzzy-swarm optimization (MOFSO), stopping criteria are chosen as the maximum number of iterations. By varying the value of the global variable in the range of [0,1] makes this algorithm fit from single objective to multi objective. This MOFSO is used to optimize process parameters to maximize objective function and minimize time complexity simultaneously. While, MOFSO is used to optimize process parameters to minimize a single function only. When the iteration stage is finished up, the algorithm yields the best search space with the best fitness.

Algorithm: Multi-objective Fuzzy Swarm Optimization

Start

Define input parameters such as number of populations, maximum iterations, switch probability, limits of lower bound & upper bound, global variables;

Initial position ($x = x_1, x_2, x_3, \dots, x_{\text{dim}}$) of all population are generated;

Stimulus Intensity I (fitness value of multi-objective function) of population are calculated;

Find the best population;

For $g = 1$ to maximum iterations

 For $h = 1$ to population size

 populationstrength is calculated using fuzzy rules;

 If random number < switch probability

 Update the position of population;

 End

 If Updated populationposition < lower bound || updated

populationposition > upper bound

 Updated populationis set equal to corresponding lower or upper bound value;

 End

 populationfitness evaluate at updated populationposition;

 If Updated populationfitness < Best populationfitness

 Best populationfitness and its position get updated;

 End

 Update sensory modality;

 End

End

Output best populationfound

End

5 Results and Discussions

In this section, different contributions of researchers are summarized. After surveying different expertise work in the field of dataset clustering and partitioning, many quantitative parameters are identified such as (Table 1):

- Accuracy
- Error Rate
- Relevence Index
- Processing Time
- Memory Utilization

Table 1 Comparative result analysis

Ref	Technique used	Results	Limitations
[20]	Fuzzy C-means	Accuracy = 96.3%	Increased computational time
[21]	Genetic algorithm	Maximal relevance is approx. 5	Diverse solution on large feature set
[22]	Distributed fuzzy rough set	Accuracy = ~94%	Scalability issue
[23]	Multilayer co-evolutionary	Accuracy = ~94%	With increase in noise level accuracy drops steeply
[24]	PSO-GWO	Accuracy = ~90%	Overfitting problem.
[25]	Genetic algorithm	Accuracy = ~94%	Increased computational time
[26]	Particle swarm optimization	–	Large computational cost
[27]	Canonical PSO	Performs better than PSO	Increased computational time
[28]	MO-PSO	Error rate = 0.5	Computational speed is affected by increasing iteration
	MOGA	Error rate = 0.47	
	MOCSO	Error rate = 0.01	
[29]	Shared nearest neighbor clustering	–	Replication of same blocks
[30]	Inclusive similarity-based clustering	Time = 50–100 s Accuracy = ~90%	Time consumption was high

Table 1 shows that most of the researchers focused their work on designing single objective as well as multi-objective clustering algorithm for data partitioning. But the change in these parameters will not be cost-effective. So, there is a requirement for the integration of some solutions with existing designs to improve their efficiency.

6 Conclusion

With the development of artificial intelligence applications, a large volume and variety of data are sensed and collected from different sources. Such a large volume of data need to be processed and analyzed by data mining tools. However, collected complex data makes the analytical process difficult with a large processing time. It is quite difficult to process such a large volume of data and to extract information out of them. Generally, these large collected data need to be categorized into meaningful clusters. For this, clustering or data partition is required. So, it is required to achieve a better solution on such complex data. As in many fields, clustering algorithms show its effectiveness. In this work, swarm optimization is used to partition data into different clusters for further processing. The proposed multi-objective method by

improving the global searching ability for the best solution will achieve high accuracy as compared to other single objective swarm algorithms even when complex data has been used.

References

1. Prasad, B.R., Bendale, U.K., Agarwal, S.: Distributed feature selection using vertical partitioning for high dimensional data. In: International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 807–813 (2016)
2. Bolon canedo, V., Sanchez, N., Cervino, J.: Toward parallel feature selection from vertically partitioned data. ESANN (2014)
3. Bakshi, K.: Considerations for big data: architecture and approach. In: IEEE Aerospace Conference, pp. 1–7 (2012)
4. Chen, X., Xie, M.: A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica* **24**(4), 1655–1684 (2014)
5. Agarwal, S., Mozafari, B., Panda, A., Milner, H., Madden, S., Stoica, I.: BlinkDB: queries with bounded errors and bounded response times on very large data. In: ACM European Conference on Computer Systems (EuroSys’13), Prague, Czech Republic, pp. 29–42 (2013)
6. Lazar, N.: The big picture: Divide and combine to conquer big data. *Chance* **31**(1), 57–59 (2018)
7. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Symposium on Operating System Design and Implementation (OSDI’04), pp. 137–150 (2004)
8. Singh, D., Reddy, C.K.: A survey on platforms for big data analytics. *J. Big Data* **2**(1) (2014)
9. Jagadish, H.V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J.M., Ramakrishnan, R., Shahabi, C.: Big data and its technical challenges. *Commun. ACM* **57**(7), 86–94 (2014)
10. R. Nair.: Big data needs approximate computing: Technical perspective. *Communications of the ACM*, **58**(1), 104–104 (2015)
11. Li, K., Li, G.: Approximate query processing: what is new and where to go? *Data Sci. Eng.* **3**(4), 379–397 (2018)
12. Sagi, O., Rokach, L.: Ensemble learning: a survey. *Data Mining Know. Discov.* **8**(4), 1–18 (2018)
13. Basiri, S., Ollila, E., Koivunen, V.: Robust, scalable, and fast bootstrap method for analyzing large scale data. *IEEE Trans. Signal Process* **64**(4), 1007–1017 (2016)
14. Das, S., Agrawal, D., El Abbadi, A., Elastras.: An elastic transactional data store in the cloud. In: Conference on Hot Topics in Cloud Computing (HotCloud’09), San Diego, CA, USA, pp. 1–5 (2009)
15. Baker, J., Bond, C., Corbett, J., Furman, J., Khorlin, A., Larson, J., Leon, J.-M., Li, Y., Lloyd, A., Yushprakh, V.: Megastore: providing scalable, highly available storage for interactive services. In: Conference on Innovative Database Research (CIDR), Asilomar, CA, USA, pp. 223–234 (2011)
16. Kamal, J., Murshed, M., Buyya, R.: Workload-aware incremental repartitioning of shared-nothing distributed databases for scalable OLTP applications. *Future Gener. Comput. Syst.* **56**, 421–435 (2016)
17. Huang, Y.-F., Lai, C.-J.: Integrating frequent pattern clustering and branch-and-bound approaches for data partitioning. *Inf. Sci.* **328**, 288–301 (2016)
18. Phansalkar, S., Ahirrao, S.: Survey of data partitioning algorithms for big data stores. In: International Conference on Parallel, Distributed and Grid Computing (PDGC), pp. 163–168 (2016)
19. Shvachko, K., Kuang, H., Radia, S., Chansler, R.: The Hadoop distributed file system. In: IEEE Symposium on Mass Storage Systems and Technologies (MSST), Incline Village, NV, USA, pp. 1–10 (2010)

20. Khan, M.A., Arshad, H., Nisar, W., Javed, M.Y., Sharif, M.: An integrated design of fuzzy C-Means and NCA-Based Multi-properties Feature Reduction for Brain Tumor Recognition. *Signal and Image Processing Techniques for the Development of Intelligent Healthcare Systems*, 1–28 (2020)
21. Siddiqi, U.F., Sait, S.M., Kaynak, O.: Genetic algorithm for the mutual information-based feature selection in univariate time series data. *IEEE Access*. **8**, 9597–9609 (2020)
22. Kong, L., et al.: Distributed feature selection for big data using fuzzy rough sets. *IEEE Trans. Fuzzy Syst.* **28**, 846–857 (2020)
23. Shaw, R.N., Walde, P., Ghosh, A.: IOT based MPPT for performance improvement of solar PV arrays operating under partial shade dispersion. In: 2020 IEEE 9th Power India International Conference (PIICON), SONEPAT, India, pp. 1–4 (2020). [10.1109/PIICON49524.2020.9112952](https://doi.org/10.1109/PIICON49524.2020.9112952)
24. El-Hasnony, M., Barakat, S.I., Elhoseny, M., Mostafa, R.R.: Improved feature selection model for big data analytics. *IEEE Access* **8**, 66989–67004 (2020)
25. Paul, S., Verma, J.K., Datta, A., Shaw, R.N., Saikia, A.: Deep learning and its importance for early signature of neuronal disorders. In: 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, pp. 1–5 (2018). <https://doi.org/10.1109/ccaa.2018.8777527>
26. Fong, S., Wong, R., Vasilakos, A.: Accelerated PSO swarm search feature selection for data stream mining big data. *Serv. IEEE Trans. Comput.* **9**, 33–45 (2016)
27. Gu, S., Cheng, R., Jin, Y.: Feature selection for high-dimensional classification using a competitive swarm optimizer. *Soft. Comput.* **22**, 811–822 (2018)
28. Yan, D., Cao, H., Yu, Y., Wang, Y., Yu, X.: Single-objective/multiobjective cat swarm optimization clustering analysis for data partition. In: *IEEE Trans. Autom. Sci. Eng.* **17**(3). 1633–1646 (2020)
29. Wang, S., Eick, C.F.: MR-SNN: design of parallel shared nearest neighbor clustering algorithm using MapReduce. In: *IEEE International Conference on Big Data Analysis (ICBDA)*, pp. 312–315 (2017)
30. Sangeetha, J., Prakash, V. S. J.: An efficient inclusive similarity based clustering (ISC) algorithm for big data. In: *World Congress on Computing and Communication Technologies (WCCCT)*, pp. 84–88 (2017)
31. Barhanpurkar, K., Rajawat, A.S., Bedi, P., Mohammed, O.: Detection of sleep apnea & cancer mutual symptoms using deep learning techniques. In: 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, pp. 821–828 (2020). <https://doi.org/10.1109/i-smac49090.2020.9243488>
32. Singh Rajawat, A., Jain, S.: Fusion deep learning based on back propagation neural network for personalization. In: 2nd International Conference on Data, Engineering and Applications (IDEA), Bhopal, India, pp. 1–7 (2020). <https://doi.org/10.1109/idea49133.2020.9170693>

Efficient Deep Learning for Reforming Authentic Content Searching on Big Data



Anand Singh Rajawat, Kanishk Barhanpurkar, S. B. Goyal, Pradeep Bedi, Rabindra Nath Shaw, and Ankush Ghosh

Abstract With the advancement of search engines, a major change has occurred in the way people are accessing data on the net. Search engines have made access to data efficient and easier as billions of pages on the net (or called big data) are suggested at once. The pages with the most significant rank generally have a higher visibility rate to people and hence every webmaster wants to push their page to higher rank. As a result, Search Engine Optimization (SEO) has become a massive business which strives in enhancing the ranking of clients' webpage. But there are many myths and misconceptions about the ranking algorithms due to inadequate knowledge about SEO's methods. Still there is a need to propose a verified algorithm that is efficient in retrieving pages from Google. The link analysis algorithm is in accordance with the link structure of any document as the page which has many links also has many connections to it and hence can increase retrieval capacity. There is another approach called the integrated ranking approach which comes under personalized web research. In the integrated ranking approach, both content and link are used as parameters to improve retrieval efficiency. This approach is used by

A. S. Rajawat (✉)

Department of Computer Science Engineering, Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore, India

K. Barhanpurkar

Department of Computer Science and Engineering, Sambhram Institute of Technology, Bengaluru, Karnataka, India

S. B. Goyal

City University, Malaysia, Malaysia

P. Bedi

Department of Computer Science Engineering, Lingaya's Vidyapeeth, Faridabad, Haryana, India

R. N. Shaw

Department of Electrical, Electronics & Communication Engineering, Galgotias University, Noida, India

e-mail: r.n.s@ieee.org

A. Ghosh

School of Engineering and Applied Sciences, The Neotia University, Sarisha, West Bengal, India

Google using Deep Learning to increase retrieval efficiency of web pages as per the user's requirement from the big data.

Keywords SEO · Deep learning · Web page ranking · Static ranking · Search engine

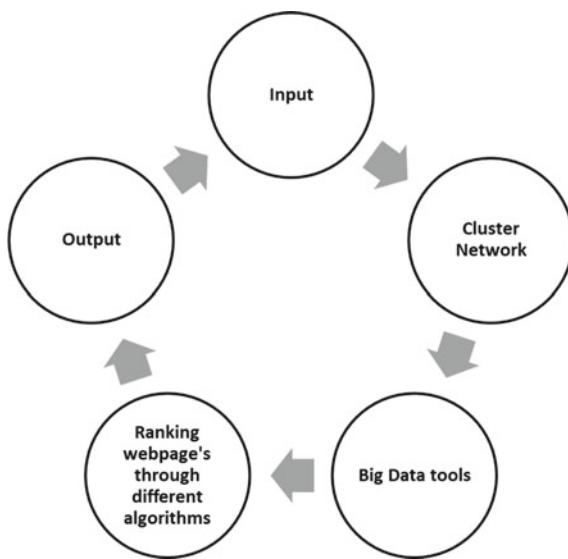
1 Introduction

The search algorithms have been continuously updated again and again since 1998. Big companies like Google tend to improve features that already exist instead of just proceeding with any new innovation [1]. As an example, the companies try to simply revise the ranking algorithms in order to better determine the content quality, instead of using scrapping content quality as the factor for ranking. These logical and revolutionary enhancements are not likely to disappear shortly, but some new approaches of rolling out these updates can have the power to modify how we use, see or take advantage of web page search ranking. This is all based on machine learning. The rising fields of web mining intend for finding and extracting suitable data from the web. Web Mining is similar to big data web mining, as it is also a multi-disciplinary effort which extracts techniques from various fields like statistics, Natural Language Processing (NLP), information retrieval, machine learning, etc. Search engines are used by a large number of users to find relevant information from the Internet as search engines play a very important role in information retrieval. Big web mining is broadly classified into three sub-areas namely: Web content mining, web structure mining and web usage mining. Finding relevant information from the massive data sources available on the internet has become very important for every user (Fig. 1).

The major issue is that most of the users (70%) try to find the appropriate result by only navigating the first two pages and ignoring the rest of the pages even without getting suitable information. There are a number of ways to resolve the issue of presenting the most useful result on the top of web page search. Showing the most relevant information at the top of the search is the most common approach to resolve the problem and this method is known as page ranking. Due to a huge amount of information present today, it has become difficult for website owners to provide the appropriate information for the internet users. If we look into the working of a typical search engine, we find that it shows the search graph of the previous web user [2].

The efficient ranking of the query words plays a major role in the process of efficient searching. The ranking of the web pages suffers from many challenges such that some pages on the web are only made for navigation purpose and some pages do not even have the quality of self-descriptiveness. Several algorithms are proposed in the literature for the ranking purpose of web pages. The objective of this paper is to examine the current important algorithms proposed for web page ranking, to determine the strengths and drawbacks of the respective algorithms and then to provide some important research in determining efficient algorithm for the ranking

Fig. 1 Workflow model of ranking webpage using big data tools



of the web page. We suggested a new algorithm in our paper which is a merger of two approaches. Deep learning algorithm and other algorithms have been implemented in this work to determine the web structure and content mining, respectively.

Overlooking PageRank. Out of many other factors, PageRank scores are considered important since the time Google was born. Many experts in SEO and other users underestimated this factor. There is only 1 out 3 lists that may include PageRank among the top ten factors. Although, the results offered by ranking algorithm of Google, PageRank score is considered among the top factors for page ranking [3].

Role of keyword in URL. If we consider the results, the suggestions that can be given to SEOs or web users include: Firstly, choose a suitable filename as that can be an important factor to select a good title. It was also discovered that the keywords used in the URL are essential as per the ranking algorithm of Google [4]. However, it was found that only a few lists like the survey consider this factor whereas other lists by SEO experts like SEOmoz'07 do not think this factor as important.

Meta-description tag counts. The ranking function of the search engine has been a matter of dispute for a long time and the reason being whether metatags are important or not. Some think that metatags can be abused by webmasters as they can embed unnecessary keywords into them and this could be contributing to the reluctant behaviour of search engines to use metatags in ranking. Recent findings show that meta keyword doesn't have much influence going by the overall rankings. However, going by the ranking algorithm of Google, Meta description certainly dominates among the ranking factors.

2 Related Work

Many search engines are using various types of ranking algorithm that has to be studied and results of Google, Yahoo and Bing are analysed properly using our own ranking system. Ao-Jan Su et al. [5] analysed the system of ranking for Bing search engine. He proved that to improve the accuracy of Bing, we need a recursive partitioning algorithm that can be beneficial. Another factor that can contribute to the prediction of Bing's result by improving the ranking system and its accuracy is the PageRank scores that were acquired from Google. Another data revealed about the dispute between the two mentioned search engines regarding various ranking features. Bing gives much importance to the meta keyword tag while Google prefers meta-description tag. Chen and Décarie [6]. This paper's approach is to use rule-based semantic analysis. It gives priority to the user's intention and context related to different areas and relates it to the search engines. It goes with the proposed model and frames to conduct analysis and extraction of digital text. Its main goal is to improve the content and its relevance. The efficiency and usefulness of information search are also improved. The description of this method lies in its architecture and various processes.

Prabha et al. [7]. All the other algorithms that were previously explained are efficient in doing the process of retrieval by extracting web pages. This algorithm is based on linking the structure of documents and is known as the link analysis algorithm. The efficiency of a page can be improved by the number of links and references it contains. The personalized web search contains an integrated ranking approach in which the link and content are combined together to improve the efficiency of the system. Similarly, segmentation of the page is done in which a page is separated into blocks that result in improving the retrieval efficiency of the web. Such algorithms are page segmentation algorithms. Hence, each type of algorithm can have its own pros and cons [8]. This research paperwork is based on studying and enforcing On-page and Off-page optimization method applied on an educational website. These techniques are widened with some additional components, for instance, displaying the web pages as graph objects.

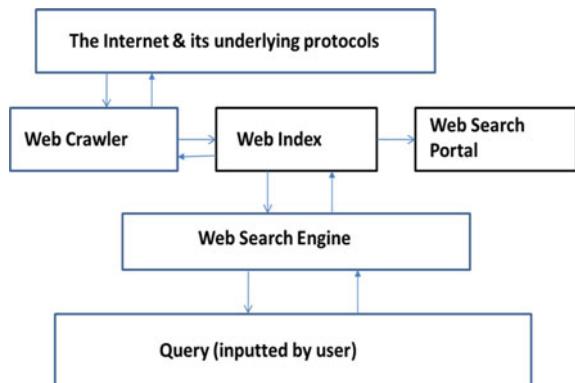
3 Proposed Approach

We improved the effectiveness and efficiency of search engines, and a machine learning technique is used which is known as PageRank. PageRank is used to calculate the need and relevancy of the page and order the pages coming back from a traditional search engine according to their priority using known keyword search. Google search engine is the best example to measure the success and effectiveness of this technique. Links on the web pages are analysed with the help of big data on the internet. The number of web pages pointing to it determines the PageRank value of that page. The calculation of PageRank value is basically the measurement of backlinks to

that particular page and backlinks are links that points web page instead of going out from the web page. Considering other techniques which analyse links, this technique shows distant features [9]. Generally, there is a misconception that this technique takes all the links as the same but it is not true. If the links are coming from high priority web pages, then weighting is used and that's why PageRank value is not just based on analysing backlinks. With the help of a deep learning technique, we created a ranking system that can meet the aforesaid design goal. The Crawler and the Ranking Engine are the two important components of our system. It is very difficult and complicated to isolate the dependency of individual features which can influence the ranking and this was clearly seen in our early unsuccessful attempts, for example, one at a time. That's why we resort to combine multiple features as a consideration. This concludes that collecting and analysing the search results and evaluating search engine can be done by search queries. The Crawler performs the task of data collection and takes the results of ranked search which queries the search engine. It queries domain information as mentioned and downloads the HTML Web pages from original Websites, additionally [10]. Second, the ranking engine extracts key properties under study from the raw web and learns to evaluate other ranking models to approximate the search engine ranking results. We have certain contributions to this part. We make sure that the ranking function from the search engine is not a usual linear function among all the features by depicting a nonlinear model outperforming the linear model. Nonlinear model's ranking is better than a simple linear model, in ranking a search engine. However, it is hard for users to accept the nonlinear model. Our second contribution is a simple linear model performing simple recursive ranking procedures and achieving comparable results to the nonlinear model. Theoretically, the nonlinear model can approximate effectively by recursive application of the linear model. In addition, human convenience and effective results are very finely covered by the linear model (Fig. 2).

Malicious spamming and static ranking the number of incorrect (e.g. phishing) sites has grown rapidly since last few years. Discovering useful information from the search engine has come to increasing reliance due to a sheer number of good and

Fig. 2 Schematic diagram of the interaction between internet and user



bad pages on Websites. The users not only want the result related to their query but also want to separate the good information from bad considering that the useful and best pages come first.

Static Ranking Types

- RankNet
- fRank is a feature-based ranking that uses RankNet for ranking.
- fRank
- Pairwise accuracy ranking measure, a pair of documents with the same human judgement.
- Pairwise Ranking.

Benefits

A large number of features can be concluded from a good Static Ranking Algorithm:

- Efficiency
- Relevance
- Crawl Priority.

RankNetRank uses RankNet algorithm.

On the issues of classification and regression, machine learning has contributed much more in terms of evaluating the need and performing the maximum task.

4 PageRank

A web link connected from one web page to another can be seen as a support of that page. This concept is the basic idea behind web PageRank. In fact, links are created by people only. Links indicate the quality of the web pages to which they point. This linkage information can be used to serialize the web pages in accordance with the observed quality. Analysis of search engine's performance regardless of important complexity is a feasible job. This is one of the contributions of our work. Traditionally, SEO optimization is the key application used for such analysis. Auditing search engines is another relevant and growing area, i.e. identifying a possible search bias. The most consistent analysis of any search engine brings the fear that they create reality instead of reflecting it. We explained below why a confined set of features are still efficient to provide high accuracy, and why a comprehensive set of features will be useful for a practical application [11]. For instance, it is possible that features like user backlink mined from search engine logs might be used by a search engine as one of the dignified features. Even if we are aware that this is the possibility, still it would be demanding to calculate exactly or guess such kind of information from the endpoint. But this type of parameter can be ignored if we carefully examine the system, i.e. by using suitable keywords (Table 1).

Table 1 Comparison of various studies carried out on applications of big data

Study	Year	Big data application	Description
Takura et al.	[12]	Healthcare industry	Data related to patient for long-term care
Bai et al.	[13]	Urban planning	Data related to the planning of major cities
Alam et al.	[14]	Urban planning	Data related to Smart city as Neom City in Saudi Arabia
Harrigan et al.	[15]	Digital Marketing	Data related to finding influencer on the Internet
Au-Yong-Oliveira et al.	[16]	Healthcare industry	Healthcare industry
Chaudhari and Sinha	[17]	Business related	Data related to startup funding in India
Shao et al.	[18]	Supply chain application	Data related to supply chain management to increase productivity

The following processes come under the method:-

Many benefits of using deep learning are to keep algorithm step-by-step progressing and it; earn from its own experiences considering each user differences. Large numbers of user behaviour can be understood just by “general” pattern, but there is always a chance for distinct individual differences to account for. For these types of exceptions, only a deep learning algorithm could give their time, speed and scalability. Algorithms having prospective to gain by themselves can update faster comparing to those that rely on manual inputs. Similarly, they have amazing scalability; last till you have the processing power to devote, billions of queries to a machine learning algorithm could be applied the same way it could be applied to a dozen [19].

The x-axis contains the public leaderboard ranking and the y-axis contains the private leaderboard ranking. The orange region depicts the region where the ranking is very low and the blue region depicts that ranking is good and that webpage is properly optimized (Fig. 3).

Unlimited application potential. Many machine learning algorithms are used by Google in the search world, but this is not the only potential application here. Any element of Google’s core search algorithm could be improved by the deep learning process. This does not look possible at this point, as deep learning algorithm also need someone to look after and need to be adjusted time by time and there may always be a place for better machine learning technology. However, the need for technical search roles could be affected by significantly increased automation in the search engine.

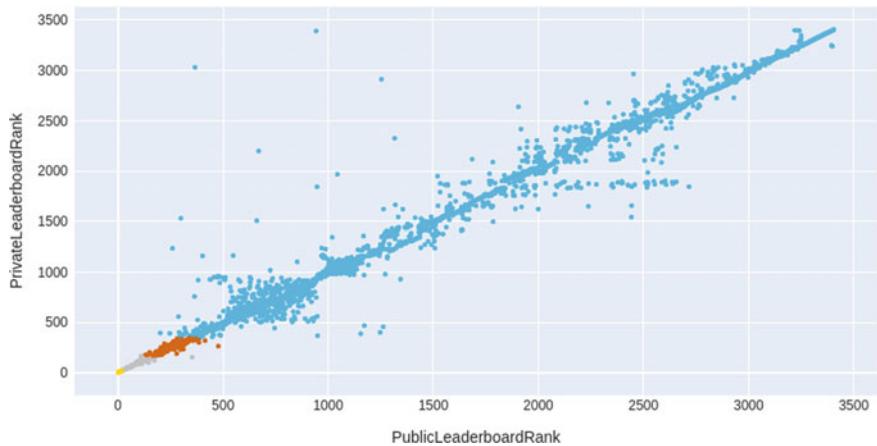


Fig. 3 The correlation between data collected for various websites based on the private leaderboard ranking and public leaderboard ranking

5 Conclusion and Future Work

In this paper, the authors have analyzed the method of big data in ranking websites and the factors that affect the overall ranking of websites. Big data technologies play a vital role in the development of many interdisciplinary domains such as urban planning, healthcare domain, project management and digital marketing. Also, the big data ranking algorithm is examined which helps to understand the working of the Search engine optimization in this process. It also concludes that big data helps in the integration of digital marketing and promotes the business among local vendors, merchants and tech-giant like Amazon, Flipkart, Walmart and Google. The author also describes the interaction between internet protocols and the user with help of a search engine. Finally, Big Data platforms will be incorporated and harmonized to facilitate a full and flawless assimilation between different Big Data tools and clients.

References

1. Matošević, G., Dobša, J., Mladenić, D.: Using machine learning for web page classification in search engine optimization. *Future Internet* **13**(1), 9 (2021)
2. Dery, L.: Multi-label ranking: mining multi-label and label ranking data (2021)
3. Kherbachi, S., Yang, Q., Khan, S.Z.: A structured approach to measuring and optimizing the organizational architecture in global product development projects. *Concurr. Eng.* **28**(3), 161–174 (2020). <https://doi.org/10.1177/1063293X20929388>
4. Wang, J., Wang, L., Ye, K., Shan, Y.: Will bid/no-bid decision factors for construction projects be different in economic downturns? A Chin. Study. *Appl. Sci.* **10**(5), 1899 (2020)
5. Ao-Jan Su, Y., Charlie, H.U.: Aleksandar Kuzmanovic, cheng-kokkoh (2014)

6. Chen, M., Décarie, M.: A cognitive-based semantic approach to deep content analysis in search engines. In: 2018 IEEE 12th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, 2018, pp. 131–139 (2018). <https://doi.org/10.1109/icsc.2018.00027>
7. Prabha, S., Duraiswamy, K., Indhumath, J.: Comparative analysis of different page ranking algorithms. World Academy of Science, Engineering and Technology. Int. J. Comput. Inf. Eng. **8**(8) (2014)
8. Krrabaj, S., Baxhaku, F., Sadrijaj, D.: Investigating search engine optimization techniques for effective ranking: a case study of an educational site. In: 2017 6th Mediterranean Conference on Embedded Computing (MECO), Bar, 2017, pp. 1–4 (2017). <https://doi.org/10.1109/meco.2017.7977137>
9. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Zettlemoyer, L., et al.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint [arXiv:1910.13461](https://arxiv.org/abs/1910.13461) (2019)
10. Kherbachi, S., Benkhider, N., Keddari, N.: Application of pagerank in virtual organization architecture. Res. J. Comput. Inf. Technol. Sci. **1**(1), 1–14 (2020)
11. Li, G., Chen, C., Zhang, G., Martek, I.: Bid/no-bid decision factors for Chinese international contractors in international construction projects. Engineering, Construction and Architectural Management (2019)
12. Takura, T., Goto, K.H., Honda, A.: Development of a predictive model for integrated medical and long-term care resource consumption based on health behaviour: application of healthcare big data of patients with circulatory diseases. BMC Med. **19**(1), 1–16 (2021)
13. Bai, Y., Yang, R., Xiao, J., Wang, Z., Xie, P., Cheng, X., Gao, J.: Exploration and practice of china unicorn multi-scene capability construction and open platform for large-scale heterogeneous data. In: Signal and Information Processing, Networking and Computers, pp. 1019–1026. Springer, Singapore (2021)
14. Alam, T., Khan, M.A., Gharaibeh, N.K., Gharaibeh, M.K.: Big data for smart cities: a case study of NEOM city, Saudi Arabia. In: Smart Cities: A Data Analytics Perspective, pp. 215–230. Springer, Cham (2021)
15. Harrigan, P., Daly, T.M., Coussement, K., Lee, J.A., Soutar, G.N., Evers, U.: Identifying influencers on social media. Int. J. Inf. Manag. **56**, (2021)
16. Au-Yong-Oliveira, M., Pesqueira, A., Sousa, M.J., Dal Mas, F., Soliman, M.: The potential of big data research in healthcare for medical doctors' learning. J. Med. Syst. **45**(1), 1–14 (2021)
17. Chaudhari, S.L., Sinha, M.: A study on emerging trends in Indian startup ecosystem: big data, crowd funding, shared economy. Int. J. Innov. Sci. (2021)
18. Shao, X.F., Liu, W., Li, Y., Chaudhry, H.R., Yue, X.G.: Multistage implementation framework for smart supply chain management under industry 4.0. Technol. Forecast. Soc. Change **162**, 120354 (2021)
19. Smith, K.G., George, W.B.: U.S. Patent No. 10,540,660. Washington, DC: U.S. Patent and Trademark Office (2020)

IoT as a Platform: For Smart Home Analysis and Monitoring of Fire Parameters



Sudip Suklabaidya and Indrani Das

Abstract There are various natural havocs where human lives and belongings are always susceptible to certain life threats. Here, fire outbreak is taken into consideration in respect of home which can be major havoc in terms of human loss and belongings. To prevent and avoid such fire outbreak, we need to cover certain fire parameters like temperature, flame, and smoke, as these parameters help to create an alert whenever there is an outbreak of fire. This paper aims at designing and implementing a low-cost system for real-time monitoring of fire alert system using an IoT platform. The proposed system deals in controlling various conditions for fire outbreak through sensors implemented and the final data collected after analysis are then sent to the webserver and can be accessed over the internet to find out the outcome. This paper projects on Sensor-Based Smart Home (SBSH) fire Monitoring techniques which measure various parameters, i.e., smoke, flame, and temperature of a fire in the home. The Controller Arduino UNO and NodeMCU are used to process the values collected from various sensors. Finally, the sensor data can be viewed graphically using ThingSpeak API. The main purpose of the paper is to achieve a smart home monitoring system with high frequency, real values, and low power.

Keywords Sensor · Arduino UNO · ThingSpeak cloud · NodeMCU

1 Introduction

The Internet of Things offers a wide range of applications where a large number of information is extracted computationally to provide a real-time and sensible output in many cases. In fact, the Internet of Things (IoT) concept and technology differs from conventional information technology because of its reusable and embedded

S. Suklabaidya (✉)

Department of Computer Science and Application, Karimganj College, Karimganj, India

I. Das

Department of Computer Science, Assam University, Silchar, India

nature that supports our day-to-day life. Apart from this, it emphasizes on various types of smart system which enhances the life pattern in the present day.

As the adaptability toward smart life has convinced people of this era for better life and convenience, time again has played a major feedback toward this approach. Since people focus more on less time and more productivity, the technological era has well utilized this fact in bringing out the best of best. Their vision and implementation for convenient life at an affordable structure have encompassed lives and livelihood in a new way which were never taken into consideration in earlier period, and with the advancement of technology and its applications, the entire world seems to be defined in a single term as “Smart World”. Among these smart implementations, IoT has paved a new way and a greater area of research. IoT, as defined from its abstract of application, is said to be a connection of physical objects over the internet wherein these objects can collect and share or exchange data among themselves. Though the area is still on its research bud, with its outspread and usage, the IoT can be better classified in the future time. Now considering IoT in terms of applications, any physical objects may be considered and implemented, it can be a Television or a Car or any other device, etc.

This system aims at providing a low-cost method to protect homes, offices, and buildings from fire outbreak in a more efficient way. The components needed to establish such a system do not incur much cost. Moreover, the installation procedure is quite easy which requires programming knowledge in NodeMCU and Arduino, and in some cases, it can be customized based on the need of the system. A strong network works as a backbone to gather data from the various sensors which in turn send data continuously to the server as long as the sensors are connected over network. It is also taken into consideration that these sensors can send data during a fire outbreak, though sometimes they may get demolished in the course of a significant fire break and some sections of the network may also fail.

In this paper, a fire detection model has been developed using sensors like LM35 (Temperature Sensor), MQ2(Smoke Sensor), YL-44 (Buzzer Module), and Flame sensor. All these sensors along with Wifi module NodeMCU are connected through Arduino Board, an 8-bit microcontroller which generates an integrated control framework. The system on receiving input is analyzed and produces real-time output through ThingSpeak Server in graphical format. The main objective of this paper is to build a system through which information can be sent remotely to the user or to the competent authority when the fire is being detected.

The paper is synchronized as follows: Sect. 2 in the paper highlights review of the previous related works done to solve the stated problem, Sect. 3 describes the proposed approach with a block diagram, Sect. 4 elaborates the working mechanism of the proposed system. Section 5 presents the results and findings and lastly, Sect. 6 concludes the paper.

2 Related Work

This section describes the analysis of earlier work on real-time monitoring of values, detection, and findings of fire-based sensors as well as other related application areas based on various sensors, Arduino UNO, Raspberry Pi controller, and other useful techniques.

Perilla et al. [1] prepare a framework for fire alert and security system where they use sensors like DH11 for temperature and humidity, flame sensor, MQ2 sensor for smoke, and PIR sensor for theft prevention. Also, they prepare a questionnaire to measure and evaluate from the various respondents about the usability of the model developed by them, where they found that it is highly usable and important in respect of efficiency, helpfulness, and learnability. Khalaf et al. [2] propose a fire alert system model to avoid risk from the fire outbreak. They used controller NodeMCU and DHT-11sensor in order to measure heat around areas. Mahzan et al. [3] also work on the design of a home fire alarm where they used Arduino UNO and LM35 as temperature sensor along with GSM module to notify the competent authority about the rise in temperature in the house. Gosrani et al. [4] also propose a low-cost fire alert system that comprises of various sensors. The system will notify all the competent authority if there is any hazard in home or any other workplace. Yadav et al. [5], Shah et al. [6], Ajith et al. [7], and Kumar et al. [8] also propose a fire detection based on Arduino board where the system will detect fire in the surroundings and will send the message or call through GSM module.

Imteaj et al. [9] proposed an IoT-based fire alarming and authentication system where they used two controllers Raspberry Pi 3 and Arduino Mega 2560 rev3. They used a camera sensor to receive snaps. The controller receives data and transfers the data to the authority through the GSM module for the validity of the message. Angeline et al. [10] also use a Raspberry Pi controller along with PIR, Gas, and Camera sensor. They tried to tackle false alarm generated from the model by using a camera sensor. The system also sends a mail to the owner of the workplace along with an attachment of a photo of the fire outbreak. Nancy et al. [11] also proposes a framework for fire detection with Raspberry Pi controller where they take readings of various fire sensors in different circumstances. They also proposed that the system can be implemented further with other sensors like flame sensor and light sensor for better accuracy. Sangam et al. [12] also prepared an IoT-based fire detection monitoring system where they used fire sensors along with a water pump and servo motor. The sensor provides value to the monitor and will send the message through the GSM module when it crosses the threshold value. They also told that the system can be further prepared more and more efficiently by including some other features. Jhuang et al. [13] prepare a smart kitchen prevention system where they used three control panels' Arduino for reading the sensors detection data, D1 Mini for notifications, and Webduino for remote controls and Wifi. They used three sensors for detecting kitchen fire. When any of the sensors exceed the threshold value, the relevant alarms, notification, etc. get activated immediately. Yuen et al. [14] prepared an IoT-based fire safety system using MQTT communication. The system consists of three parts:

detector, processing unit, and surveillance. Raspberry Pi is used to run the Node-RED application which processes and monitors data. The message, voice call, and location will be sent to the owner on the occurrence of fire.

The author Chelliah et al. [15] used IoT Microcontroller, Wifi module, and various types of sensors to measure the temperature and humidity of closed areas so that they can be controlled at normal standard values. They apply the machine learning algorithm polynomial regression in order to predict the upcoming temperature based on past values.

Banka et al. [16] build a medical aid system for monitoring patient medical information using IoT Kit. They used various medical sensors to measure the heart rate, blood pressure, etc. of patients in order to analyze and detect first hand disease using data mining techniques. The information collected from the system will provide the patient and medical community with taking decisions.

Singh [17] in his paper builds an alert system for a farmer using IoT kit where they could get information regarding the weather condition of the field. The author used various sensors to measure the different climatic parameters of the field. Girija et al. [18] also proposed a model in order to predict the climatic situations of a particular place, cities, and industrial areas. The data collected from the system can be accessed from any place through the internet. Wadhwani et al. [19] build smart home management with various sensors where they tried to automate the home appliances such as sensing, monitoring, and controlling the system. Tarone et al. [20] work on “Cost-effective Smart Garbage monitoring system”, using ESP8266 where the system will monitor the waste products and will alert the concerned authority when the garbage level marks the threshold value.

3 Proposed System

3.1 Working Mechanism

The working mechanism of the proposed model emphasizes limiting the drawbacks which are present in the existing system. Here, in this proposed model, we have used Arduino UNO as the central controller and various sensors to observe the fire parameters. The architecture of the proposed system is represented in Fig. 1. Arduino Operating System executes on the Arduino UNO that serves to achieve various types of equipment including sensors and so on. Here, we concentrated on connecting different sensors with Arduino UNO and NodeMCU to measure different parameters of fire in the home. The Arduino UNO then accesses the data from the sensors used in this working mechanism and processes the data. The collected sensor data is then viewed on the cloud using ThingSpeak App. The graphical format that is generated from the analog data and received from the sensors is later described with accurate values.

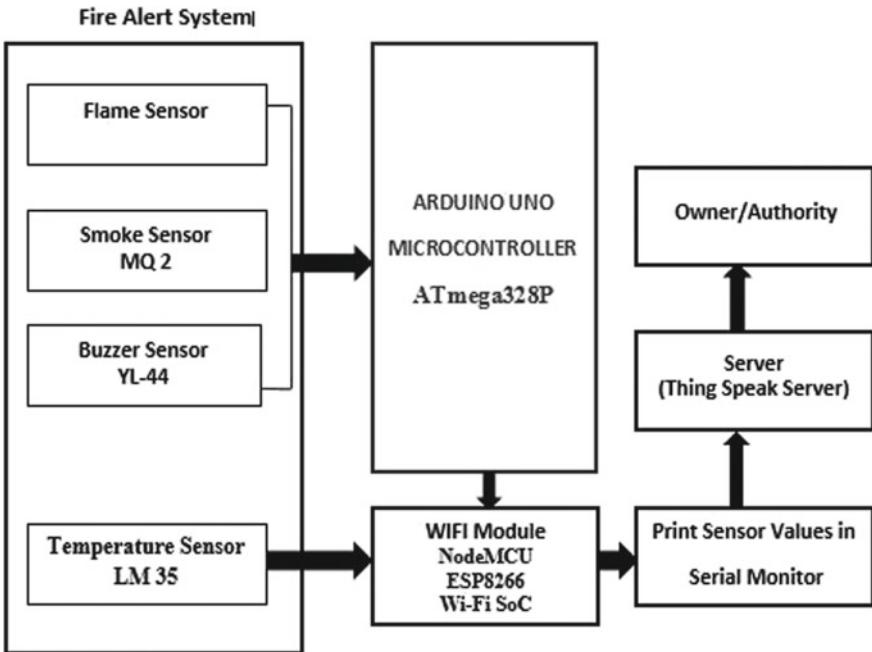


Fig. 1 Architecture of the proposed system

4 Experimental Setup and Implementation

The various components of the proposed model are integrated primarily in a breadboard. As NodeMCU is having only one analog pin, so flame sensor, smoke, and buzzer are connected with Arduino and the temperature sensor LM 35 is kept connected with A0 pin of NodeMCU. The flame sensor is connected with Arduino A0 pin, the smoke sensor (MQ2) is connected with Arduino A1 pin, and the buzzer sensor with Arduino digital 8 pin. Also Ground (Gnd) and power supply of flame, buzzer, and smoke are connected with Arduino, and the ground and power supply of LM35 are connected with NodeMCU. The flame and smoke on sensing, will send data from Arduino to NodeMCU via serial communication. Here, Arduino is used as a transmitter that is transmitting data from sensors, and NodeMCU is used for receiving data through serial communication in JSON format. The digital pin 5 and digital pin 6 are connected with D5 and D6 pin of NodeMCU for serial communication. NodeMCU will then send data to the ThingSpeak server. The serial monitor will display all the values of the sensors along with the presence of fire or not in the monitor, shown in Figs. 2 and 3.

Moreover, a dataset of around 1000 is being created from the readings of all the three sensors generated from the fire alert system, which can be used for further analysis to find out the accuracy, precision, etc. A sample data is depicted in Table 1

```

COM5
Send
Autoscroll Show timestamp
Newline 9600 baud Clear output

14:22:06.549 -> Received Smoke: 261.00
14:22:06.549 -> Received Flame: 948.00
14:22:06.597 -> Temperature = 29.70 °C
14:22:06.597 -> Fire Detect:: 0
14:22:06.643 -> -----
14:22:08.554 -> Sensors Data Recieved
14:22:08.554 -> Received Smoke: 221.00
14:22:08.589 -> Received Flame: 948.00
14:22:08.650 -> Temperature = 29.70 °C
14:22:08.650 -> Fire Detect:: 0
14:22:08.650 -> -----
14:22:10.589 -> Sensors Data Recieved
14:22:10.589 -> Received Smoke: 215.00
14:22:10.636 -> Received Flame: 947.00
14:22:10.636 -> Temperature = 29.70 °C
14:22:10.636 -> Fire Detect:: 0
14:22:10.636 -> -----

```

Fig. 2 Sensor values along with the absence of fire

```

COM5
Send
Autoscroll Show timestamp
Newline 9600 baud Clear output

14:22:59.473 -> Sensors Data Recieved
14:22:59.520 -> Received Smoke: 368.00
14:22:59.520 -> Received Flame: 76.00
14:22:59.567 -> Temperature = 29.40 °C
14:22:59.567 -> Fire Detect:: 1
14:22:59.613 -> -----
14:23:02.534 -> Sensors Data Recieved
14:23:02.534 -> Received Smoke: 354.00
14:23:02.581 -> Received Flame: 59.00
14:23:02.581 -> Temperature = 29.70 °C
14:23:02.630 -> Fire Detect:: 1
14:23:02.630 -> -----

```

Fig. 3 Sensor values along with the presence of fire

Table 1 Real-time readings

Readings	Smoke (ppm)	Flame (Hz)	Temperature (C°)	Fire Observ
1	285	200	27	1
2	412	954	27	1
3	325	58	28.8	1
4	394	86	28.8	1
5	290	915	29.7	1

showing the real values of all the three sensors, and Table 2 presents the measured values along with the presence or absence of fire.

Simultaneously, a personal account is created in the ThingSpeak server, an IoT operation platform with Channel ID and password. An Application Program Interface

Table 2 Sensor readings from the system

Readings	Smoke (ppm)	Flame (Hz)	Temperature (C°)	Fire detect
1	234	991	27	0
2	412	954	27.90	1
3	325	58	28.8	1
4	394	86	28.8	1
5	290	915	29.7	0

(API) key is generated after the new channel log in. The write API key is then inserted along with all the credentials such as Channel Id and Service Set Identifier, i.e., the network name, network password, and server address into the NodeMCU. The ThingSpeak library function also needs to be included in Arduino and NodeMCU for sending various sensors measured values to the server. The system will then send data after every 15 s from Arduino and NodeMCU to the server.

5 Results and Analysis

The proposed system consists of three module smoke, flame, and temperature which is here in Figs. 4 and 5 illustrated as field 1, field 2, and field 3, respectively. Field 4

**Fig. 4** Representing fire system showing the absence of fire



Fig. 5 Representing fire system showing the presence of fire

represents the complete fire system which shows the presence or absence of fire in the surrounding areas. In the smoke and flame sensor, when it crosses the threshold value, the buzzer sensor connected with Arduino UNO is automatically invoked with an alarming sound that the fire has been detected. The values are passed from the NodeMCU to the server which are graphically represented below as “Fire Detection: 0”, i.e., the absence of fire and “Fire Detection 1”, i.e., the presence of fire (either the presence of smoke, flame, or both) in the surroundings.

6 Conclusions

The paper “IoT-Based Low-Cost Monitoring of Real-Time fire parameter” aims at analyzing the values that are performing high in real time and accurate. In our proposed system, we have gathered values of smoke, flame, and temperature sensor and had sent the data to the server for further notification to the user for the fire outbreak. The system is designed for constantly monitoring the values of the sensor used in the smart home and the presence or absence of fire in a room. The primary purpose of the paper is to develop a fire alert monitoring system that could provide accurate values. The results or values obtained from the system are relatively efficient for the fire detection system. The dataset obtained from the model can be used further by using several techniques in finding out the error rate, precision, accuracy,

loss function, etc. The proposed prototype can be useful in smart cities such as offices, factories, and firms due to its reliability, flexibility, and ease of handling. However, with the development and increase of sensor range, the system could be more resourceful and effective. Also, a GPS module can be included in the model that can specify the location of the affected region and can notify the competent authority immediately.

References

1. Perilla, F.S., Jr Villanueva, G.R., Cacanindin, N.M.: Fire safety and alert system using arduino sensors with IoT integration. In: 7th International Conference on Software and Computer Applications February 2018, pp. 199–203 (2018)
2. Khalaf, O.I., Abdulsahib, G.M., Zghair, N.A.K.: IOT fire detection system using sensor with Arduino. In: REVISTA AUS 26–1 (Sept 2019). <https://doi.org/10.4206/aus.2019.n26-7>. [www.ausrevista.com/editor@ausrevista.com](http://ausrevista.com/editor@ausrevista.com)
3. Mahizan, N.N., Enzai, N.I.M., Zin, N.M., Noh, K.S.S.K.M.: Design of an Arduino-based home fire alarm system with GSM module. In: 1st International Conference on Green and Sustainable Computing (ICoGeS) 2017
4. Gosrani, S., Jadhav, A., Lekhak, K., Chheda, D.: Fire detection, monitoring and alerting system based on IoT. Int. J. Res. Eng. Sci. Manag. **2**(4), 2019
5. Yadav, R., Rani, P.: Sensor based smart fire detection and fire alarm system. In: Proceedings of the International Conference on Advances in Chemical Engineering (AdChE) (2020)
6. Shah, R., Satam, P., Sayyed, M.A., Salvi, P.: Wireless smoke detector and fire alarm system. Int. Res. J. Eng. Technol. (IRJET) **06**(01) 2019. e-ISSN: 2395-0056
7. Ajith, G., Sudarsaun, J., Dhilipan Arvind, S., Dr. Sugumar, R.: IOT based fire deduction and safety navigation system. Int. J. Innov. Res. Sci. Eng. Technol. **7**(2) (2018)
8. Kumar, N., Kumari, P.: Intelligent fire detection and visual guided evacuation system using Arduino and GSM. IJRTER **04**(12) (2018). ISSN: 2455-1457
9. Imteaj, A., Rahman, T., Hossain, M.K., Alam, M.S., Rahat, S.A.: An IoT based Fire alarming and authentication system for workhouse using Raspberry Pi 3. In: International Conference on Electrical, Computer and Communication Engineering (ECCE), 16–18 Feb 2017, Cox's Bazar, Bangladesh (2017)
10. Angeline, R., Adithya, S., Narayanan, A.: Fire alarm system using IOT. Int. J. Innov. Technol. Explor. Eng. (IJTEE) **8**, (2019). ISSN: 2278-3075
11. Andrea, L., Abirami, R., Diviya, M., Silviya Nancy, J.: Framework for fire detection and mitigation using IoT. Int. J. Pure Appl. Math. **118**(18), 1801–1811 (2018)
12. Sangam, K., Prasanna, T., Bramaramba, K.: An IoT based fire detection, precaution & monitoring system using Raspberry Pi3 & GSM. Int. J. Eng. Res. Technol. (IJERT) **8**(07), 2019
13. Hsu, W.-L., Jhuang, J.-Y., Huang, C.-S., Liang, C.-K., Shiao, Y.-C.: Application of internet of things in a kitchen fire prevention system 27 Aug 2019
14. See, Y.C., Ho, E.X.: IoT-based fire safety system using MQTT communication protocol. Int. J. Integr. Eng. **12**(6) 207–215 (2020)
15. Chelliah, B.J., Anand, A., Kaul, A., Pathak, M.: Temperature capstone and humidity monitoring using IoT with machine learning algorithm. Int. J. Eng. Adv. Technol. (IJEAT) **9**(2) (2019). ISSN: 2249-8958
16. Singh, A.K.: Applications of IoT in agricultural system. Int. J. Agric. Sci. Food Technol. 26 May 2020. ISSN: 2455-815X
17. Girija, C., Grace, S.A., Harshalatha, H., Pushpalatha, H.P.: Internet of Things (IOT) based weather monitoring system. Int. J. Eng. Res. Technol. (IJERT). ISSN: 2278-0181. Published by, www.ijert.org. (NCESC-2018 Conference Proceedings)

18. Wadhwani, S., Singh, U., Singh, P., Dwivedi, S.: Smart home automation and security system using Arduino and IOT. Int. Res. J. Eng. Technol. (IRJET) **05**(02) (2018). e-ISSN: 2395-0056, p-ISSN: 2395-0072
19. Tarone, A.V., Katgube, A.A., Shendre, H.H., Ghugal, R.P., Bobade, N.P.: IOT based smart garbage monitoring system using ESP8266 with GPS link. Int. Res. J. Eng. Technol. (IRJET) **05**(03) (2018). e-ISSN: 2395-0056, p-ISSN: 2395-0072
20. Anuradha, T., Bhakti, C.R., Pooja, D.: IoT based low cost system for monitoring of water quality in real time. Int. Res. J. Eng. Technol. (IRJET) **05**(05) (2018). e-ISSN: 2395-0056, p-ISSN: 2395-0072

Death Prediction in the Current Pandemic Scenario and Cluster Classification Using Soft Computing Techniques



Loshima Lohi and Maya L. Pai

Abstract Coronavirus is a profoundly irresistible infection and has ruinous impacts far and wide. This virus affects both the economic and health sectors of all nations. It also freezes the day-to-day life in the world. In this paper, predictive and classification models for COVID-19 in Indonesia using machine learning, deep learning, and genetic algorithm techniques are developed. In predictive models, we used machine learning algorithms like Logistic Regression, Support Vector Machine, and Recurrent Neural Network. Among the machine learning algorithms, feature scaling with Logistic Regression (94.75%) performed well. Decision Tree Classifier and Genetic algorithms are used to classify the clusters like green, yellow, orange, and red conforming to the new death of each province in which Genetic Algorithm gave better accuracy (93%).

Keywords COVID-19 · Mortality · Death prediction · Machine learning · Neural networks · Genetic algorithm

1 Introduction

Toward the end of 2019, Corona infection flare-ups became exposed. The first Corona case was reported in Wuhan, China. China informed World Health Organization (WHO) some pneumonia instances of an obscure reason are accounted for in Wuhan. Step by step the illness spread to more cities of China. Later on, it spread all over the world. WHO proclaimed it as a pandemic situation. 8,71,61,708 confirmed and 18,82,643 deaths are reported all around the world till the first week of January. Indonesia 7,88,402 confirmed cases and 23,296 deaths were recorded. The first positive case of Indonesia was reported in the first week of March 2020. All nations were recommending various lockdowns to control this pandemic situation. Social distancing, wearing masks, and washing hands properly are guidelines given by

L. Lohi (✉) · M. L. Pai

Department of Computer Science & IT, Amrita School of Arts & Sciences, Kochi, India

Amrita Vishwa Vidyapeetham, Bengaluru, India

World Health Organization (WHO). WHO is working in collaboration with scientists, businesses, and global health organizations to overcome the pandemic situation.

Agbelusi and Olayemi [1] proposed a model for the prediction of the mortality rate of COVID-19 patients in Nigeria using Naïve Bayes, Decision Tree, and Multilayer Perceptron (MLP). Multilayer Perception is effective and reliable and is recommended to forecast the rate of mortality of patients infected with coronavirus [1]. Arora et al. [2], predicted COVID-19 positive cases using Long Short-Term Memory (LSTM). In which Adam Optimizer and Relu were used as activation functions. In view of prediction errors, bi-directional LSTM gives the best outcomes [2]. Shastri et al. [3] developed a model for forecasting COVID-19 cases using deep learning models for the time series data. In this study, they have compared the cases of two nations namely India and USA. They have observed that Convolutional LSTM (Conv-LSTM) is the best model with high accuracy [3]. Kumar [4] clustered COVID infections in India using hierarchical cluster analysis [4]. Robu and Holban [5] used different aspects of genetic algorithms and data mining for classification [5]. Khakharia et al. [6] have discussed various techniques for the outbreak prediction of COVID-19 in their literature. Auto-Regressive Moving Average (ARMA) gave better performance with accuracy 99.93% [6]. Hazarika and Gupta [7] proposed a model for modeling and forecasting using Wavelet-Coupled Random Vector Functional Link (WCRVFL) networks for COVID-19 spread, WCRVFL forecast COVID-19 spread in proper way [7]. Lalmuanawma et al. [8] reviewed various details related to the application of Machine Learning and Artificial Intelligence. In their study, they are focusing treatment, medicines, prediction and forecasting process for COVID-19 pandemic situation [8]. Dhamodharavadhani et al. [9] paper intends to investigate appropriate Statistical Neural Network (SNN) models and their crossover form for COVID-19 death rate forecast in India [9]. Kavadi et al. [10] proposed comparisons of two different models using partial derivative regression and nonlinear machine [10]. James and Menzies [11] analyze the evolution of multivariate time series using a cluster-based method [11]. Carrillo-Larco and Castillo-Cara [12] used an unsupervised machine learning algorithm (k-means) to classify data-driven clusters of different countries [12]. Chakraborty and Ghosh [13] proposed a novel hybrid Auto-Regressive Integrated Moving Average–Wavelet Based Forecasting (ARIMA-WBF) model for forecasting and also used an optimal regression tree to identifying the fatality rate [13]. Djalante et al. [14], assessed the current reactions to COVID-19 in Indonesia for the time of January to March 2020 [14]. Wang et al. [15] proposed a hybrid prediction model dependent on Logistic and Prophet to predict the pandemic pattern of COVID-19 worldwide [15]. Farooq and Bazaz [16] proposed an Artificial Neural Network (ANN)-based adaptive steady learning procedure for model parameter learning and model updatations [16]. Kelly and Davis [17] describe a method for hybrid Genetic Algorithms and k-nearest neighbors for Classification [17]. Sujath et al. [18] proposed a machine learning forecasting model using Logistic Regression(LR), Multilayer Perceptron (MLP), and Vector Auto Regression(VAR) for the COVID-19 pandemic in India. They found MLP gives high performance with 95% of accuracy [18]. Michelozzi et al. [19] discussed the mortality effects of the COVID infection flare-up by sex and age in Italy, from 1 February to 18 April 2020 [19].

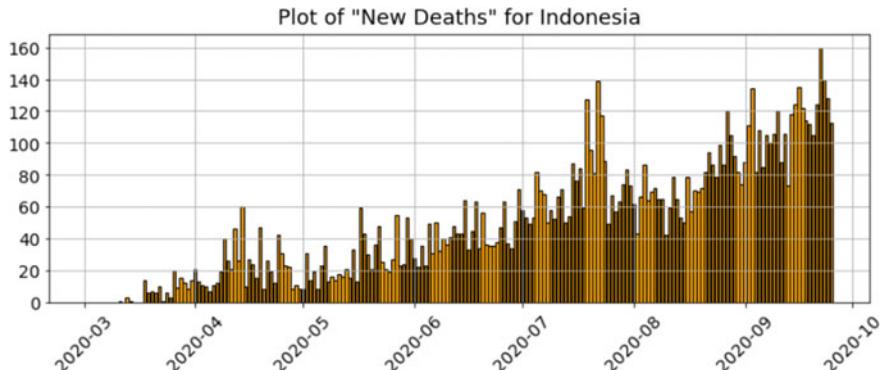


Fig. 1 New deaths in Indonesia

2 Data

The COVID-19 Indonesian time series dataset of sample points 6537 and 37 location (latitude and longitude)- based attributes were accessed from Kaggle.

We obtained the raw data from Kaggle preprocessed using the following steps.

- (1) Ignore the Location parameters to avoid redundancy and only latitude and longitude are considered.
- (2) Avoid all the less relational attributes from the dataset by taking univariate feature selection techniques.

We used univariate feature selection with ANOVA f-test, attributes with F-Score less than 1 dropped. According to date factor, total death, active cases, new deaths, new recovered, etc., are the main parameters. After preprocessing, 22 attributes were used for further work. We used Python 3.0 and open source libraries like Numpy, Pandas, Keras, and Tensorflow.

Exploratory Data Analysis (EDA) was done to overview data. Figures 1 and 2 represent the new deaths and distribution plot for total deaths of Indonesia, respectively.

Figures 3 and 4 depict the weekly increase in the number of confirm and death cases.

Figure 5 plots the top 15 provinces as per the death cases reported in Indonesia. Using the active case tracker function, track the active cases based on location and number of days.

3 Methodology I

To find the best model for this dataset, first convert the dataset values to equal normally distributed values using skewness.

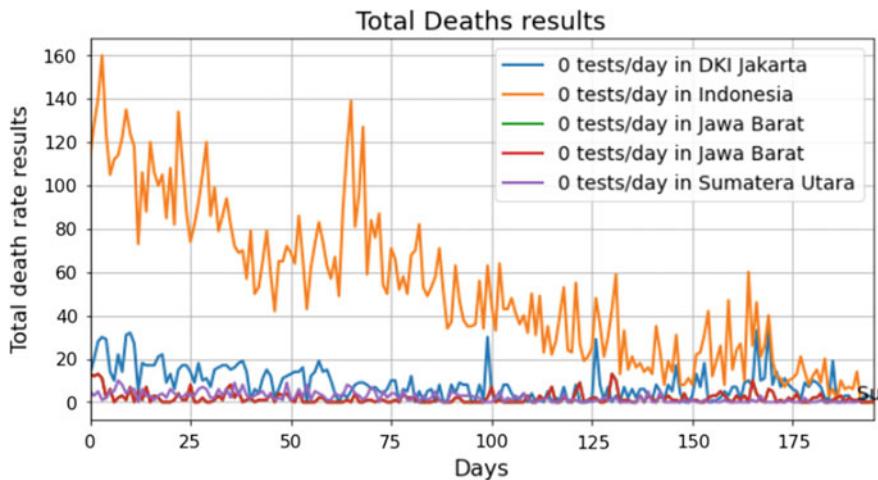
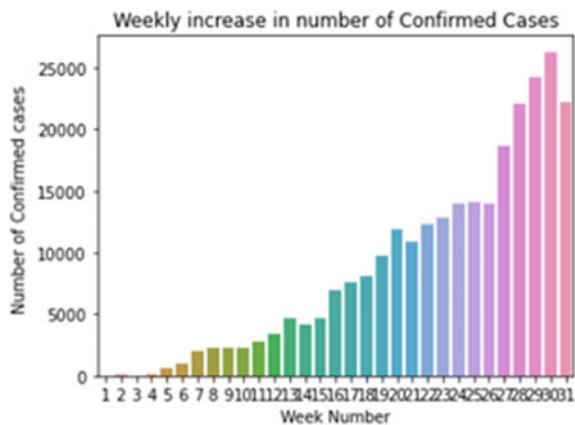


Fig. 2 Distribution plot for total death results

Fig. 3 Weekly increase in the number of confirm cases



$$\text{Skewness} = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}} \quad (1)$$

In this study, we have compared Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree, and Random Forest using GridSearchCV, one of the hyper parameter tuning method. Logistic regression is highly suited for our project, based on the continuous behavior of our dataset. In the coding of LR, ‘C’ (1,5,10) as an Inverse regularization parameter which is inversely positioned to the Lambda regulator. $C = 1/\lambda$. Here, Gini and entropy criterion were used for splitting the nodes of a decision tree. And we tried values 5 and 10 for max-depth. In Random forest, Gini criterion and n_estimators (10, 15, 20, 50, 100, 200) are used to decide the trees

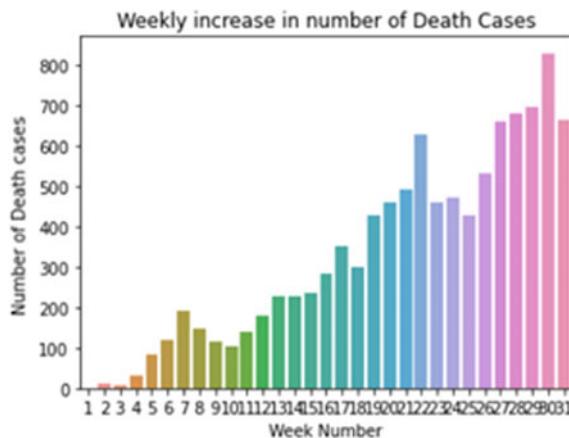


Fig. 4 Weekly increase in the number of death cases

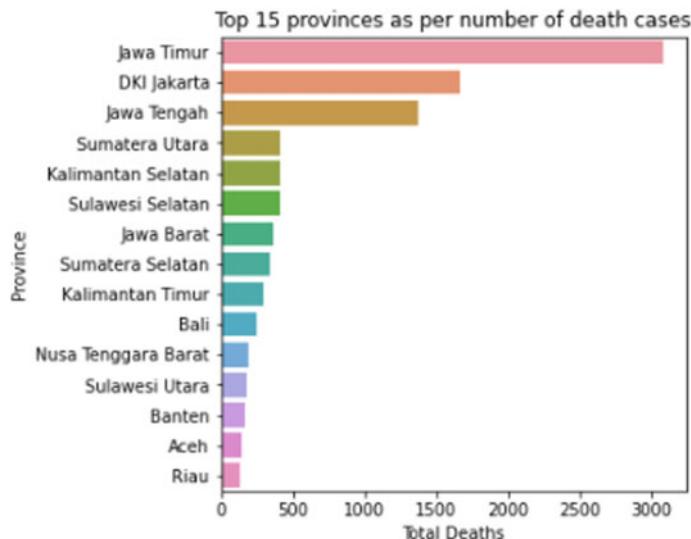


Fig. 5 Top 15 provinces as per the number of death cases

for making decisions. Gaussian Radial Basis Function (RBF) and the linear kernel are used in SVM. And we tried 1, 10 and 20 for C.

Table 1 represents the result of best model calculation. In this study, we have observed that the best model is decision tree (88.87%) which is shown in bold in Table 1.

The dataset is divided into two subsets namely: training and test set (70:30), respectively. We develop predictive models (Fig. 6) using Logistic Regression (LR), Support Vector Machine (SVM), and Long short-term memory (LSTM) algorithms.

Table 1 Result of best model calculation

Model	Best parameter	Score
Logistic regression	{‘C’: 10}	0.795630
Decision tree	{‘criterion’: ‘entropy’, ‘max_depth’: 10}	0.888711
Random forest	{‘n_estimators’: 200}	0.837145
Support vector machine	{‘C’: 20, ‘kernel’: ‘rbf’}	0.838310

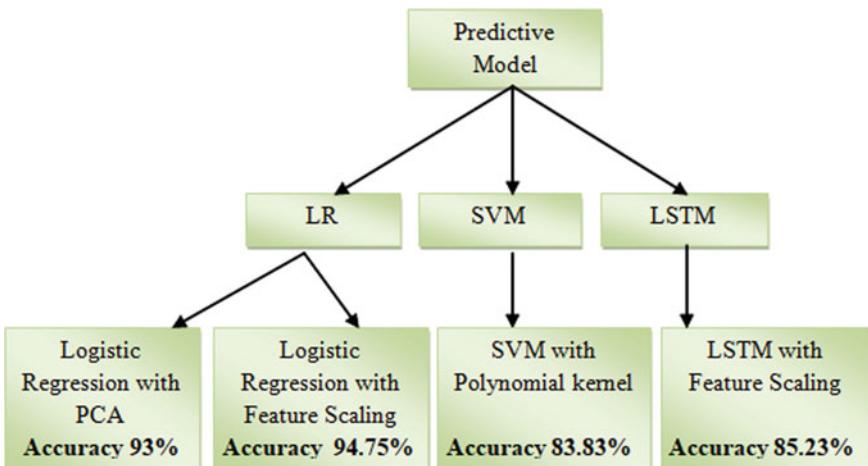


Fig. 6 Flow diagram of predictive model

3.1 Logistic Regression

The LR model predicts the probability of the real-valued esteemed input having a place with the default (class 0). If the probability is >0.5 we can accept the yield as a forecast for the default (class 0), in any case, the expectation is for the other (class 1). If an event is occurring its probability (between 0 and 1) is Y otherwise 1-Y.

According to Fig. 6, we apply two variations of logistic regression. First, we applied Principal Component Analysis (PCA) to reduce the dimensionality of our dataset. We got 66% of accuracy after PCA. If we lift the accuracy using the bagging classifier, the outcome gave 93% accuracy.

Next, we apply feature scaling for data normalization. The Standard Scaler takes your data regularly distributed inside each feature and will scale them with the end goal that the appropriation is presently based on 0, with a standard deviation of 1. Feature scaling is applied to independent features for information preprocessing. It assists with normalizing the data inside a specific reach. Now and again, it moreover helps in accelerating the algorithm in a calculation. Sklearn preprocessing package is used for this purpose. The use of feature scaling weighs all the features equally.

Dataset for this study is not linearly separable, so we have used sigmoid function as the activation function. The output of the LR is passed through a sigmoid (logit) function that can map value between 0 and 1. Use the standard scalar function and fit the training dataset and testing dataset. For this dataset, the logistic regression has three coefficients simply like linear regression, for instance, output

$$Y = b_0 + b_1x_1 + b_2x_2 \quad (2)$$

$$p = \frac{1}{1 + e^{-y}} \quad (3)$$

After applying sigmoid function, Eq. (3) and Eq. (2) will reduce to Eq. (4)

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1x_1 + b_2x_2 \quad (4)$$

In LR with PCA and LR with feature scaling, better performance is produced by LR with feature scaling with accuracy 94.75%.

3.2 Support Vector Machine

The system is exposed to train the Polynomial Kernel Support Vector Regression (SVR) model on the training set. SVM algorithms use a set of mathematical functions, define as a kernel. The general equation of Kernel is given below.

$$K(\bar{x}) = \begin{cases} 1 & \text{if } \|\bar{x}\| \leq 1 \\ 0 & \text{Otherwise} \end{cases} \quad (5)$$

If we consider that our datapoint taken from our dataset is a and b, then the equation for this polynomial kernel should be written like this

$$k(X_1, X_2) = (a + X_1^T \cdot X_2)^b \quad (6)$$

where b is degree of kernel and a is the constant term. If first datapoint is (a_1, a_2) and second datapoint is (b_1, b_2) , then consider a variable X_a and set the first data point to that variable and place the second data point to X_b .

$$X_a = (a_1, a_2) \quad (7)$$

$$X_b = (b_1, b_2) \quad (8)$$

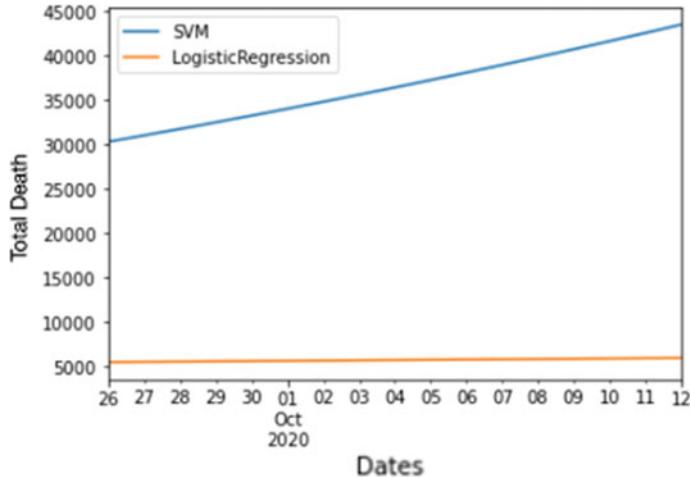


Fig. 7 Predicted death plot by LR and SVM

Presently on the off chance that we need to plan our data into higher dimension suppose in Z space which is 6-dimensional it might appear as

$$Z_a = \emptyset(X_a) = (1, a_1, a_2, a_1^2, a_2^2, a_1 * a_2) \quad (9)$$

$$Z_b = \emptyset(X_b) = (1, b_1, b_2, b_1^2, b_2^2, b_1 * b_2) \quad (10)$$

To solve SVM we want to find $Z_a^T * Z_b$. Traditional and kernel trick methods are used for this purpose. Here, we adopt kernel trick method, then applying polynomial kernel it should be considered as a degree of values, i.e., calculated on the basis of degree and $Z_a^T * Z_b$ combinations of degree polynomial values and data points.

$$Z_a^T Z_b = k(X_a, X_b) = (1 + X_a^T X_b)^2 \quad (11)$$

$$Z_a^T Z_b = 1 + a_1 b_1 + a_2 b_2 + a_1^2 b_1^2 + a_2^2 b_2^2 + a_1 b_1 a_2 b_2 \quad (12)$$

Figures 7 and 8 represent predicted death plot and predicted deaths by LR and SVM.

3.3 Long Short-Term Memory

LSTM is a variation of a Recurrent Neural Network (RNN) derived from feedforward networks. RNN exhibits similar behavior to how human brains function. LSTM

Fig. 8 Predicted deaths by LR and SVM

	Dates	LogisticRegression	SVM
0	2020-09-26	5444	30313
1	2020-09-27	5473	31028
2	2020-09-28	5502	31757
3	2020-09-29	5532	32500
4	2020-09-30	5561	33257
5	2020-10-01	5590	34028
6	2020-10-02	5620	34814
7	2020-10-03	5649	35615
8	2020-10-04	5679	36430
9	2020-10-05	5708	37261

replaces the hidden layers of RNN with memory cells. It is also capable to learn long-term dependencies [3]. The equation of the recurrent neural network is

$$h_t = f(h_{t-1}, X_t) \quad (13)$$

h_t is the new state, h_{t-1} is the previous state while X_t is the current input.

We input a sequence of input attributes and train the model to predict the death rate. Figure 9 represents new death rate graph based on the date.

The input will be converted to an array of values like a matrix before being passed into an LSTM layer. Create features and labels from sequences. Build LSTM models with Embedding, LSTM, and Dense layers. MinMaxScaler is used to transform features by scaling each feature to a given range. Dropout regularization method is used to reducing the over fitting. We are using the Keras Sequential Application

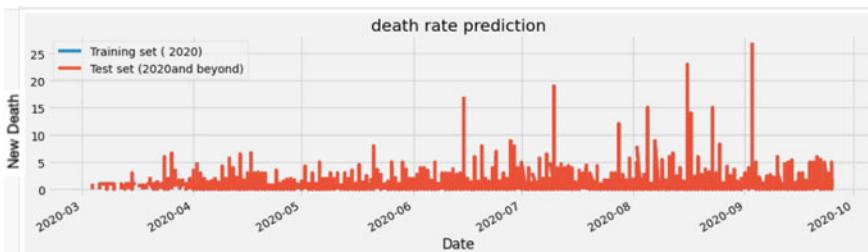


Fig. 9 New death rate graph based on date

Programming Interface (API) which means we build the network up to one layer at a time. It is possible to access the hidden state output for each input time step. A classic LSTM cell contains two sigmoid functions and one hyperbolic tangent (\tanh) function. A fully connected dense output layer produces a probability for learning the death rate from the label using these activation functions. Dense works according to the following operation: $\text{output} = \text{activation}(\text{dot}(\text{input}, \text{kernel}) + \text{bias})$, where activation is an activation function passed as the activation argument.

As a trial and error method, we started with 1 layer and 500 neurons. Then, we repeated the training for four layers. After inputting same number of neurons, we tried applying different numbers of neurons (500, 2000, 4000, and 500) in each layer. Then, we found that the system produces same accuracy in 4 layers with 500 neurons and different neurons. But the minute difference is shown in the train accuracy as shown in Fig. 10. This process is repeated with 7 layers for 500, 2000, 4000, and 500 neurons. We observed the accuracy for each case, and we found a slight decrease in the accuracy after the fourth layer. Thus, we can say that the fourth layer gives the best accuracy among the whole process of LSTM. To get better performance, Adam Optimizer is used which increases computational efficiency. Figures 10 and 11 plot the train and test accuracy of four hidden layers with 500 neurons.

Fig. 10 Train accuracy of 4 hidden layers with 500 neurons

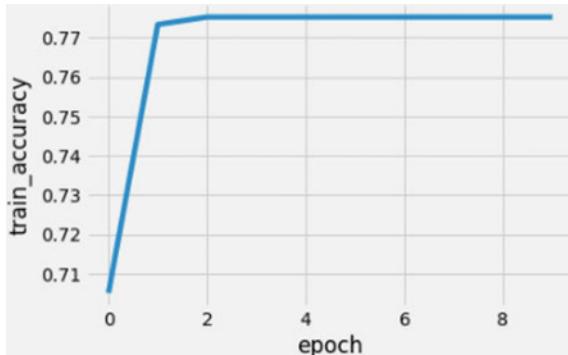
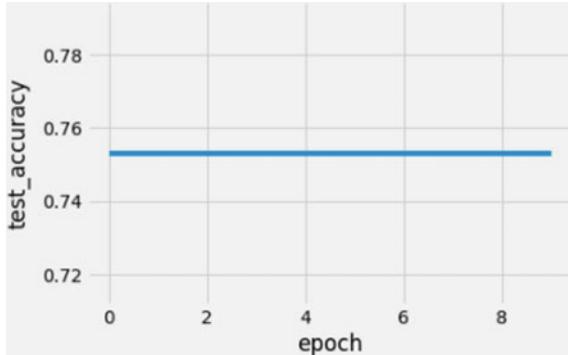


Fig. 11 Test accuracy of 4 hidden layers with 500 neurons



Figures 12 and 13 plot the train and test accuracy of four hidden layers with different neurons.

Figures 14 and 15 represent epoch vs accuracy plot of train and test in four hidden layers with 500 and different neurons.

Fig. 12 Train accuracy of 4 hidden layers with different neurons

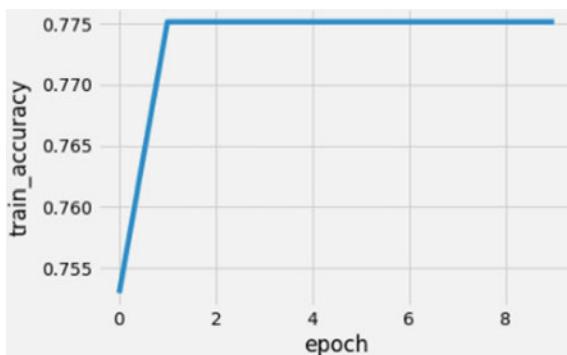


Fig. 13 Test accuracy of 4 hidden layers with different neurons

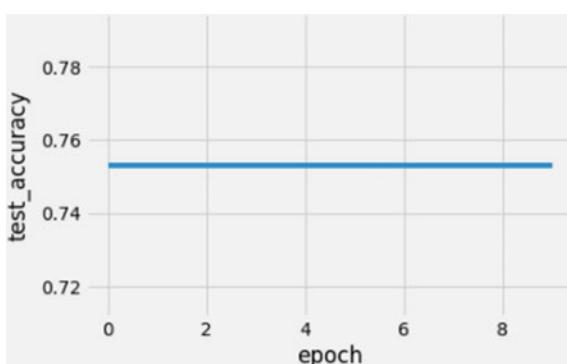


Fig. 14 Epoch versus accuracy plot of train and test in 4 hidden layers with 500 neurons

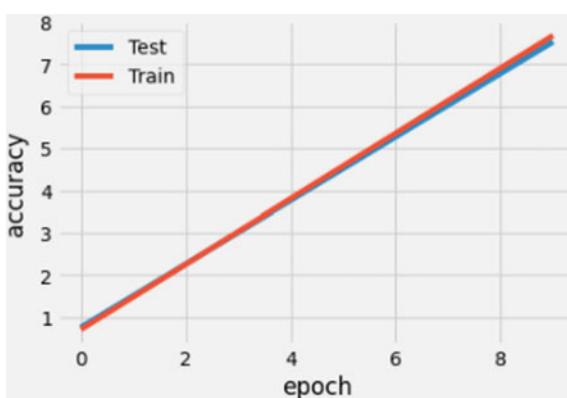


Fig. 15 Epoch versus accuracy plot of train and test in 4 hidden layers with different neurons

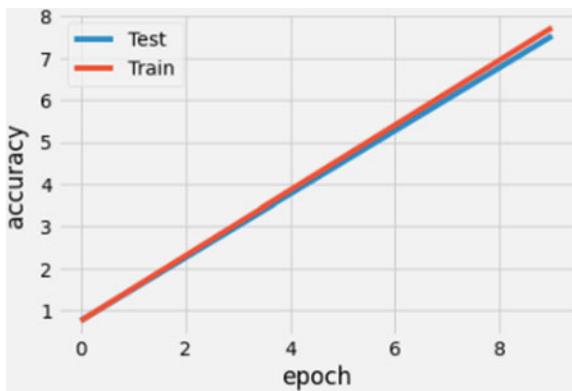
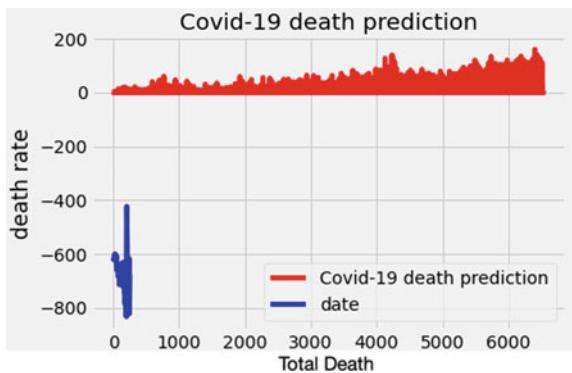


Fig. 16 Death prediction by LSTM

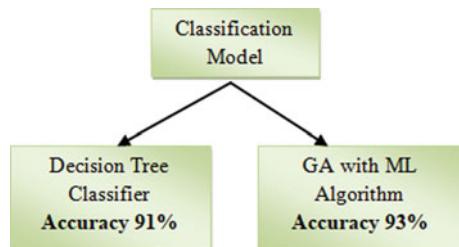


The model accuracy of four layers after implementation is 76.85%. Adaptive Boosting (AdaBoost) is used to increase the accuracy (85.23%) of this model. Figure 16 depicts the predicted death by LSTM.

4 Methodology II

In this methodology, we classify the clusters of each province per the new death. For this, we propose two models (Fig. 17) using the Decision Tree Classifier and Genetic Algorithm (GA). The province into each cluster according to the severity of the new death rate is classified.

Fig. 17 Flow diagram of classification model



4.1 Decision Tree Classifier

In the learning process, the decision tree learns from a collection of categorized training samples to create a decision tree. A decision tree combines several independent variables with direct or indirect relations to a target variable with a tree structure, created by repetitively partitioning the data into several classes. Decision tree classifier checks new death conditions of Indonesia for better prediction. The decision says that most of the new death rate of Indonesia is less than or equal to 100. Decision tree classifies the provinces into green, yellow, orange, and red regions according to the new death. We obtained an accuracy of 91% for decision tree classifier which is higher than 77% in Agbelusi and Olayemi [1] work. Figure 18 represents the final decision plot of weekly progress of confirmed and death cases.

The output of cluster classification using decision tree is given below.

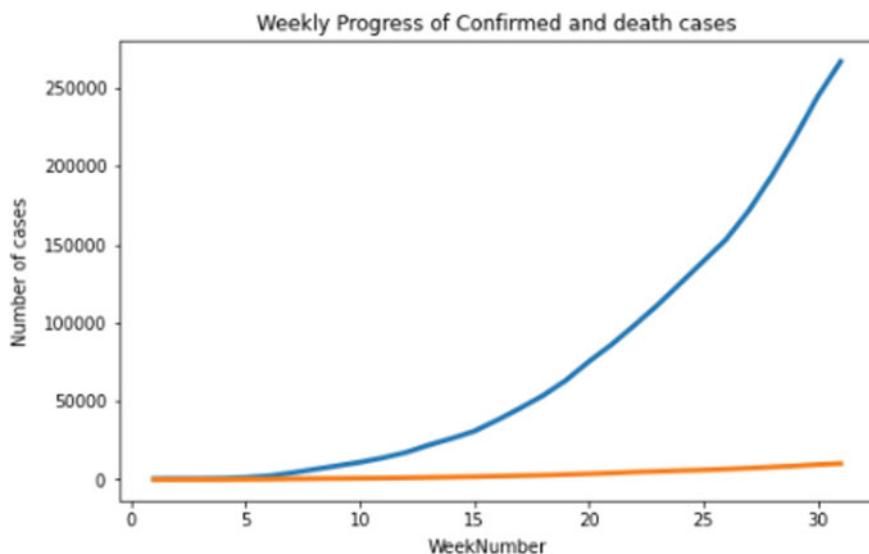


Fig. 18 Final decision plot—Weekly progress of confirmed and death cases

```

Enter the value of death rate: 20
Enter the date: 01/01/2021
enter Your location: Bali
2021-01-01 Bali The Death rate is increased and it will goes to Yellow alert
Please Use Mask and Sanitiser and be careful Thank you

```

4.2 Genetic Algorithm

Machine learning algorithm is used to build a genetic algorithm model. The best machine learning model for the Indonesian dataset is a decision tree classifier (Accuracy 91%). So we select decision tree classifier to develop this genetic algorithm.

In the Genetic Algorithm first, we start with the initial population, here we select parents from the population. Next find out the fitness score of the population. And go through selection, crossover and mutation operations on the parent to generate a new generation. At last, these new off-springs replace the existing generation in the population, and the process repeats. Genetic algorithms are based on a biological metaphor [20].

Using this model, we obtained the highest performance with the help of mutation and crossover techniques (Accuracy 93%). The genetic algorithm classifies the clusters with the highest performance concerning the new death of each province.

5 Results

Table 2 shows the accuracy of different models used in this study.

In this study, prediction and classification have been done. Applying machine learning algorithms we have observed that the best algorithm for COVID death prediction is feature scaling with logistic regression with 94.75% accuracy. which is shown in bold in Table 2. Initially, we started with polynomial kernel SVR and

Table 2 Accuracy of predictive and classification models

No.	Model	Accuracy (%)
Predictive model		
1	Logistic Regression	94.75
2	Long ShortTerm Memory	85.23
3	Support Vector Machine	83.83
Classification models		
1	Genetic Algorithm	93
2	Decision Tree	91

got an aggregate accuracy of 83.83%. As a next level, we have trained the LSTM with 4 hidden layers with the same and different number of neurons in which no accuracy change is observed in both the cases. This process is repeated up to 7 layers but slight decrease in accuracy is observed after the fourth layer. In LSTM model, fourth layer is considered as the model with best accuracy (76.85%). Model accuracy is increased using AdaBoost (85.23%). LSTM performed better than the SVR algorithm.

In the classification model, genetic algorithm and decision tree are considered. According to Table 1, the decision tree is the best model for this dataset which is used to classify the clusters. Decision tree classify the clusters with 91% accuracy according to the new death rate. The genetic algorithm performed better than decision tree. Decision tree algorithm is used to build a genetic algorithm model which in turn gives an accuracy of 93%.

6 Conclusion and Future Work

In this paper, we applied machine learning, deep learning, and genetic algorithms to predict and classify COVID-19 dataset for Indonesia. In the case of prediction, Logistic Regression performed well when compared to SVM and LSTM. When classification of clusters of COVID-19 dataset of Indonesia is considered, GA gives best results than decision tree. Therefore, LR and GA can be applied further for prediction and classification of COVID-19 cases for other regions and with other datasets.

In future, we can apply optimizers in GA for better output. Also other algorithms can be applied to check whether better performances are obtained.

References

1. Agbelusi, O., Olayemi, O.C.: Prediction of mortality rate of COVID-19 patients using machine learning techniques in nigeria. *Int. J. Comput. Sci. Softw. Eng.* **9**(5), 30–34 (2020)
2. Arora, P., Kumar, H., Panigrahi, B.K.: Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. *Chaos, Solitons Fractals* **139**, (2020)
3. Shastri, S., Singh, K., Kumar, S., Kour, P., Mansotra, V.: Time series forecasting of Covid-19 using deep learning models: India-USA comparative case study. *Chaos, Solitons Fractals* **140**, (2020)
4. Kumar, S.: Monitoring novel corona virus (COVID-19) infections in India by cluster analysis. *Ann. Data Sci.* **1** (2020)
5. Robu, A., Holban, S.: A genetic algorithm for classification. In: Proceedings of the 2011 International Conference on Computers and Computing (2011)
6. Khakharia, A., Shah, V., Jain, S., Shah, J., Tiwari, A., Daphal, P., Mehendale, N.: Outbreak prediction of COVID-19 for dense and populated countries using machine learning. *Ann. Data Sci.* 1–19 (2020)

7. Hazarika, B.B., Gupta, D.: Modelling and forecasting of COVID-19 spread using wavelet-coupled random vector functional link networks. *Appl. Soft Comput.* **96**, (2020)
8. Lalmuawma, S., Hussain, J., Chhakchhuak, L.: Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: a review. *Chaos, Solitons Fractals*, 110059 (2020)
9. Dhamodharavadhani, S., Rathipriya, R., Chatterjee, J. M.: Covid-19 mortality rate prediction for India using statistical neural network models. *Front. Public Health* **8** (2020)
10. Kavadi, D.P., Patan, R., Ramachandran, M., Gandomi, A.H.: Partial derivative nonlinear global pandemic machine learning prediction of covid 19. *Chaos, Solitons Fractals* **139**, (2020)
11. James, N., Menzies, M.: Cluster-based dual evolution for multivariate time series: analyzing COVID-19. *Chaos Interdiscip. J. Nonlinear Sci.* **30**(6), 061108 (2020)
12. Carrillo-Larco, R.M., Castillo-Cara, M.: Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: an unsupervised machine learning approach. *Wellcome Open Res.* **5**(56), 56 (2020)
13. Chakraborty, T., Ghosh, I.: Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: a data-driven analysis. *Chaos, Solitons Fractals* 109850 (2020)
14. Djalante, R., Lassa, J., Setiamarga, D., Mahfud, C., Sudjatma, A., Indrawan, M., Surtiari, I.G.A.: Review and analysis of current responses to COVID-19 in Indonesia: period of January to March 2020. *Prog. Disaster Sci.* 100091 (2020)
15. Wang, P., Zheng, X., Li, J., Zhu, B.: Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. *Chaos, Solitons Fractals* **139**, (2020)
16. Farooq, J., Bazaz, M.A.: A novel adaptive deep learning model of Covid-19 with focus on mortality reduction strategies. *Chaos, Solitons Fractals* **138**, (2020)
17. Kelly Jr, J.D., Davis, L.: A hybrid genetic algorithm for classification. In: IJCAI 1991, vol. 91, pp. 645–650 (1991)
18. Sujath, R., Chatterjee, J.M., Hassanien, A.E.: A machine learning forecasting model for COVID-19 pandemic in India. *Stoch. Environ. Res. Risk Assess.* 1 (2020)
19. Michelozzi, P., de'Donato, F., Scortichini, M., De Sario, M., Noccioli, F., Rossi, P., Davoli, M.: Mortality impacts of the coronavirus disease (COVID-19) outbreak by sex and age: rapid mortality surveillance system, Italy, 1 February to 18 April 2020. *Eurosurveillance* **25**(19), 2000620 (2020)
20. Luger, G.F.: Artificial Intelligence: Structures and Strategies for Complex Problem Solving, 2nd edn. Addison-Wesley, Boston (2002)

Taala Classification in Carnatic Music Using Machine Learning Algorithms and Deep Neural Networks



Amrutha Jayakumar and Maya L. Pai

Abstract In Carnatic music, taala plays a vital role. Taala classification in Carnatic music is an area in which very few works have been reported so far. For any audio sample, taala is the most unique factor. Thus, it is important to study the variations of taala in an audio to know which category it belongs to. In this paper, we aimed at classifying four major taalas namely Aadi, Rupaka, MishraChapu and KhandaChapu in Carnatic music by using various machine learning algorithms and Deep Neural Networks (DNNs). We found out the correlation between the attributes and identified the best attribute that contribute well to the classification of taala from the dataset. We implemented the algorithms such as Random forest, K- Nearest Neighbour, Principal Component Analysis (PCA) with Logistic Regression (LR) and Support Vector Machine (SVM) with kernel Gaussian Radial Basis Function(RBF) and observed that PCA with LR and SVM with RBF performed the best classification with 98.11% and 98.1% accuracies, respectively. Also, we worked with DNN incorporating six hidden layers and resulted in a better performance of 97.50% to classify the taalas.

Keywords Taala · Principal component analysis · Deep neural networks · Logistic regression

1 Introduction

Taala in Carnatic music is a proportion of time-frame or in any case a musicality cycle. As heart beat offers life to a man, taala is the life of a show. It is said ‘Sruthi Maatha Laya Pitha’, which implies that the tone originated from the Tampura is mother to the music and taala resembles father. Taala is a progression of checks made by wave of hand or tap of the hand on the lap or by utilizing both the hands taking after an applaud. All the more profoundly, taala is the cadence system in Carnatic music and gives a wide structure to interpretation and redundancy of melodic and musical expressions, thought processes and impromptu creations. It comprises fixed

A. Jayakumar (✉) · M. L. Pai

Department of Computer Science & IT, Amrita School of Arts & Sciences, Kochi, Amrita Vishwa Vidyapeetham, India

length time cycles called Avartana. An Avartana of a taala is partitioned into essential time units called Akṣaras. The Akṣaras are assembled into longer time units called beats, which are outwardly shown through hand motions. The sub-beat structure (which chooses the quantity of Akṣaras contained per beat) is called nade. The main Akṣara beat of an Avartana is known as the Sama. Most Carnatic creations are set to a particular taala and start on the Sama. Nonetheless, the creations can likewise begin after (or previously) the cycle starts, and this balance of the beginning of the structure comparative with the sama is called Edupu, and is regularly estimated in Akṣaras. An Avartana is likewise isolated into conceivably inconsistent segments called Anga, which characterize the structure of the taala. The Angas characterize the visual signals that used to show the movement through the taala.

Renteria and Llano [1] considered the relationship between the features of the song and their genre to provide data mining tools for genre and sub-genre classification. They have used Random forest classifier, fully connected neural networks and also Logistic regression for the classification. They obtained a better accuracy for the random forest than that of fully connecter neural networks, but still contained overfitting [1]. Jacoby and McDermott [2] focused on the reproduction of random temporal sequences by producing random seed rhythms. They efficiently mapped out the stability of such reproducible patterns of rhythms [2]. Rao et al. [3] used machine learning methods to obtain the classification of phrases on manually processed audio. They have used Dynamic time warping and classification based on HMM for their work [3]. Anglade et al. [4] propose an automatic technique using frequent sequences of code on audio data. They used representation of first-order logic and used 2 way and 3 way classification which retained an accuracy less than 79% and 72%, respectively [4]. Kan et al. [5] compared the machine learning approaches to classify taala. They used CNN and K-NN to depict both the statistical and deep learning approaches to the classification. As a result they obtained 92% accuracy for 2-dimensional CNN as the best prediction method for their work [5]. Morris [6] describes in their paper the various theoretical tools taken from Western music, which are capable to classify taalas in Carnatic music. They tried to categorize the rhythmic structures into different classes based on their similarity and equivalence. They have also used isomorphism to the beat classes inorder to depict the time point classes as a better interpretation of unordered pitch class set [6]. Srinivasamurthy and Serra [7] focused on tracking metrical cycles and the hierarchical metrical structure of Carnatic music. They targeted on akshara pulse period, aksharas and samas of Carnatic music dataset [7]. Nitha and Suraj [8] formulated an algorithm to detect taala in Carnatic music to make it useful in music therapy applications. They made use of features of the patterns of the music signals. They used beat detection algorithm, sound energy algorithm and frequency selected sound algorithm to perform the classification. They mainly focused on the low-level features for the identification of taala. They have also trained an artificial neural network for better results [8]. Heshi et al. [9] made an effort to analyze the rhythm and features of timbre in order to identify raga and also taala from a selected audio of Carnatic music. The features like flux, spectral roll-off, entropy, etc. were used as features of timbre while patterns and histograms of rhythm were used as features of rhythm. They made use of T test and Gaussian Mixture Model

(GMM) for further classification [9]. Srinivasamurthy et al. [10] formulated a beat detection method to describe the rhythm in Indian classical music. They generated a matrix of beat similarity and histogram of inter onset interval and then proposed an algorithm based on it that can easily extract the main and sub-beat structures of a musical piece. The algorithm was tested on manually annotated music dataset as well as on an Indian light classical music dataset and compared the results [10]. Pulijala and Gangashetty [11] used audio thumbnailing in their paper to classify taala. They used SVM and trained CNN RNN for the classification. They utilized different built in packages and libraries of Python to complete the classification. They also calculated the time taken to compute a thumb nail for an average recording audio of about six minutes [11]. Ramirez and Flores [12] have discussed various techniques for the classification of music. They focused on Decision trees, Support vector machines, CNN, FCNN, RNN, etc. and compared their observations of each method. They tried to present a survey on these models so that the reader gets a clear idea about the challenges of this area, recent datasets to work on and the opportunities in the music classification field [12]. Tian et al. [13] proposed a music recommendation system which is based on Logistic Regression and XGBoosting. They have insisted on the integration of Logistic regression and XGBoosting in their paper and thereby built a hybrid LX recommendation algorithm. They obtained a very good result in their work [13]. Randall and Greenberg [14] used deep networks like ImageNet model for the study of features from the audio spectrograms. They observed that the networks with the separation of input sources give the best result [14]. Bahuleyan H. [15] have compared two models CNN and a model that utilizes frequency and time factors of music. Among the classifiers used, they got better result for an ensemble classifier [15]. Vishnupriya et al. [16] have used a CNN for training the model and they have selected MFCC as their feature. They obtained an accuracy of 76% for the model [16]. Al Mamun et al. [17] have used six different genres of Bangla Music and they have considered the factors such as time and frequency of the audios. Then they have used a deep learning approach for the classification [17]. Chillara et al. [18] have introduced many methods in their work. They have considered both the spectrogram and feature based methods to classify the audio samples. And they got better result for CNN [18]. Steinmetz et al. [19] have used a different style of approach that make use of Temporal CNN that give an efficient result in their work. When compared to other methods, their approach gave better result [19].

2 Dataset

In this study, we have used CompMusic Carnatic Rhythm dataset [20], a collection of rhythm related analysis for automatic rhythm classifications and interpretation tasks in Carnatic Music which is shown in Table 1. The gathering consists of different length audio excerpts from the Comp Music Carnatic analysis, which is processed manually which indicates the development of sequences of taala, and also the related information. Since no rhythm annotated Carnatic audio dataset existed, we accessed a data set that consists of 176 excerpts of Carnatic music sampled from the Comp

Table 1 Description of the CompMusic Carnatic Rhythm dataset [20]

Taala	Beats, Akshara	Pieces	Total minutes in hours	Median length of a piece (in min)	Annotated beats	Samas	Annotated Positions in Taala Cycle
Aadi	8, 32	50	252.78 (4.21)	4.85	22793	2882	1, 2, 3, 4, 5, 6, 7, 8
Rupaka	3, 12	50	267.45 (4.45)	4.62	22668	7582	1, 2, 3
MishraChapu	7, 14	48	342.13 (5.7)	6.59	31055	7795	1, 2, 3, 4, 5, 6, 7
KhandaChapu	5, 10	28	134.62 (2.24)	4.41	13111	4387	1, 2, 3, 4, 5
Total		176	996.98 (16.62)	5.06	89627	22646	

Music assortment with 14 attributes. The items span the four preferred taalas in Carnatic music within which a majority of item's area units composed.

3 Data Preprocessing

The initial preprocessing of data is done through Exploratory Data Analysis. The dataset is divided into labels and features. We checked for the missing values, and found no null values in our data. After observing the shape and statistical summary of the dataset, the attribute Unique Identifier(UID) is taken out as it will return the unique identity of each values of dataset. Among the 14 attributes, we dropped out the irrelevant attributes such as 'MBID of the recording', 'Name', 'Artist', 'Release + Volume', 'Lead Instrument Code' and 'Raaga' as they contribute nothing to the process. Thus ,we have used the resultant 8 attributes for further work. For the ease of computation, the categorical values are converted into numerical values and we counted the number of taalas in our dataset. Table 2 below shows the each of the taala, their numerical value associated and the number of samples of each taala in the dataset.

Table 2 Taalas, their numerical value and count of samples

Name of Taala	Numerical value	No. of samples in dataset
Aadi	0	50
Rupaka	1	50
MishraChapu	2	48
KhandaChapu	3	28

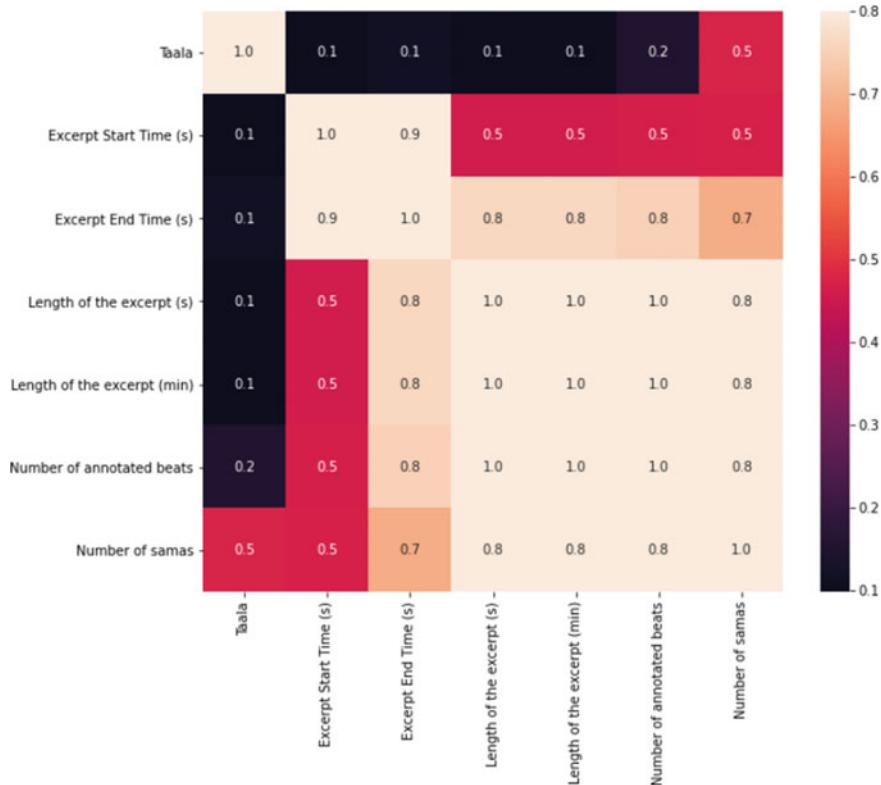


Fig. 1 Heat map generated to depict the correlation between the attributes

Inorder to establish the correlation between the attributes in our dataset, we used Pearson's Correlation Coefficient. The heatmap generated for the correlation between the attributes is depicted in Fig. 1.

Among the attributes, each of them show relationship between themselves. The attribute ‘Number of samas’ shows very good relationship with all others and that attribute contribute much to the classification process. Figure 2 represents the relation between the number of samas and taala, in order to plot the peakness of the data. Also we checked the relation between each taala with the number of samas. Figure 3 depicts the peaks in each of the taala with the attribute number of samas.

The data analysis as well as the further computations were performed using jupyter notebook with Python 3.9. The open source libraries such as Numpy, Pandas, Matplotlib and Keras were made use of.

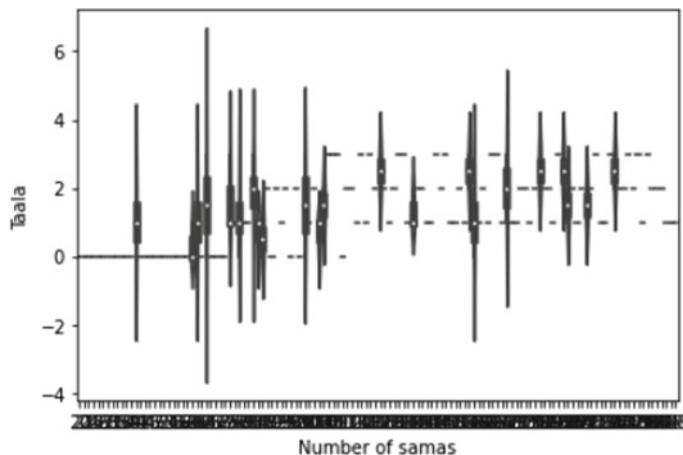


Fig. 2 Relation between the attributes taala and number of samas

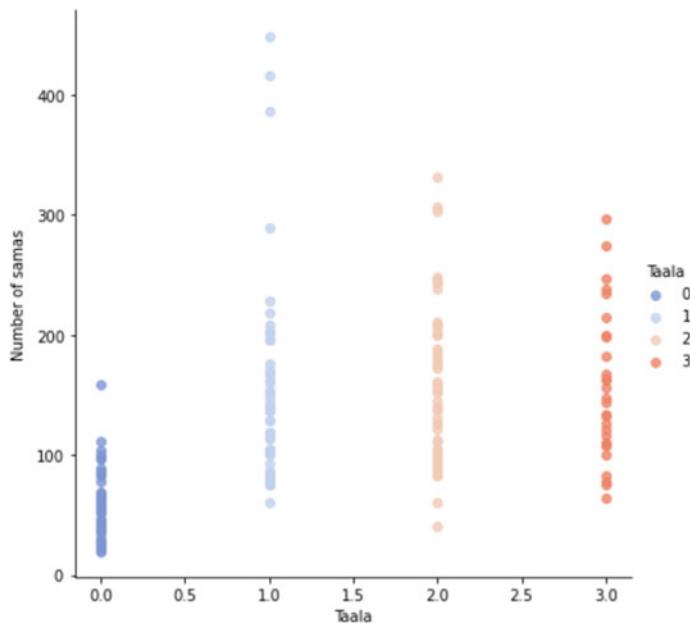


Fig. 3 Relation of the attribute ‘number of samas’ with each type of taala

4 Proposed Methodology

The primary philosophy of the methodology followed in this paper is the classification of Taalas in Carnatic music. Among the different Taalas, this paper focuses

on the classification of the four popular Taalas-Aadi, Rupaka, Khandachappu and Misrachappu. The data is divided as 70% training set and 30% testing set. The preprocessed data is fed to the machine learning algorithms like Kernel SVM, Random Forest, K-Nearest Neighbour, PCA and Logistic Regression and observed the accuracy. Later the system is fed into Deep neural networks with six hidden layers to obtain better accuracy. Following a trial and error method we trained 8 layers. We checked the accuracy of layers with 50, 100, 500 and 1000 neurons for each layer from 1 to 6 and we also applied different number of neurons in each layer for 4–8 layers. We found that the sixth layer shows the best accuracy than others and further layers show a stable accuracy. We then observed that SVM with kernel RBF and PCA with Logistic Regression functions well for the classification among the machine learning algorithms. In this paper, we have used four machine learning algorithms to classify the Taalas.

4.1 Random Forest

Random forest algorithm is used to build decision trees based on data and then obtain the prediction from each of the trees constructed and then select the better solution by a process called voting. It is one of the ensemble method which is quite better than a single tree as it decreases the chance of overfitting and it finds averages of the result. After splitting the data into training and testing phase, Feature scaling is performed by StandardScaler. The StandardScaler assumes that the data is distributed normally and it scales the data such that the distribution is centred around the value 0, with a standard deviation of value 1. We have used a package `sklearn` and imported the function as `sklearn.preprocessing`. The standardization replaces the values by their Z scores and feature scaling helps to weigh all the features equally.

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}} \quad (1)$$

The creation of `estimators_list` of Decision Tree Classifier was followed by the collection of fitted sub-estimators. We checked the functionality of Random Forest algorithm for 10, 20 and 30 number of estimators, respectively. When `n_estimators` = 10, the accuracy was 88.67%. When `n_estimators` = 20, the accuracy was 92.45% and similarly when `n_estimators` = 30 and above the accuracy stood same as 92.45%.

An ensemble classifier uses different decision trees. Initially, we start with the selection of random samples from our dataset. Then, the algorithm starts constructing a decision tree for each sample. After building trees, it predicts the result from every decision tree. Then, voting is performed for every predicted result. Finally, the result with more number of votes was selected as the final result of the prediction. We got 92.45% accuracy for this algorithm.

Figure 4 shows the resultant data with respect to UID and Fig. 5 shows the result with respect to Taala

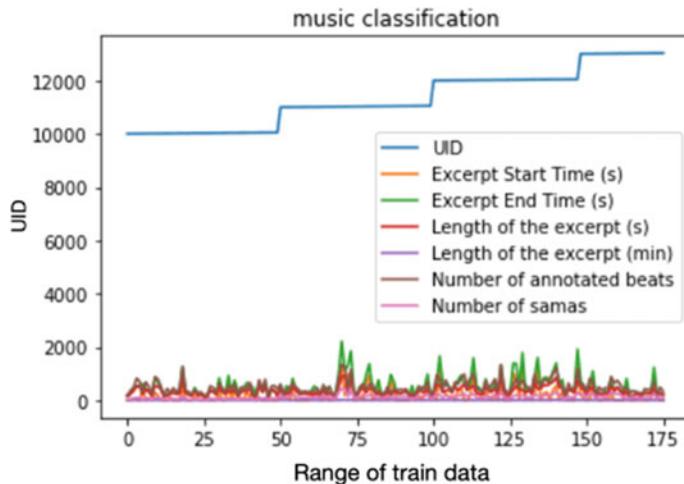


Fig. 4 Result based on UID

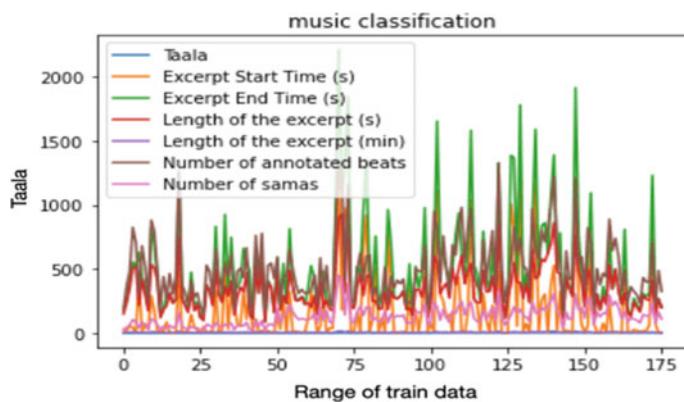


Fig. 5 Result based on Taala

4.2 Support Vector Machine

We used Kernel SVM using the RBF function in this paper. The function of kernel is to receive the inputs and to convert it into the intended form. Different SVM algorithms use different Kernel functions. Here, we used the RBF function as it has localized and finite response throughout the entire x-axis. The general equation of a Kernel function is as follows:

$$K(\bar{x}) = \begin{cases} 1 & \text{if } \|\bar{x}\| \leq 1 \\ 0 & \text{Otherwise} \end{cases} \quad (2)$$

The value of this function returns a 1 if it is centred around the origin or otherwise it gives a 0.

In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various learning algorithms which make use of Kernels. In particular, it is commonly used in SVM classification. RBF is a general-purpose kernel which is used when we have no previous knowledge about the data. The equation used is

$$k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2); \gamma > 0 \quad (3)$$

where $\|x_i - x_j\|$ is the Euclidean distance between the points x_i and x_j and γ is used only for RBF kernel. We obtained 98.1% accuracy for SVM model with kernel RBF.

4.3 K-Nearest Neighbour (K-NN)

The K-NN algorithm assumes that the objects exhibiting similarity in their nature appears close to each other. In this paper, we started with selection of number of neighbours, the value for K. Here, we have chosen the value $k = 5$ as we intend to check the 5 nearby neighbours before letting the new point to a specific category. Then, we calculated the Euclidean distance of the neighbours and took the K-nearest neighbours based on the calculated Euclidean distance. We counted the number of data points in each category and assigned the new data point into a category in which there are maximum number of neighbours. The equation to find Euclidean distance is

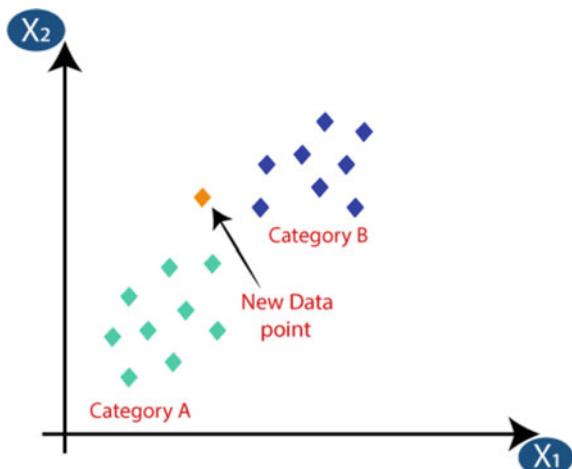
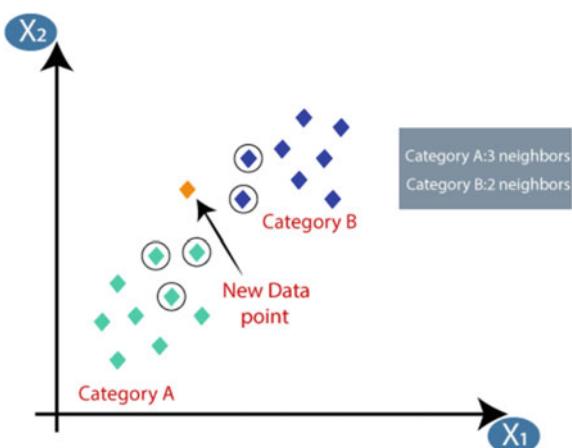
$$d(x, y) = \sqrt{\sum_{i=1}^m (X_i - Y_i)^2} \quad (4)$$

By calculating the Euclidean distance, we got the nearest neighbours, as three nearest neighbours in category A and two nearest neighbours in category B as shown in the Figs. 6 and 7. We got 90.56% accuracy by using K-NN.

4.4 PCA and Logistic Regression

In this paper, we have done Logistic Regression with Principal Component Analysis, or PCA. Mathematically, the standardization process in PCA can be done by subtracting the mean and dividing by each value of each variable. The general formula to perform standardization is given in the Eq. (1).

The data has been normalized using StandardScaler. When the normalization is done, all the factors will be changed to a similar scale. We have used

Fig. 6 The new data point**Fig. 7** Categories of classification

`sklearn.decomposition` package to import PCA and `sklearn.linear_model` to import Logistic Regression. The L R model accepts the real-valued inputs and predicts so as to the probability of the input be included in the class 0 (default class). If the probability is greater than 0.5 then we can consider the output as a prediction for the class 0, otherwise the prediction is for the other class (class 1). For this dataset, the logistic regression has three coefficients just like linear regression:

$$\text{output} = b_0 + b_1x_1 + b_2x_2 \quad (5)$$

where b_0 , b_1 and b_2 are the trials for the prediction and x_1 , x_2 are the outputs obtained.

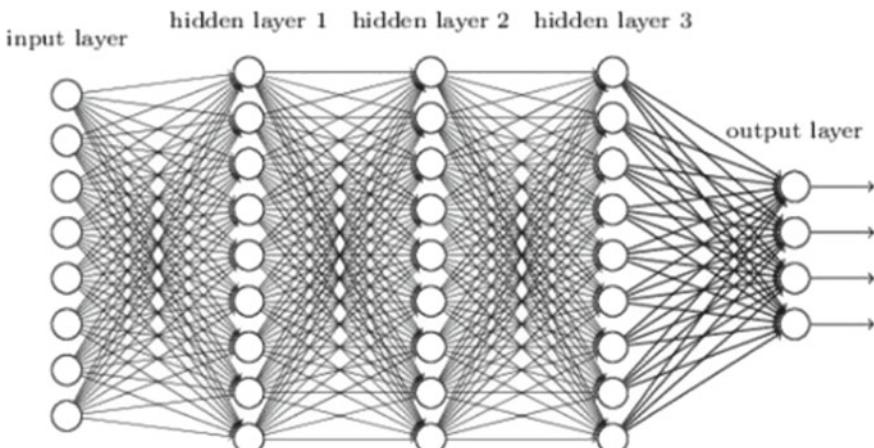


Fig. 8 Deep neural network

The chance for an occasion to happen is the division of times you hope to see that occasion in numerous preliminaries. If an event is occurring its probability will be Y otherwise $1-Y$. Probabilities always range between 0 and 1. We observed an accuracy of 98.11% for this model.

4.5 Deep Neural Networks

A DNN is an Artificial Neural Network (ANN) with intermediate layers in between the input and output layers. We have used a feedforward network which accepts inputs and multiply the inputs with weights. It uses no backpropagation throughout the process. As the system may forget the training at times, we use a history function to help the system to remember the training done and process accordingly. Figure 8 shows the general diagram of DNN.

In this paper, DNN are applied using Keras, which is an easy and open source library of Python to construct and evaluate deep learning models. It encapsulates the most efficient numerical libraries of Python such as Theano and TensorFlow and allows us to define and train the models in a fewer lines of code. In order to normalize the data, we have used the StandardScalar function.

As hidden layers contribute much to greater accuracy, we have used six hidden layers for our work. Experimenting with each hidden layer and checking the accuracy of learning, we came to a decision that six hidden layers give the most better accuracy for our problem. The sequential API allows us to generate models for most of the problems each layer after layer.

As a trial and error method, we started with 1 layer and 50 neurons with ReLU activation, the Rectified Linear activation function which is a linear function that displays the output directly from the given input if it is positive or otherwise displays a zero. Then, we repeated the training with 100, 500 and 1000 neurons for the same layer. This process is repeated with 8 layers for 50, 100, 500 and 1000 neurons. After inputting same number of neurons, we tried applying different number of neurons in each layer. Then, we found that the system produces better accuracy when exposed to different number of neurons than that of same number of neurons. We observed the accuracy for each case, and we found that the accuracy is stable after the sixth layer for our dataset. Thus, we can say that the sixth layer gives the best accuracy among the whole process of deep neural networks. For the better performance, we made use of Adam optimizer which requires little memory requirements and is computationally efficient. DNN allows the usage of Dense function which implements the operation: $\text{output} = \text{activation}(\text{dot}(\text{input}, \text{kernel}) + \text{bias})$ where activation is the element-wise activation function which is passed as the activation argument.

Along with the ReLU activation, we have used Softmax function in each layer. Softmax transforms a real vector to a vector of categorical probabilities. The output vector are in the range (0, 1) and sum to 1. Softmax is normally used as the activation in the last layer of a classification network because we could get the result in the form of a probability distribution and it is also referred to as a Softmax Loss. Table 4 shows the loss and accuracy obtained in each hidden layers.

From Table 3, it is observed that 6 hidden layers with different neurons give the best and stable accuracy of 97.50%

Figure 9 shows the epoch vs accuracy graph of 4 hidden layers with different neurons.

Figure 10 shows the plot of 6 hidden layers with different neurons.

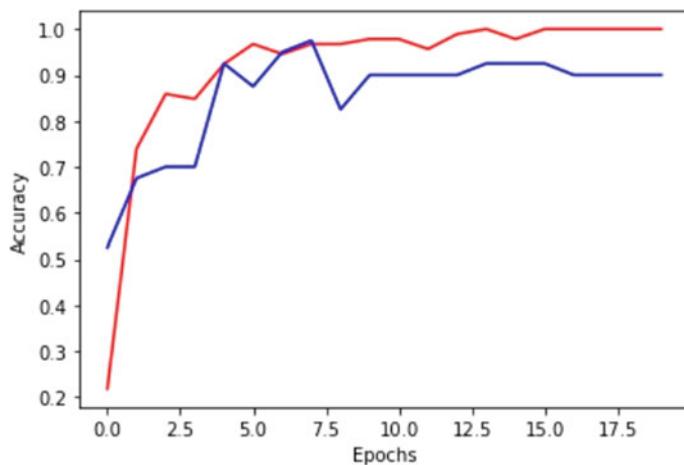
Table 3 Validation loss and Validation accuracy of each hidden layers

Hidden layers	Validation loss	Validation accuracy
1 layer with 50 neurons	1.8610	0.3750
1 layer with 100 neurons	1.4760	0.6250
1 layer with 500 neurons	0.9380	0.6750
1 layer with 1000 neurons	0.6450	0.7000
2 layer with 50 neurons	1.4960	0.5750
2 layer with 100 neurons	1.5678	0.6500
2 layer with 500 neurons	0.5289	0.8750
2 layer with 1000 neurons	0.7239	0.6750
3 layer with 50 neurons	1.7277	0.5250
3 layer with 100 neurons	1.5964	0.6250
3 layer with 500 neurons	1.5745	0.5500
3 layer with 1000 neurons	0.7249	0.6750
4 layer with 50 neurons	1.5938	0.5500
4 layer with 100 neurons	1.4203	0.6750
4 layer with 500 neurons	0.9367	0.6750
4 layer with 1000 neurons	0.9815	0.6750
5 layer with 50 neurons	1.7169	0.5000
5 layer with 100 neurons	1.6090	0.6750
5 layer with 500 neurons	1.0059	0.6750
5 layer with 1000 neurons	0.6825	0.6750
6 layer with 50 neurons	1.9881	0.5250

(continued)

Table 3 (continued)

Hidden layers	Validation loss	Validation accuracy
6 layer with 100 neurons	1.4739	0.6000
6 layer with 500 neurons	0.9095	0.6750
6 layer with 1000 neurons	0.6855	0.6750
4 layer with different neurons in each layer	1.8296	0.9000
5 layer with different neurons in each layer	0.6743	0.9000
<i>6 layer with different neurons in each layer</i>	<i>0.0330</i>	0.9750
7 layer with different neurons in each layer	0.02980	0.9750
8 layer with different neurons in each layer	0.02330	0.9750

**Fig. 9** Accuracy versus Epoch of 4 hidden layers with different neurons

5 Results

In this work, classification of taalas in Carnatic music has been done. Among different taalas, we focused on four major taalas for the classification. We dealt with the audio samples from the renowned dataset. Applying the machine learning algorithms, we could understand the best algorithms for taala classification are Kernel SVM and PCA with Logistic Regression with 98.11% accuracy. Initially, we started with the Random forest algorithm and we got an aggregate accuracy of 92.45%. The algorithm started

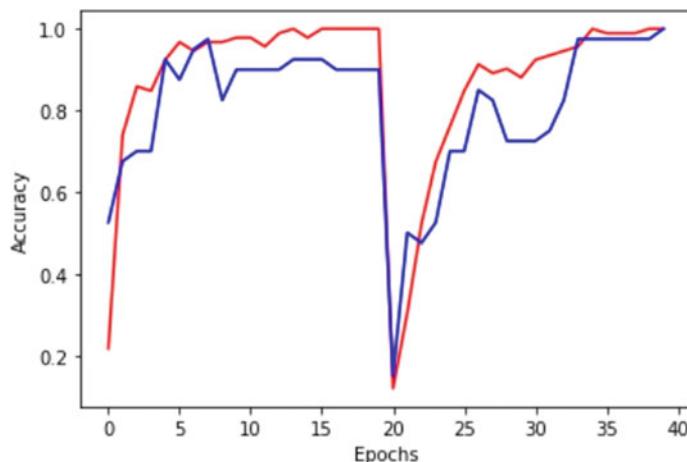


Fig. 10 Accuracy versus Epoch of six hidden layers with different neurons

with $n_estimator = 10$ and got an accuracy of 88.67%. Later we tried the value of $n_estimator$ to 20, 30 and so on. We observed 92.45% of accuracy for all those values. Kernel SVM performed better than the Random forest algorithm. The Gaussian RBF function added more sense to get an accuracy of 98.1%. The K-NN algorithm worked less when compared to the other algorithms. It offered 90.56% accuracy based on the calculation of nearest neighbours. Applying Logistic Regression classifier to PCA resulted in a better score of 98.11%, and it is one among the best classification algorithms for our dataset. We also trained the DNN with 8 hidden layers. Training with same number of neurons in each layer showed a gradual increase of accuracy, and we observed that the accuracy was better when different number of neurons were applied. Both 4 and 5 hidden layers with different neurons resulted in an accuracy of 90%. In the 6, 7 and 8th layer we got a stable accuracy of 97.50%, which is considered as the best among the whole neural network. The Kernel SVM and LR with PCA are the best classification models and their accuracies are described in Table 4.

Figure 11 shows a comparison graph of all methods used in the paper with their accuracies.

Table 4 Classification Models and their accuracies

Model	Accuracy (%)
Random Forest	92.45
Kernel SVM	98.1
K-NN	90.56
LR with PCA	98.11
DNN	97.50

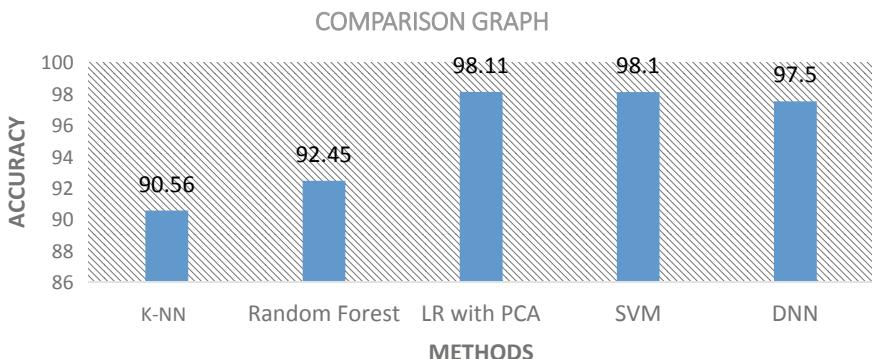


Fig. 11 A comparison graph of all used methods with their accuracies

6 Conclusion and Future Work

Machine learning algorithms like Kernel SVM and PCA with Logistic Regression gave the best result for our classification. We could classify the four taalas namely Aadi, Rupaka, Misrachappu and Khandachappu with possibly best and accurate result percentage of 98.1 for the above two algorithms. DNN also gave us better accuracy of 97.5% when compared to other methods. As the Carnatic music data is not much popular and open, we had to limit the dataset with only 176 samples of recordings. It might take more time to collect greater number of audio recordings from the concerts and stage events. Thus, we would like to extend the work with more sample data so that it may positively affect the accuracy and observations. Also we would also intend to work with the extensive set of taalas in Carnatic music.

References

1. Renteria, S.S., Llano, J.L.: Data-driven techniques for music genre recognition (2020)
2. Jacoby, N., McDermott, J.H.: Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction. *Curr. Biol.* **27**(3), 359–370 (2017)
3. Rao, P., Ross, J.C., Ganguli, K.K., Pandit, V., Ishwar, V., Bellur, A., Murthy, H.A.: Classification of melodic motifs in raga music with time-series matching. *J. New Music Res.* **43**(1), 115–131 (2014)
4. Anglade, A., Ramirez, R., Dixon, S.: Genre classification using harmony rules induced from automatic chord transcriptions. In: ISMIR, pp. 669–674 (2009)
5. Kan, A., Sankar, E.A., Sundhar, S., Yang, E.A.: A comparison of machine learning approaches to classify tala COMP 562 (2017)
6. Morris, R.: Sets, scales, and rhythmic cycles: a classification of talas in Indian music. In: 21st Annual Meeting of the Society for Music Theory (1998)
7. Srinivasamurthy, A., Serra, X.: A supervised approach to hierarchical metrical cycle tracking from audiomusic recordings. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5217–5221. IEEE (2014)

8. Nitha, K.P., Suraj, E.S.: An algorithm for detection of tala in carnatic music for music therapy applications (2019)
9. Heshi, R., Suma, S.M., Koolagudi, S.G., Bhandari, S., Rao, K.S.: Rhythm and timbre analysis for carnatic music processing. In: Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics, pp. 603–609. Springer, New Delhi (2016)
10. Srinivasamurthy, A., Subramanian, S., Tronel, G., Chordia, P.: A beat tracking approach to complete description of rhythm in indian classical music. In: Proceedings of the 2nd CompMusic Workshop, pp. 72–78. (2012)
11. Pulijala, A.S., Gangashetty, S.V.: Tala classification in carnatic music using audio thumbnailing (2020)
12. Ramirez, J., Flores, M.J.: Machine learning for music genre: multifaceted review and experimentation with audioset. *J. Intell. Inf. Syst.* **55**(3), 469–499 (2020)
13. Tian, H., Cai, H., Wen, J., Li, S., Li, Y.: A Music recommendation system based on logistic regression and eXtreme gradient boosting. In: International Joint Conference on Neural Networks (IJCNN) 2019, pp. 1–6. IEEE (2019)
14. Randall, R., Greenberg, A.S.: Principal components analysis of musicality in pitch sequences (2016)
15. Bahuleyan, H.: Music genre classification using machine learning techniques (2018). [arXiv: 1804.01149](https://arxiv.org/abs/1804.01149)
16. Vishnupriya, S., Meenakshi, K.: Automatic music genre classification using convolution neural network. In: International Conference on Computer Communication and Informatics (ICCCI) 2018, pp. 1–4. IEEE (2018)
17. Al Mamun, M.A., Kadir, I., Rabby, A.S.A., Al Azmi, A.: Bangla music genre classification using neural network. In: 8th International Conference System Modeling and Advancement in Research Trends (SMART) 2019, pp. 397–403. IEEE (2019)
18. Chillara, S., Kavitha, A.S., Neginalhal, S.A., Haldia, S., Vidyullatha, K.S.: Music genre classification using machine learning algorithms: a comparison (2019)
19. Steinmetz, C.J., Reiss, J.D.: Efficient neural networks for real-time analog audio effect modeling (2021). [arXiv:2102.06200](https://arxiv.org/abs/2102.06200)
20. Srinivasamurthy, A., Holzapfel, A., Cemgil, A.T., Serra, X.: Particle filters for efficient meter tracking with dynamic bayesian networks. In: Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015), pp. 197–203. Malaga, Spain (2015)

Utilizing the Data Mining Techniques for Obesity Prognosis Based on Eating and Lifestyle Routines of Adolescents and Adults



P. Vineetha Sankar and K. Sreekumar

Abstract Since 1980, obesity has increasingly become a global health issue—it is increasingly prevalent among children with serious implications for their health and their later well-being as adults. The World Health Organization considers obesity a non-communicable disease which may lead to severe diseases such as diabetes, cardiovascular diseases, and various kinds of cancers. Detrimental dietary patterns and the absence of physical activity are the root causes of obesity. Data mining is a quickly developing field with numerous applications in the field of medical services. One of its several utilizations is the forecasting of sicknesses. The proposed work intends to examine varied machine learning techniques: K- Nearest Neighbor, Decision Tree, Ensemble Random Forest, and Deep Neural Network to prognosticate obesity. The inference analysis shows that the ensemble random forest classification with XG Boosting is the most precise procedure to prognosis obesity.

Keywords Obesity · Non-communicable disease · Data mining · Ensemble random forest · Deep neural network

1 Introduction

Being overweight or obese is characterized by the accretion of unnecessary and/or abnormal fat—this may have serious effects on health. Viewing the world as a whole, it is seen that obesity causes more deaths than undernourishment. Overweight is a greater health hazard than underweight.

Some recent WHO global estimates follow:

- In 2016, in excess of 1.9 billion adults were overweight. Out of these 1.9 billion, over 650 million adults were found to be obese.

P. Vineetha Sankar (✉) · K. Sreekumar

Department of Computer Science and IT, Amrita School of Arts and Science, Kochi, Amrita Vishwa Vidyapeetham, India

K. Sreekumar

e-mail: sreekumar@asas.kh.amrita.edu

- In 2016, overweight was observed among 39% of adults.
- As a whole, approximately 13% of the world adults (11% of men and 15% of women) have been categorized as obese in 2016.
- The occurrence of obesity in the world as a whole has grown 3-fold during 1975–2016.

The mismatch between the intake of calories and energy expenditure is the root cause of obesity. Internationally, one sees

- Increase in the consumption of fat-rich, sugar-rich, high-energy food.
- Many jobs are getting more sedentary. Superior transport, greater urbanization, etc., have caused a reduction of physical activity.

Obesity is a major risk factor for non-communicable illness—heart disease, diabetes, certain types of cancer (for example, cancers afflicting endometrium, breast, ovaries, prostate, liver, gallbladder, colon, kidney, etc.).

Body Mass Index (BMI) is a straightforward indicator of weight-for-height that is generally used to group overweight and obesity in grown-ups. Rather than checking the BMI ratio to classify the overweight and obesity class, this work intends to predict obesity based on the dietary and lifestyle habits. The eating pattern of people has changed considerably over the past years. There's a rise in physical idleness due to the increasing desk-bound jobs, improved methods of transportation, and expanding urbanization. Hence, BMI may not appropriately outline threat of comorbid conditions. Many works made in the past to predict obesity were restricted to the calculation of BMI only. Our approach is to analyze the dietary patterns, physical habituating along with weight and height index to foretell obesity. The aim of the proposed work is to accomplish a framework that anticipates obesity in adults and adolescents using different data mining strategies. The study incorporates examination of some related works, dataset estimation, execution of algorithms to evaluate the performance and finally drawing inferences from the outcome of the study. An experimentation approach is followed to build the model Python 3.0 which is used for implementation along with various Python libraries like Numpy, Panda, Keras, and Tensorflow.

2 Related Works

Machine learning and Artificial Intelligence in medical care is one such domain which is gaining fast acknowledgment in the health care industry. Machine Learning (ML) is already been used to facilitate aids in different circumstances in healthcare. ML handles and interprets information of great variety and has been found great use especially in diagnostics.

Quite a few earlier researchers have explored the use of data mining in determining and predicting obesity.

Chatterjee et al. [1] build a model for early foretelling of childhood obesity for children with age upwards of 3 years using data from available patient clinical records. They implement machine learning algorithms such as SVM, KNN, and ANN. Application of these on a standard dataset for performance evaluation yielded the highest accuracy of 97.77%—achieved by the KNN algorithm [1].

Joshi et al. [2] proposed techniques such as Random Forest, Logistic Regression, Decision Tree, and Neural Network. Based on the ATUS dataset, they concluded that Random Forest works best as classifier for obesity prediction—its accuracy levels touch 91% [2].

Singh and Tawfik [3] build a model to prognosticate obesity in young people. They use data collected during the ages of 3, 5, 7, and 11. The algorithms they use include KNN, J48, Random Forest, SVM, MLP, etc. [3].

Taghiyev et al. [4] aimed to develop a hybrid classification model to determine the factors contributing to obesity in the country of Turkey. Their study uses a two-stage hybrid model for classification. They achieved an accuracy of 91.4% using Decision Tree and Logistic Regression to create a hybrid model [4].

Thomas et al. [5] proposed a system based on random forest for Credit Card Fraud Detection. They employ the Borda Count in preference to the Majority Voting technique [5].

De-La-Hoz-Correa et al. [6] use a dataset with details on lifestyle and habits from sections of the population of Peru and Columbia and proposed a model. They employed SEMMA data mining methodology for selecting, exploring, and modeling the dataset. The three algorithms they used for building the models are Decision Tree (J48), Naïve Bayes, and Logistic Regression. The J48 algorithm yielded the highest accuracy of 97.4% [6].

Chishti and Awan [7] examine a model using deep neural network to categorize default users of credit cards—they also describe the construction of a comprehensive deep neural network model. They observed that deep neural network can estimate the faults with much more accuracy than the simple neural network or logistic regression models [7].

Aswathi Anand and Pai [8] formulated a model that uses Artificial Neural Network for the prediction of early diabetes [8].

Hossain et al. [9] build a model to foretell the obesity risk factor in people. They have used various supervised machine learning algorithms to build the model. They observed that Naïve Bayes method has the best prediction using WEKA [9].

Ahn et al. [10] formulated a system to foretell obesity using Artificial Intelligence techniques. They have come to a conclusion that systems based on fuzzy rule is the best model to classify obesity [10].

Abdullah et al. [11] suggest a system that categorizes the obesity in school children of age 6 years. They discuss several machine learning algorithms to predict obesity in childhood using a dataset of Malaysian school children. The study used algorithms Bayesian Network, Decision Tree, Neural Network, SVM, etc. The best results were achieved with Decision Tree [11].

Benuwa et al. [12] examine deep learning methods in their paper titled “A Review of Deep Machine Learning.” This work elaborates the implementation of deep neural architecture [12].

Dugan et al. [13] aim to foretell obesity in children after age two. They use only data accessed before age two from Child Health Improvement through Computer Automation (CHICA). Their analysis of six different machine learning methods: Random Tree, Random Forest, J48, ID3, Naïve Bayes, and Bayes modeled on CHICA data exhibit that a precise, liable model can be developed [13].

Sivarajanji [14] constructs a model for comparing the ID3 and KNN approaches for obesity prediction. Her observations show that ID3 algorithms show greater accuracy in obesity prediction –92.17% [14].

Pochini et al. [15]. Used a dataset of school children in US to formulate a system that predict risks related to obesity and overweight. They have concluded that regular work outs and timely eating helps to prevent obesity. Whereas smoking and over use of sweetened beverages cause obesity [15].

3 Proposed Methodology

Machine learning builds applications that try to interpret and learn from data and progressively enhance their accuracy over time—without explicitly being programmed. The primary goal of our model shown in Fig. 1 is to construct an intelligent model for obesity prognosis. Initially, the data preprocessing is done, followed by feature selection process and the model is built for the training and testing phase. Divers machine learning techniques are examined for the experimentation process of this study, the algorithms used are KNN, Decision Tree and Ensemble Random

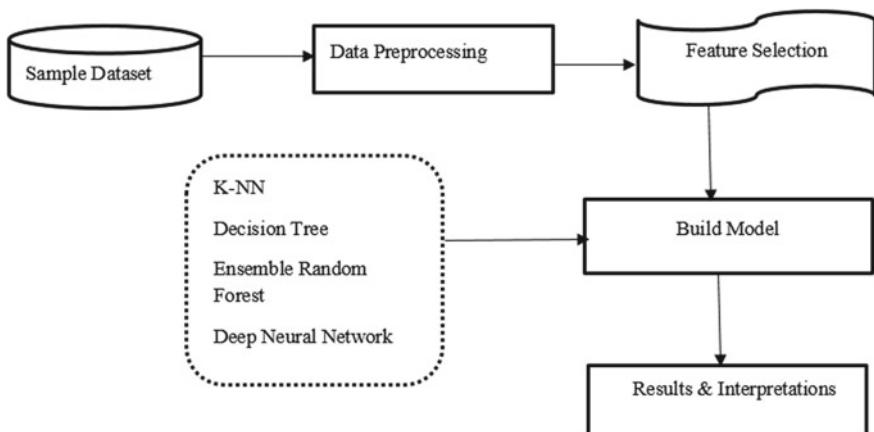


Fig. 1 Methodology flow diagram

Forest. A DNN is also implemented to analyze how best neural network can perform to predict obesity. After a few training and testing measures, the learned model will be constructed.

3.1 Dataset

This paper presents information for obesity level estimation among people from Mexico, Peru and Colombia, in view of their dietary patterns and physical condition. The dataset contains 2111 records of data and 17 attributes which are related to eating habits and physical conditions. The attributes in the dataset fall into three categories, dietary patterns, physical habituating, and responder features. Dietary patterns include information like the frequency of consumption of calorie rich food, frequency of vegetables intake, count of main meals, and number in between meals, daily water intake, and intake of alcohol. The attributes in physical habituating comprises frequency of workouts, mode of transportation using, habit of smoking, time spend in electronic devices and responder features incorporates age, gender, weight, height, family history of obesity. The dataset is accessed from UCI Repository.

Exploratory Data Analysis is carried out on the dataset which includes data visualization, analysis, and manipulation of the dataset. Initially, we check whether there are null values or redundant attributes in the dataset. We observed there was no missing value in the dataset. Followed by dividing the dataset into features and labels. Since there are 17 attributes in the dataset, we performed the feature selection to identify the best features out of the given attributes. Feature selection refers to identification and selection of a subset of input features that has greatest relevance to the target variable. The scikit-learn machine library delivers an ANOVA f-test implementation in the `f_classif()` function. This function could be employed as a strategy for feature selection, to determine the top k most pertinent features (largest values) using the `SelectKBest` class. Figure 2 indicates the features in the dataset and their F-scores and Fig. 3 portrays the plot between attributes in the dataset and their corresponding F-score.

$$F = \text{Variation between sample means}/\text{Variation within the samples} \quad (1)$$

Based on the feature selection, 9 features exhibited robust relationship with the target variable. Feature 9 and Feature 13 have been identified as the least related features and they have been dropped from dataset. Henceforth 14 features were selected for the implementation of model.

For the ease of computation, the categorical values are converted into numerical values. The normal and obesity type labels have been converted to numerical format for better evaluation and speed up the preprocessing step. The obesity type labels have been altered to Extremely Weak-0, Weak-1, Normal-2, Overweight- 3, Obesity- 4, extreme Obesity-5, highly effected: 6. The dataset used for the study should be a balanced one, if not the data mining techniques tend to learn and predict wrongly.

Fig. 2 F-Scores for the features

Feature 0:	116.651239
Feature 1:	49.677613
Feature 2:	24.516503
Feature 3:	1345.190158
Feature 4:	99.719712
Feature 5:	33.807217
Feature 6:	83.229177
Feature 7:	16.854648
Feature 8:	76.588993
Feature 9:	4.344741
Feature 10:	12.812574
Feature 11:	16.358249
Feature 12:	13.618358
Feature 13:	4.894826
Feature 14:	25.345751
Feature 15:	16.650519

Fig. 3 Plot for F-scores and features

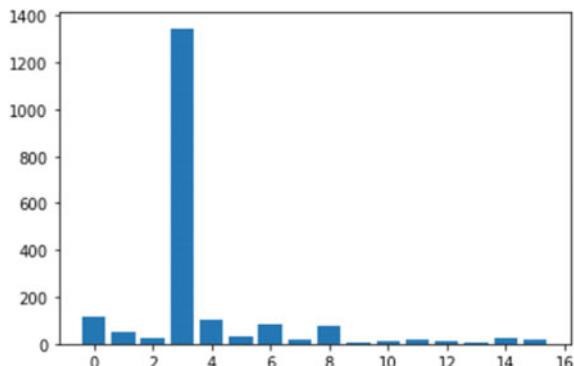


Figure 4 shows the proportion of different obesity class labels in the dataset. From the figure shown below, it is evident that the dataset used for the study is a balanced one the deductions made out of the study are more dependable and accurate.

The feature selection reveals that the weight attribute (Feature 3) in the dataset has the highest F-score; subsequently, it is the best element that contributes a lot to foretell the obesity. Figure 5 represents the relation between the weight and the target/label class obesity (obedad).

3.2 K-Nearest Neighbor (KNN)

The KNN classifier is one of the least complex yet most commonly used classifiers in supervised machine learning. KNN is frequently viewed as a lethargic learner; it doesn't in fact prepare a model to make forecasts. Rather a perception is anticipated

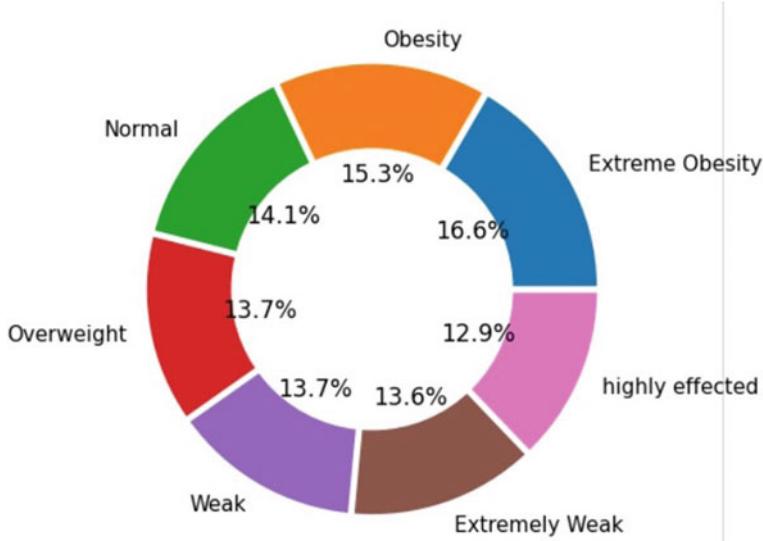
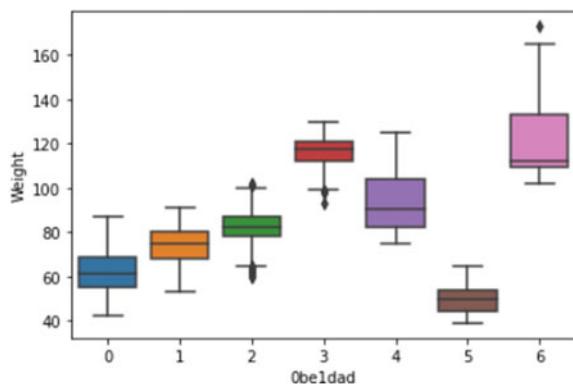


Fig. 4 Pie plot depicting proportion of different obesity types in the dataset

Fig. 5 Box plot between weight and obesity class



to be the class of that of the largest proportion of the k nearest observations [16, p. 251]. The KNN calculation accepts that comparable things exist in closeness. Otherwise, similar things are close to one another.

Dataset is divided into training set and testing set, feature scaling is performed using the StandardScaler. The StandardScaler makes the assumption that your data has a normal distribution within each specific feature and then, will scale them such that the distribution centers about 0, and the standard deviation becomes 1. The standardization replaces the values by their Z-scores and feature scaling helps to weigh all the features equally.

$$Z = (x - \mu)/\sigma \quad (2)$$

To implement KNN, initially choose the number of neighbors. We set $k = 5$, calculate the Euclidean distance between the data points using the below equation,

$$\text{Euclidean Distance } d(x, y) = \sqrt{\sum_{i=1}^m (X_i - Y_i)^2} \quad (3)$$

Sort the calculated distances and compute the data points in each classification. For the test data, allocate the data points to that class which has the maximum number of neighbors. The execution of the KNN resulted with an accuracy of 79.49%.

3.3 Decision Tree

Decision trees are viewed as classification algorithms with high efficiency. A decision tree is a tree-like structure, which begins from root attributes, and ends with leaf nodes. In general, the branches of a decision tree consist of different attributes, and the leaf node on each branch represents a class or a kind of class distribution.

A Decision Tree is developed by asking a series of inquiries regarding a record of the dataset we have. Each time an answer is obtained, a subsequent query is posed until a decision about the class label of the record is achieved. The inquiries and their potential answers can be coordinated as a decision tree, which is a hierarchical structure comprising of nodes and directed edges.

To train the decision tree classification model, the root node is identified. The total entropy for the classes is calculated. Perform the split by creating new internal nodes. The non-terminal nodes, which incorporate the root and other internal nodes, contain test conditions. Then, the entropy and information gain for every split is calculated. This process is continued until all the features are executed, each leaf node specifies a class label. Thus, the decision tree is generated [17].

$$\text{Entropy } E(S) = \sum_{i=1}^c -p_i \log_2 p_i \quad (4)$$

where “Pi” is the frequentism of a component or class “i” in the data.

Subtract the entropy of Y given X from the entropy of only Y to compute the reduction of variability about Y given an additional piece of information X about Y.

$$\text{Information Gain, IG(Y, X)} = E(Y) - E(Y|X) \quad (5)$$

Since entropy is the measure of uncertainty, it should be reduced. As the entropy decreases the information gain of the decision tree increases. The implementation

of decision tree algorithm indicated a precision rate of 95.43% which is a better accuracy than 89.14% Taghiyev et al. model [4].

3.4 Ensemble Random Forest

Random forest is an algorithm for supervised learning. It is used to perform classification as well as regression. In a random forest, numerous decision trees are trained, yet each tree just gets a bootstrapped sample of observations (i.e., a random sample of observations with substitution that matches the original number of perceptions) and every node possibly considers a subset of features while deciding the best split. This forest of randomized decision trees (henceforth the name) votes to decide the predicted class [16, p. 239, 18].

In ensemble random forest classification, we are using multiple decision trees as base learners. At first, we start with generating 10 trees in the forest (`n_estimators = 10`, i.e., `n_estimators` is the required number of trees in the Random Forest.) and the model is trained using the bootstrap sampling and the accuracy is observed as 92.59%. Later we increased the number of trees in the forest, which will eventually results in rise in the number of decision trees also. We created a list of trees from 100 to 2000, and the accuracy is noted at each point. As the number of trees increased in the forest, it is observed that the accuracy of the model also tend to rise since it minimizes the issues of over fitting by averaging the outcomes. It is noticed that when the count of trees was set to 200, 500, 1000, and 1500 we observed the most extreme exactness for the framework with an accuracy of 95.58% which is a better precision than 91% Joshi et al. system [2].

With the aim of increasing the prediction accuracy of the system further, we used XGBoost (Extreme Gradient Boosting). It is a boosting algorithm that uses the framework of Gradient Boosting Machine (GBM). Boosting is a sequential technique which functions based on the ensemble principle. It synthesizes a set of weak learners and delivers improved prediction accuracy. With the XG Boosting the overall accuracy of the model increased to 97.95%, giving the best results among the machine learning algorithms implemented.

3.5 Deep Neural Network (DNN)

It comprises neurons more than 3 layers, for example, input, output, and dense (also known as hidden) layers are called a DNN. In a general sense, DNN depends on the possibility of progressive or pile of layers of portrayal, these layers characterize how many layers are combined to build the data model-which gives the depth of the model [9]. We work with a feedforward neural network which accepts inputs and multiply them with loads and compute the results without backpropagation.

In this work, DNN is implemented using Keras, it is a famous Python library to construct, train, and evaluate a variety of neural networks. Keras is a high-level library, utilizing different libraries like TensorFlow and Theano as its “engine.” The advantage of Keras is that we can focus on network design and training, leaving the specifics of the tensor tasks to different libraries. [16, p. 298]

Since we followed an experimentation strategy, after the data preprocessing using StandardScaler function for normalizing the data, implement DNN using 1 hidden layer with 1 neurons and ReLU activation. ReLU gives the output the inputted values, if it is +ve, else gives zero. Adam optimizer is applied to resolve the optimization issues by minimizing the function. Epoch was set to 20 to train the data and Verbose was set to 1 outputting a progress bar. Figure 6a and b depicts the accuracy and loss observed at layer one with one neuron.

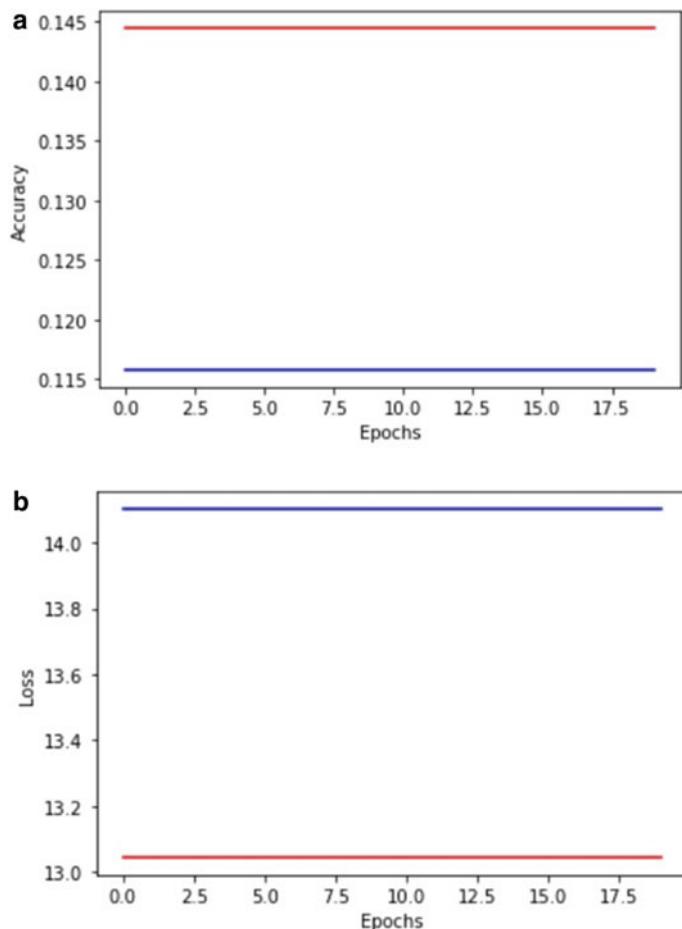


Fig. 6 **a** Accuracy versus Epochs for 1 hidden layer with 1 neuron. **b** Loss versus Epochs for 1 hidden layer with 1 neuron

The number of dense layers and neurons is increased to 2 and 50 each, respectively, with ReLU activation. Increase in hidden layer and number of neurons have brought a slight rise in the precision rate and drop in validation loss. In order to boost the accuracy further, again we increased the dense layers in the DNN to 3 and count of neurons to 100 each with ReLU activation. It is observed that the accuracy tends to elevate slightly as the number of hidden layers and neurons increase. But the rise is very slow. In the fourth layer, we utilized distinctive number of neurons at each layer and saw there is a high rise in the accuracy. Figure 7a and b represents the accuracy and loss plots with respect to the epochs for layer 4 with different count of neurons (500, 2000, 1000, 500). As the number of dense layers increased by 5 with distinct

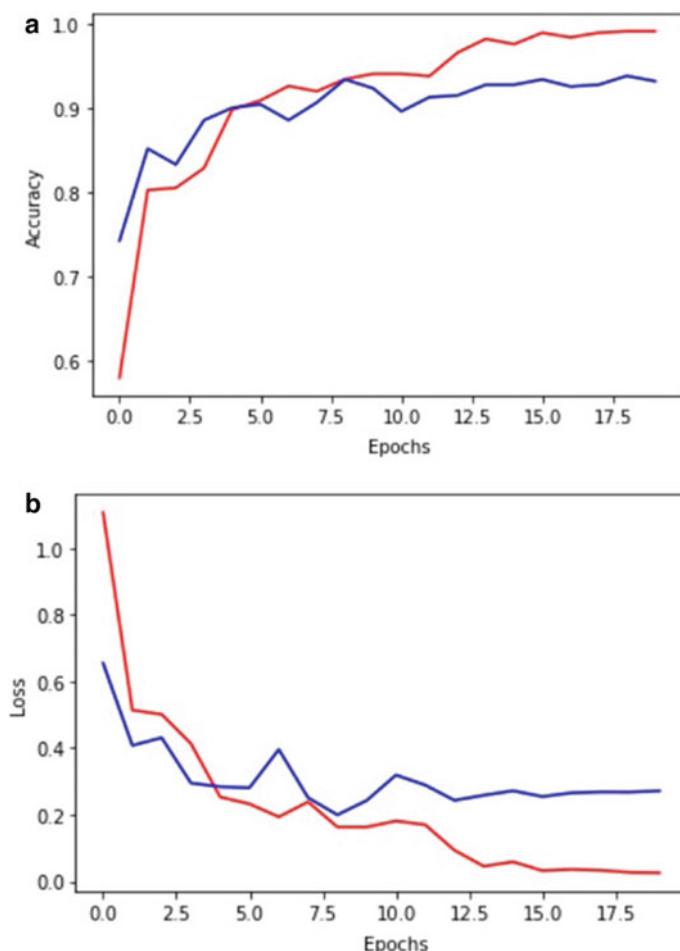


Fig. 7 **a** Accuracy versus Epochs for 4 hidden layers with different neurons at each layer. **b** Loss versus Epochs for 4 hidden layers with different number of neurons at each layer

Table 1 Validation Loss and Validation accuracy for each layer implemented in DNN

Number of hidden layers and neurons at each layer	Validation loss	Validation accuracy
One hidden layer with 1 neuron	14.1011	0.1158
Two hidden layers with 50 neurons at each layer	1.7957	0.1537
Three hidden layers with 100 neurons at each layer	1.8509	0.1558
Six hidden layers with 1000 neurons at each layer	1.5513	0.1495
Four hidden layer with distinct number of neurons at each layer	0.2717	0.9326
Five hidden layer with distinct number of neurons at each layer	0.2436	0.9347
Six hidden layer with distinct number of neurons at each layer	0.2843	0.9474

number of neurons (500, 2000, 4000, 2000, 500), we observed a very low rise in the accuracy and fall in the loss when compared with previous layer. Hence, it is evident that by expanding the layers and count of neurons results in rise the accuracy also. Hence, in the next step, the number of dense layers increased by 6 and number of neurons to 1000 each with ReLU activation. But it is noticed that the increase of dense layers with same number of neurons at each layer caused the system to exhibit a fall in the accuracy.

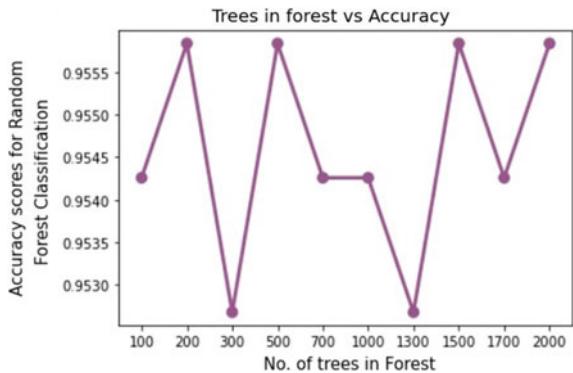
Execution of 4 and 5 dense layers with different number of neurons showed high leap in the precision rate. With the aim of attaining best prediction result for the system, the DNN with 6 hidden layers and different number of neurons (500, 8000, 4000, 2000, 1000, 500) in each hidden layer with ReLU activation function is executed. And it is observed that there is a leap in the accuracy of the system. The output layer uses the Softmax function, which is used in a classification network for the activation of the final layer to interpret the result as a probability distribution. Table 1 displays the validation loss and validation accuracy observed for different hidden layers implemented.

4 Results and Discussion

This section discusses the results that have been deducted from the implementation of the system. From the original dataset, we have identified the relationship of each feature with the target variable using the univariate feature selection. This process is called as Analysis of Variance (ANOVA). Thus, the least dependent features have been dropped from the dataset. The analyses are completed in Python 3.0. Python libraries like Numpy, Pandas, Keras, and Tensorflow are used.

Records of those selected attributes have been used to train our model. The dataset have been divided into 70% for training the model and 30% for testing the model.

Fig. 8 Change in accuracy as the N-estimators increases



The confusion matrix is used for the evaluation and visualization of the performance of supervised machine learning algorithms. The accuracy of the model is calculated using the formula [19].

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (6)$$

where TP—True Positive, TN—True Negative, FP—False Positive, FN—False Negative. These values will be evaluated from the confusion matrix.

Based on the study, it is observed that KNN predicts obesity with an accuracy of 79.49%, whereas Decision Tree has an improved accuracy rate of 95.43%. The ensemble random forest classifier predicts the obesity with much higher accuracy than any of the previously mentioned algorithms. When the number of trees in the random forest was set to default value 10, the model exhibited an accuracy of 92.59%. But when the N-estimators value set to higher range, we see the model achieve a saturation level with improved accuracy. It is obvious that as the number of trees in the forests increases that accuracy also increases by minimizing the overfitting issues. The ensemble Random Forest achieved an accuracy of 95.58%. Figure 8 depicts the accuracies obtained with respect to different counts of trees.

As a means of achieving a finer accuracy for the system further, we have used XGBoost (Extreme Gradient Boosting), boosting is a strategy which operates on the principle of ensemble. Accordingly achieved an improved accuracy of 97.95% for the model.

In the proposed system, we have implemented DNN to understand the efficiency of neural network in foretelling obesity. Out of the different layers implement in the study, we observed that the precision rate is high when we use distinct number of neurons rather than same number of neurons at each layer. Table 1 shows the results made out in the study of DNN. The model showed more exactness when we used 6 hidden layers and different neurons (500, 8000, 4000, 2000, 1000, 500) at each layer, respectively. The DNN achieved an accuracy of 94.74% for the model. Figure 9a and b shows the accuracy and loss for the final layer with respect to epochs.

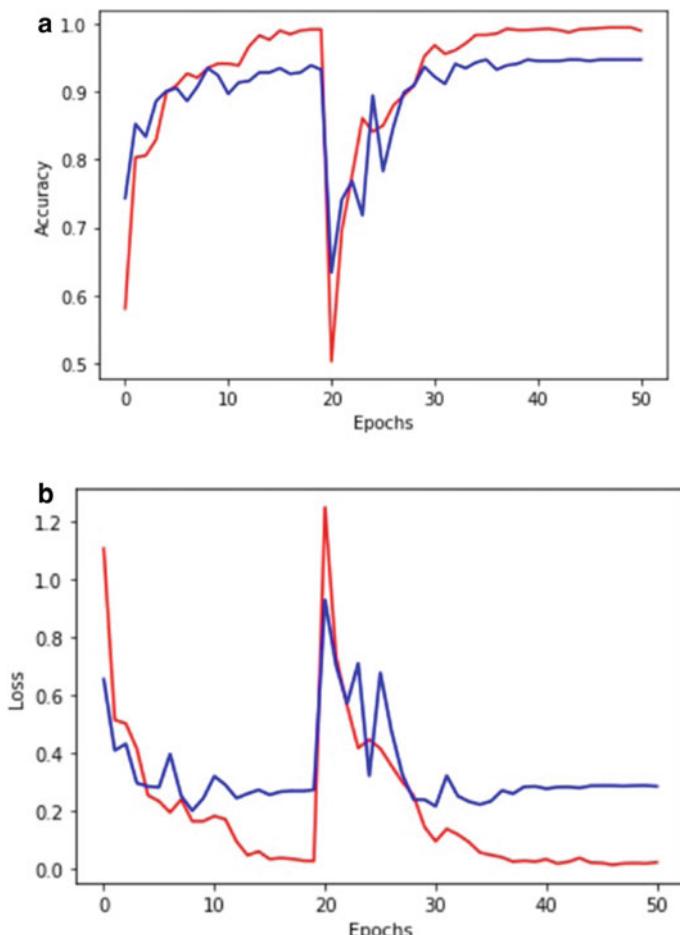


Fig. 9 **a** Accuracy versus Epochs for 6 hidden layer with distinct number of neurons at each layer. **b** Loss versus Epochs for 6 hidden layers with distinct number of neurons at each layer

Table 2. Illustrates the precision rates observed for the various models implemented. Hence based on the different analyses made on the system, we can say

Table 2 Accuracy of prediction with different model

Model	Accuracy (%)
K-Nearest neighbor	79.49
Deep neural network	94.74
Decision tree	95.43
Ensemble Random Forest with XgBoost	97.95

that an ensemble random forest classifier with XG Boost Technique has the highest prediction capacity and it is more desirable for our prediction model.

5 Conclusion and Future Work

Based on the study, it is evident that people who follow an unhealthy diet and sedentary life style tend to become overweight or obese. An eating pattern that's calorie rich, extensive use of sweetened beverages, excessive alcohol consumption, under supply of fruits and vegetables contributes to gaining weight. The desk-bounded job culture and absence of workouts are also found vital reasons for people to become obese. It is noticed that people who have a family history of obesity have higher chance to become an obese. A healthy dietary pattern and systematic physical activities can lessen the risks of being obese.

The proposed obesity prognosis study using various machine learning procedures: KNN, Decision Tree, Ensemble Random Forest classifier, and DNN resulted in a high prediction rate. Ensemble Random Forest with XGBoost has the highest precision rate among all the algorithms. Hence, the analysis helps the healthcare professionals to analyze the life style routines of people from different age groups and aid them to predict obesity. Thus, we can give awareness to people to follow healthy lifestyle routines and reduce overweight/obesity and its associated risks.

As a future work, we can include details of pregnant women as they gain weight during their pregnancy but which cannot be considered as obesity. Also the study can be expanded with information from handicapped people as they will not have body weight as compared to a normal person of their height. With a large dataset having more attributes related to lifestyle can be used for future study in order to get better prediction of obesity in terms of lifestyle habits. It has been observed that very less work has been made using genetic algorithms to predict obesity. So various genetic algorithms can be used to predict the obesity in future.

References

1. Chatterjee, K., Jha, U., Kumari, P., Chatterjee, D.: Early prediction of childhood obesity using machine learning techniques. In: *Advances in Communication and Computational Technology*, pp. 1431–1440. Springer, Singapore (2021)
2. Joshi, A., Choudhury, T., Sabitha, A. S., Raju, K. S.: Data mining in healthcare and predicting obesity. In: *Proceedings of the Third International Conference on Computational Intelligence and Informatics*, pp. 877–888. Springer, Singapore (2020)
3. Singh, B., Tawfik, H.: Machine learning approach for the early prediction of the risk of overweight and obesity in young people. In: *International Conference on Computational Science*, pp. 523–535. Springer, Cham (2020, June)
4. Taghiyev, A., Altun, A.A., Caglar, S.: A hybrid approach based on machine learning to identify the causes of obesity. *J. Control Eng. Appl. Inf.* **22**(2), 56–66 (2020)

5. Thomas, N., Jayalakshmi J., Sreelakshmi, E.S., Namboothiri, L.V.: Implementation of random forest and proposal of borda count in credit card fraud detection. *Int. J. Emerging Technol.* **11**(2), 536–540 (2020)
6. De-La-Hoz-Correa, E., Mendoza Palechor, F., De-La-Hoz-Manotas, A., Morales Ortega, R., Sánchez Hernández, A. B.: Obesity level estimation software based on decision Trees (2019)
7. Chishti, W. A., Awan., S. M.: Deep neural network a step by step approach to classify credit card default customer. In: 2019 International Conference on Innovative Computing (ICIC), pp. 1–8. IEEE. (2019, November)
8. Aswathi Anand, P., Pai, M. L.: Artificial Neural Network Model for Identifying Early Readmission of Diabetic Patients
9. Hossain, R., Mahmud, S.H., Hossin, M.A., Noori, S.R.H., Jahan, H.: PRMT: predicting risk factor of obesity among middle-aged people using data mining techniques. *Procedia Comput. Sci.* **132**, 1068–1076 (2018)
10. Ahn, S.H., Wang, C., Shin, G.W., Park, D., Kang, Y. H., Joibi, J.C., Yun, M.H.: Comparison of clustering methods for obesity classification. In: 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), pp. 1821–1825 (2018, December)
11. Abdullah, F.S., Abd Manan, N.S., Ahmad, A., Wafa, S.W., Shahril, M.R., Zulaily, N., Ahmed, A.: Data mining techniques for classification of childhood obesity among year 6 school children. In: International Conference on Soft Computing and Data Mining, pp. 465–474. Springer, Cham (2016, August)
12. Benuwa, B.B., Zhan, Y.Z., Ghansah, B., Wornyo, D.K., Banaseka Kataka, F.: A review of deep machine learning. *Int. J. Eng. Res. Afr.* **24**, 124–136 (2016). (Trans Tech Publications Ltd.)
13. Dugan, T.M., Mukhopadhyay, S., Carroll, A., Downs, S.: Machine learning techniques for prediction of early childhood obesity. *Appl. Clin. Inf.* **6**(3), 506 (2015)
14. Sivaranjani, T.: Comparative study on Obesity based on ID3 and KNN. *Int. J. Adv. Res. Comput. Sci. Manag. Stud.* **2**(9), 389–396 (2014)
15. Pochini, A., Wu, Y., Hu, G.: Data mining for lifestyle risk factors associated with overweight and obesity among adolescents. In: 2014 IIAI 3rd International Conference on Advanced Applied Informatics, pp. 883–888. IEEE (2014, August)
16. Albon, C.: Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning, 1st edn. O'Reilly Media, Inc., Japan (2018)
17. Ali, J., Khan, R., Ahmad, N., Maqsood, I.: Random forests and decision trees. *Int. J. Comput. Sci. Issues (IJCSI)* **9**(5), 272 (2012)
18. Goel, E., Abhilasha, E., Goel, E., Abhilasha, E.: Random forest: a review. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **7**(1), 251–257 (2017)
19. Han, J., Kamber, M., Pei, J.: Data Mining Concepts and Techniques., 3rd edn. Morgan Kaufmann Publishers is an Imprint of Elsevier. USA (2011)

An Optimization of Feature Selection for Classification Using Modified Bat Algorithm



V. Yasaswini and Santhi Baskaran

Abstract Data mining is the action of searching the large existing database in order to get new and best information. It plays a major and vital role nowadays in all sorts of fields like Medical, Engineering, Banking, Education, and Fraud detection. In this paper, Feature selection which is a part of Data mining is performed to do classification. The role of feature selection is in the context of deep learning and how it is related to feature engineering. Feature selection is a preprocessing technique which selects the appropriate features from the data set to get the accurate result and outcome for the classification. Nature-inspired Optimization algorithms like Ant colony, Firefly, Cuckoo Search, and Harmony Search showed better performance by giving the best accuracy rate with less number of features selected and also fine f-Measure value is noted. These algorithms are used to perform classification that accurately predicts the target class for each case in the data set. We proposed two different techniques from Wrapper-based feature selection methods namely Forward/Backward Elimination and Exhaustive Search to find best optimal solution to perform classification using different Nature-Inspired Algorithms out of which Modified Bat Algorithm is proposed in this paper. We applied new and recent advanced optimized algorithm named Modified Bat algorithm on UCI datasets that showed comparatively equal results with best performed existing firefly but with less number of features selected. The work implemented using JAVA and the Medical dataset (UCI) has been used. These datasets were chosen due to nominal class features. The number of attributes, instances, and classes varies from chosen dataset to represent different combinations. Classification is done using J48 classifier in WEKA tool. We demonstrate the comparative results of the presently used algorithms with the existing algorithms thoroughly.

Keywords Optimization · Meta-heuristic · Feature Extraction · Deep learning

V. Yasaswini (✉)

Computer Science and Engineering Department, Pondicherry Engineering College, Puducherry, India

e-mail: yasaswini2907@pec.edu

S. Baskaran

Information Technology Department, Pondicherry Engineering College, Puducherry, India

1 Introduction

Data Mining [1] is the way of searching important information from the huge present all over in the repository. Data Mining falls into two ways namely Association and Classification analyzing methods.

Optimization algorithm provides a systematic way of developing and leveling new solutions to gain an optimal result. The optimization process must only be used in those problems where there is a specific need of accomplishing a quality or a competitive work. It is expected that the solution obtained through an optimization method is better than other results in terms of the selected objective [2].

This paper shows the Bat algorithm and Modified Bat algorithm accuracy rates when compared to existing algorithms namely Firefly, Cuckoo search, and Harmony Search algorithms that showed almost equal results of the best accuracy rates in existing work [3]. There are various applications with respect to data mining and optimization techniques in different fields. This method proves the better analysis which gives the best results and improved accuracy. The following are the different fields of applications.

1. Network Security
2. Computer Vision and Processing
3. Nature-Inspired fields.
4. Medical Fields
5. Transition Probabilities for Radio Systems
6. Intrusion Detection
7. Education
8. Financial Banking.

2 Overview on Datamining

Data mining process involves the following stages.

- (a) **Problem definition.** In this stage, the analysis of the problem in the business problem is done and tries to get the clear idea of the problem to be solved. This takes some time to make an exact definition of the problem, and it does not require any data tools.
- (b) **Exploration of Data.** In this stage, data is explored by identifying quality problem to understand the metadata meaning. It is next level of problem definition stage which frequently exchanges the data.
- (c) **Preparation of Data.** In this stage, data model is built after the exploration of data. Collect the data, clear the unwanted data and arrange the data in a format like tables and records.
- (d) **Data Modeling.** At this stage after preparation of information, different mining functions are applied to the same kind of data. A high quality of mining model

is prepared based on the changes in the parameters until we get optimal data model. Finally, the good quality model is built and evaluated.

- (e) ***Evaluation of the Model.*** In this stage, the evaluated model is checked and tested whether the quality is good or not and objective is satisfied or not?
- (f) ***Deployment.*** In this stage after the evaluation of data, the exporting of the data is done and the results are checked into database tables.

3 Description of Algorithms

3.1 Existing Work Algorithms-Firefly Algorithm

Firefly Algorithm (FA) [4] being a Nature-Inspired algorithm works based on flashing nature of the fireflies. The main purpose of Fireflies is to show its flashing behavior which scatters into the system and looks for other fireflies. The algorithm was implemented to perform Feature Selection (FS) for Image Processing and Eigen Value Optimization [5] problem along with other related domains and has been performed so that better results are obtained. In order to achieve the best optimal feature subset increases the predictive accuracy of the classifier. In this algorithm, along with considering the brighter firefly to obtain the predictive accuracy of the dataset, we have also considered a comparatively brighter firefly and the predictive accuracy of that method have also been calculated.

- **Choosing a Brightest Firefly**
- **Choosing a Comparatively Brighter Firefly**

It includes previously chosen firefly solution and the newly selected brighter firefly solution in our computation. The latest solution is found by solving the below formulation

$$NewX_i = X_i + \beta_0 \exp(-\gamma(r_{ij})^2)(X_j - X_i) + \alpha \epsilon_i$$

where **Parameter Settings of Firefly Algorithm.**

X_i	Variable for present firefly (accuracy for classification)
X_j	The solution pointed by the brightest firefly (First Method) The solution pointed by the comparatively brighter firefly (Second Method)
β_0	Between 0 and 1
γ	Between 0.1 and 10
r_{ij}	Distance is fixed to 1
α	Parameters selected within the range [0,1] randomly

Implementation of Firefly Algorithm

1. **First iteration (i)**
2. **FA arguments are initialized**
3. **Calculate the Individual feature fitness**
4. **Iterate it again**
5. **Build the results from the fireflies (X_i)**
 - **Construct the “BRIGHTEST FIREFLY”**
 - **Subset of features is assigned for configuration (binary bit string) to each firefly**
 - **Develop latest feature subset**
 - **Traverse the developed feature subset to the present classifier**
 - **Calculate the fitness of the feature subset by computing the new value of current firefly**
 - **Reset the value of current firefly is based on either consideration of rejection of the “BRIGHTEST FIREFLY”**
6. **Evaluate the good feature subset in the present iteration**
7. **Iteration $i = i + 1$**
8. **Iterates till it reaches its final range**
9. **Utilize the same procedure of searching to develop the best optimal feature subset**

3.2 Existing Work Algorithm-Cuckoo Search Algorithm (CSA)

Cuckoo Search Algorithm (CSA) [6] is recognized as good optimization algorithm [7] to solve Hybrid Model Problems. This Algorithm is based on the behavior of cuckoo breeding. As this algorithm showed good results for engineering problems, we proposed this algorithm for feature selection to perform classification. Therefore, Cuckoo Search Algorithm (CSA) for FS is implemented. In this Cuckoo Search Algorithm, three different processes happen namely Hatching of eggs, Leaving the Nest, and Survival of the eggs. This algorithm uses its past memories for the place and condition of the eggs laid by those cuckoos.

This algorithm mainly concentrates on replacing not good nests with the potentially good nests. The location of the present egg replaces the new location of the eggs in nest or not and finally the eggs have been removed or not.

- **Hatching of eggs—Survival of the Cuckoo.**
- **Abandoning the nest—Host bird abandons its nest and migrates to some other place to build another nest.**
- **Evicting the eggs—The host bird throws the cuckoo bird’s eggs.**

The below-mentioned equation is used as computational formula of Cuckoo Search Algorithm.

$$F_{ij} = \frac{\{(\alpha [I_i(next) - I_i(bp)]) * iter\}}{maxcuckoo}$$

Parameter settings for Cuckoo Search Algorithm

F_{ij}	Fitness function used to find the alpha value of the cuckoo
$I_i(next)$	Accuracy of the cuckoo bird being selected
$I_i(bp)$	Accuracy of the host cuckoo bird
Iter	Fixed to 1
∞	Between 0 and 4
Maxcuckoo	Maximum number of cuckoos in a particular dataset

Implementation of Cuckoo Search Algorithm

1. **Iteration(i) = 1**
2. **The CSA arguments are initialized**
3. **Generate the initial population of Host Birds' nests and the Cuckoo Birds' nests**
4. **Evaluate the status of eggs with its own feature**
5. **Iterate it again**
6. **Build the new result based on fitness of the cuckoos eggs (F_{ij})**
 - Construct the “STATUS OF THE EGGS – Eviction, Abandon, Survive”
 - Choose the cuckoo according to the alpha value computed in the fitness function by comparing with the alpha value of the other features
 - Subset of features is assigned for configuration (binary bit string) to each egg of the cuckoo
 - Develop latest feature subset
 - Traverse the developed feature subset to the present classifier
 - Calculate the fitness of the feature subset by computing the new value of F_{ij}
 - Compute the new solution of the feature and find the new solution based on its accuracy done through classification
 - Reset the value of F_{ij} based on the rejection either by eviction or abandon or survive
7. **Evaluate the good feature subset in the present iteration**
8. **Iteration i = i + 1**
9. **Iterates till it reaches its final range**
10. **Utilize the same procedure of searching to develop the best optimal feature subset.**

3.3 Existing Work Algorithm-Harmony Search Algorithm (HSA)

- Harmony Search Algorithm (HSA) is a music-based meta-heuristic optimization algorithm which is conceptualized using the musical process to search for the perfect state of harmony and based on value of frequency the algorithm is built.
- Musical performances seek to find to be a pleasing harmony determined by the aesthetic standard, based on the 3 parameters
 - Pitch
 - **Amplitude**
 - Timbre

In HSA, individual music player (variable which takes decision) runs a bit of music (value) which finds the best harmony (global optimum) at the last.

The below-mentioned equation is used as computational formula by Amplitude of Harmony Search Algorithm (HSA)

$$\lambda = \frac{C}{h_i[\text{next}] * h_i[\text{music}]}$$

Parameter Settings for Harmony Search Algorithm (HSA)

λ	Lambda value (Amplitude)
C	Velocity of Light
$h_i(\text{next})$	Accuracy of newly selected tune
$h_i(\text{music})$	Accuracy of the existing tune that needs to be compared with the new tune

Harmony Search Algorithm

1. **Iteration(i) = 1**
2. **The HSA arguments are initialized**
3. **Generate initial population tune for a perfect harmony**
4. **Evaluate the amplitude of the tone (Lambda) of its own feature**
5. **Iterate it again**
6. **Build the result by the Lambda value of the tones (λ)**
 - Construct the “AMPLITUDE OF THE TUNES – Noise or Melody”
 - Compare frequency of each harmonic tune and generate similar frequency of harmonic tunes
 - Choose the tune according to the lambda value being computed through the amplitude by comparing with the lambda value of the other tunes (features)
 - Develop latest feature subset

- Traverse the developed feature subset to the present classifier
 - Calculate the fitness of the feature subset by computing the new value of λ
 - Compute the new solution of the feature and find the new solution based on its accuracy done through classification
 - Get the latest harmonics (solutions) if it shows good results
 - Calculate the latest solution and find the new solution based on its accuracy done through classification
 - Reset the value of λ based on the rejection either filtering by noise or by melody
7. Evaluate the good feature subset in the present iteration
 8. Iteration $i = i + 1$
 9. Iterates till it reaches its final range
 10. Utilize the same procedure of searching to develop the best optimal feature subset.

4 Proposed Work Algorithm- Modified Bat Algorithm Applied in Feature Selection [11]

Objective Function: Count of Bats $X_i = X_{i1} \text{ to } X_{iD}$ to the power of T . where I belongs to the range $[1, N_p]$ [13]

1. Initialize the count of Bats in the Search space X_i and V_i
2. Initialize the frequencies (f_i), Pulses (r_i) and the Loudness (A_i)
3. while ($t < \text{Max number of iterations}$)

Generate new solutions by adjusting frequency,
Update velocities and locations/solutions

4. if ($\text{rand} > r_i$)

Select a solution among the best solutions
Generate a local solution around the selected best solution
end if
Generate a new solution by flying randomly
5. if ($\text{rand} < A_i \& f(x_i) < f(x^*)$)

Accept the new solutions
Accept the new solutions
Increase r_i and reduce A_i
6. Rank the bats and find the current best x^*

end while

In this algorithm [14], a little change in the computation of emission of pulse rate and loudness of the bats is improved which showed superior results when evaluated with the Bat algorithm.

The following pseudocode of the Modified Bat algorithm is explained.

- i. Firstly initialize the objective function of the algorithm and assign the best solution x_{best} in the count of bats.
- ii. In the second, new solutions are generated according to the movement of the virtual bats in the search space.
- iii. In the third step, best solution is determined by using random walks in local search.
- iv. In the fourth step, Evaluation of the new solution is moved out.
- v. In the fifth step, bank the current best solution.
- vi. In the last step, compute the best solution and update it.

The echolocation behavior of microbats is explained in this algorithm. The primary use of Bat Algorithm (BA) [8] is variant behavior of the microbats based on the frequency tuning, velocity v_i^t , and its location x_i^t ; and calculates based on the iteration t with d -dimensional search or solution space. According to Yang, the following mathematical equation is updated and written as

$$\begin{aligned} f_i &= f_{min} + (f_{max} - f_{min})\beta \\ v_i^t &= v_i^{t-1} + (x_i^t - x*)f_i \\ x_i^t &= x_i^{t-1} = +v_i^t \end{aligned}$$

where $\beta \in [0,1]$ which uniformly distributes the random vector is a random vector drawn from a uniform distribution. A direct exploitation for searching local solutions modifies the present good result.

$$x_{next} = x_{prev} + \epsilon A$$

From the above equation, ϵ refers to range of $[-1, 1]$ which can be any number between -1 and 1 . A^t refers to the calculation of overall best loudness mean. During the iterations, the emission of pulse from bats and its loudness is variant which is calculated based on the following equation,

$$\begin{aligned} A_i^{t+1} &= a A_i^t \\ \gamma_i^{t+1} &= \gamma_i^0 [1 - \exp(-\gamma t)] \quad \text{Where } 0 < a < 1 \text{ and } \gamma > 0 \text{ are constants.} \end{aligned}$$

The main modifications in this algorithm are changing the ranges of the parameters defining $A_0 = 1$ and **Amin = 0, fmin = 0 and fmax = 2, $\alpha = \gamma = 0.9$ to 0.975**.

Table 1 Dataset description

Datasets	Instances	Features	Class
Heart-C	303	14	02
Dermatology	366	34	06
Hepatitis	155	19	02
Lung Cancer	32	56	02
Pima Indian diabetes	768	08	02
Iris	150	04	03
Lymphography	148	18	04
Diabetes	768	09	02
Heart-Stalog	270	13	02
Audiology	226	74	10

5 Experimental Setup and Result Analysis

Fourteen standard datasets drawn from the UCI collection were used in the experiments. These datasets were chosen due to nominal class features. The number of attributes, instances, and number of classes vary in the chosen dataset to represent different combinations. All the features in tenfold cross-validation run through Weka tool to get classification accuracy. The classifier used for evaluating the feature subsets generated is J48, Naïve Bayes, and Logistic. Feature subset (FS) generation by Firefly Algorithm (FA), Cuckoo Search Algorithm (CSA), and Harmony Search Algorithm (HSA) is implemented using Net Beans IDE in the existing work. Feature subset generation by Nature-inspired algorithms named Modified Bat algorithm (MBA) has been implemented using Net Beans IDE while the UCI dataset of Medical field is run through WEKA tool to get the classification rate which is processed through J48 classifier in the proposed work.

Table 1 describes the different types of diseases in medical field and instances, Features and Class of each disease. For example, in this dataset Iris disease can be diagnosed by 4 features and it falls into 3 classes.

It is clear from Table 2 that the accuracy of the Modified Bat algorithm is best when compared to other existing and the proposed Bat algorithm.

From Fig. 1, we can observe that the Modified Bat algorithm which is obtained for 10 UCI datasets showed better accuracy when compared to other algorithms.

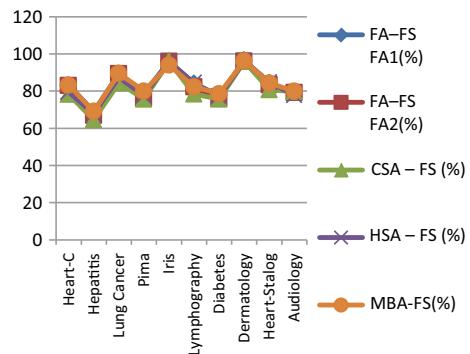
In this Table 3, it is clear that the number of features selected for Modified Bat algorithm is best when compared to other existing and the proposed Bat algorithm.

6 Inference

- From Table 2, We can infer that Modified Bat Algorithm applied for UCI dataset for FS with respect to Accuracy gives better as well equal accuracies

Table 2 Comparison of all the accuracy of existing with proposed algorithms

Datasets	FA-FS FA1(%)	FA-FS FA2(%)	CSA-FS (%)	HAS-FS (%)	MBA-FS(%)
Heart-C	83.15	83.07	78.217	79.53	83.191
Hepatitis	69.03	67.00	64.516	67.74	69.072
Lung Cancer	89.50	89.33	84.375	87.37	89.443
Pima	78.70	76.10	75.651	77.47	79.981
Iris	96.00	96.00	96.00	96.00	94.00
Lymphography	84.10	82.25	78.378	84.43	82.295
Diabetes	77.47	76.00	75.65	77.21	78.583
Dermatology	96.72	96.17	95.901	96.07	96.210
Heart-Stalog	84.44	83.39	80.74	84.81	84.345
Audiology	79.051	79.20	79.646	77.87	79.942

Fig. 1 Graphical representation of accuracy of all these algorithms**Table 3** Comparison of all the features selected in existing with the proposed algorithms

Datasets	FA-FS	CSA-FS	HAS-FS	BA-FS	MBA-FS
Heart-C	6	7	5	6	5
Hepatitis	8	10	9	8	9
Lung Cancer	14	20	18	16	18
Pima	3	3	4	4	3
Iris	2	3	2	2	2
Lymphography	8	7	10	8	7
Diabetes	4	5	6	5	4
Dermatology	23	22	25	21	23
Heart-Stalog	7	6	7	8	7
Audiology	53	55	49	56	51

- By comparing Firefly Algorithm (FA), Cuckoo Search Algorithm (CSA), and Harmony Search Algorithm (HSA) for FS with respect to Features: we can infer that the features for Modified Algorithm get reduced as well get increased in some of the datasets in other Algorithms.

7 Conclusion

In the recent times, Data mining and Data Science are the leading techniques to be used by almost all the different fields to get the best results mainly in Medical field and Banking field. Data mining mainly concentrates on feature selection and optimization problems which are the most common problem nowadays. The proposed system for FS Optimization to perform Classification is applied, and the results have been obtained using UCI datasets. The datasets are taken from UCI repository, and Table 1 describes the 10 datasets that we have used. The UCI datasets are used in continuing works in literature; such as Classification, FS and Classifier Ensemble, Stacking Ensemble, so we have adopted these datasets for our work. This paper has attempted to give accuracy results of the several popular Meta-Heuristic algorithms. Finally, analyzing the proposed algorithms compared to existing algorithms, we found that Modified Bat algorithms showed better results in some particular data in the dataset.

References

1. Tan, S.K.: Introduction to Data Mining (2005)
2. Bergh, F., Engelbrecht, A.P.: A study of particle swarm optimization particle trajectories. *Inf Sci* **176**, 937–971 (2006)
3. Basturk, B., Karaboga, D.: An artificial bee colony (ABC) algorithm for numeric function optimization. In: IEEE Swarm Intelligence Symposium, 12–14 May, Indianapolis (2006)
4. Hassan, A.E. et.al.: Extensions of Dynamic Programming for Combinatorial Optimization and Data Mining (2018)
5. Rao, R.V.: Teaching learning based optimization algorithm. And Its Engineering Applications. Springer (2015)
6. Sunil, K.: Datamining and optimization techniques. *Int. J. Statistica Mathematica*. ISSN. 2277-2790, E-ISSN. 2249-8605, **6**(2), 70–72 (2013)
7. Xin-She, Y., He, X.: Bat algorithm: Literature review and applications. *Int. J. Bio-Inspires Comput.* **5**(3), 141–149 (2013)
8. Nidhi, T., Amit, K.M.: A survey on data mining optimization techniques. *Int. J. Sci. Technol. Eng.* **2**(06) (2015). ISSN (online). 2349–784X
9. Rao, R.V., Kalyankar, V.D.: Parameter optimization of modern machining processes using teaching–learning-based optimization algorithm. *Eng. Appl. Artif. Intell.* **26**(1), 524–531 (2013)

A Low-Cost Web Interface for Object Tracking Based on a Wireless Sensor Network



Juan P. Carvajal, Arturo Fajardo, and Carlos Paez

Abstract This paper describes the design of a minimum cost web interface for object tracking applications based on open (or free) tools. Furthermore, the different blocks and parts that compose it were detailed; this paper was written as a tutorial to motivate engineers and students with basic knowledge on programming tools to integrate these web interfaces in their projects. The platform features were validated with a study case, which is based on a low-cost LoRa Network and the developed web platform. The implemented features exhibit complete functionality.

Keywords Web services · Databases · Open-source software · Global positioning system · Wide area networks · Wireless sensor networks

1 Introduction

Nowadays a large number of projects solve social problems using Wireless Sensor Networks (WSNs) [1], which are used to recollect information and to bring services to final users or to collect environmental data and trigger a response regarding it. In particular, the Internet of Things (IoT) based on WSN is becoming popular in tracking objects applications (i.e., tracking of bicycles and portable computers). This approach allows reducing the number of cases of theft [2, 3]. The conventional WSN implementation for tracking applications involves 2G/3G mobile network [4]. However, Low Power Wide Area Networks (LPWANs) technologies offer a lower power consumption [4, 6]. In particular, LoRa allows a wide coverage area (i.e., 1–10Km [5]), with low consumption [7] using unlicensed spectrum bands (i.e., low-

J. P. Carvajal · A. Fajardo (✉) · C. Paez
Pontificia Universidad Javeriana, Bogotá, Colombia
e-mail: fajardoa@javeriana.edu.co

J. P. Carvajal
e-mail: carvajal_juan@javeriana.edu.co

C. Paez
e-mail: paez.carlos@javeriana.edu.co

ers system operating costs [8]). Furthermore, LoRa is being adopted within novel applications, such as the smart campus concept [9, 10]. This paradigm is supported on WSNs, which collects data (e.g., air quality data of the campus) in the network nodes, and sends the sensed data to the cloud by a Gateway. This information is stored in a database, which feeds a user interface for visualization purposes [10]. The successful deploying an IoT system (i.e., smart campus concept) involves at least a communication infrastructure (e.g., a WSN) and end-user tools, as is shown in Fig. 1. The conventional user tools are interfaces, customization tools, and generic monitoring platforms (e.g., supervise and collect platforms).

The designer of a low-cost web IoT systems finds several options available to implement the communication infrastructure (e.g., Arduino-like platforms). However, it must overcome the lack of development (and exploration) of low-cost tools for the end-user [1], especially for specific features (i.e., to generate an alert in a tracking). Consequently, the specialized IoT systems (e.g., an object tracking system) based on WSN solutions have been limited only for a selected group of users (e.g., large companies, educational institutions).

This paper presents the development of a minimum-cost web interface for object tracking based on a wireless sensor network, which is based on open-source tools or free-tools (with limitations). The platform features were validated with a study case, which is based on a low-cost LoRa Network and the developed web platform. The implemented features exhibit complete functionality. The outline of this paper is as follows. Section 2 presents the proposed network system, Sect. 3 presents an overview of the interface architecture. Sections 4 and 5 explain the characteristics and development of the general and specific features of the system. Section 6 shows tests carried out and its results. Finally, Sect. 7 presents the conclusions.

2 Proposed Network Architecture

A general architecture of a Lora Network for object tracking system is shown in Fig. 1, and the localization data is sent by the transmitting node [11, 12]. These data (i.e., coordinates) are received by the gateway through a wireless communication link (based on the LoRa protocol). Then, the data is sent to a computer by serial port. Finally, the computer executes a code to order and to forward the data to the cloud server. Finally, this server delivers and updates data of the node's position to the web interface. Furthermore, a new data entry in the web interface generates and updates the map, which is used to show the most recent position of the tracking object to the final user.

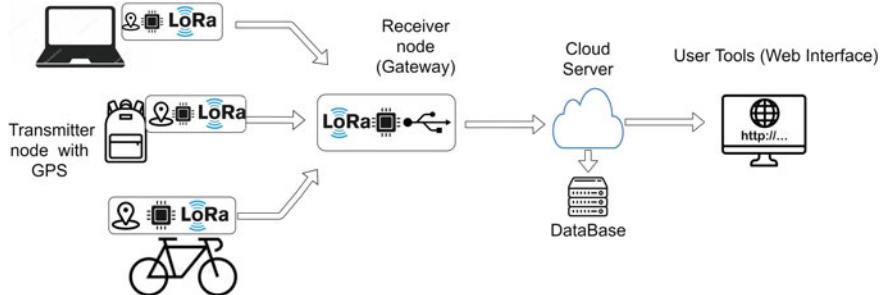


Fig. 1 Block diagram of the web interface

2.1 Basics of LPWAN and LoRa Networks

LPWAN networks are a type of technology oriented to low consumption and long range, with a low data transmission rate. This technology is being an increasingly adopted over the last few years in IoT solutions because of these characteristics. Furthermore, LoRa technologies guarantee long battery run-time and high network coverage [13, 14].

LoRa is a patented spread spectrum modulation scheme that is derived from Chirp Spread Spectrum (CSS) modulation. Furthermore, the signal spectrum is achieved by generating a pulse frequency modulated signal [15]. This modulation allows exchanging data rate for range or power consumption [16, 17]. LoRa modulation also includes a variable error correction scheme, called a Forward Error Correction (FEC) mechanism derived from Hamming (7, 4) encoding that adds three additional check bits to every four data bits in the message [18]. This mechanism allows the verification of the messages in the receiver without retransmission of the original information and is conditioned by the Coding Rate (CR), which can be varied according to the requirements of the system, since this also affects the time transmission and with this energy consumption. Other parameters that directly affect transmission in LoRa are the propagation factor or Spreading Factor (SF), which defines the amount of frequency modulations contained in each symbol. There are a total of six spreading factors (i.e., SF7 to SF12). A higher spreading factor produces longer over the air time and larger consumption [19]. It is also possible to configure other parameters such as the operating bandwidth (BW) and the transmission power (PWR).

3 Description of the Web Interface Architecture

The architecture of the proposed web interface is shown in Fig. 1. The cloud storage receives the incoming data to be sent to the website. It sorts and stores the information by stream processing and database sub-blocks, respectively. The web page block

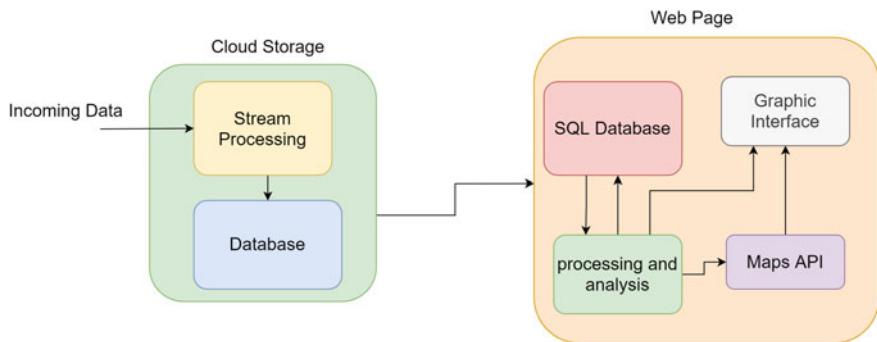


Fig. 2 Proposed system block diagram

Table 1 Table of chosen solutions

Sub-block	Solution	Plan type
Stream processing and database	Google firebase	Free plan (Spark plan)
Maps API	Mapbox	Open source
SQL Database	MySQL	General public license (Free)
Processing and analysis	PHP, JavaScript	Free
Graphic Interface	CSS, PHP	Free
Web hosting	125 mb.com	Free plan

performs the following processes, the data is sent to specific functions to be analyzed in the processing and analysis sub-block, data is visualized through the maps API sub-block, the control and customization of the application can be done in the graphic interface sub-block and user authentication through the SQL database.

As a cloud storage platform, Google Firebase service is chosen because it has a free plan that meets the requirements of the project, allowing real-time databases with 1Gb of storage, up to 10Gb in downloads of data per month, up to 100 simultaneous connections to the database and unlimited writes to the database. To display the maps, a JavaScript API provided by Mapbox was used. The maps available with this API could be customizable to the required user interface. These maps are fed by the information stored in the database (e.g., nodes position saved in variables in firebase). Additionally, Cascading Style Sheets (CSS) and Hypertext Preprocessor (PHP) are used to design the rest of the components of the web page, such as headers, menus, images, content for the map, buttons, and registration to the web page. The chosen software of the proposed low-cost web interface is summarized in Table 1.

Table 2 Table of general features

General Features	Description
Main Page	Allows access to the various network functions
Login and Registration System	This is to control the information that is presented, since it can be confidential content of the system and cannot be accessed by anyone
Application	A section dedicated to the final application with the possibility of having graphics, additional menus, or buttons to modify or visualize data from the sensor network
Information	Allows users to know what possibilities and parameters can be modified in the application

Table 3 Table of specific features

Specific Features	Description
Operating Range visualization	Shows the ranges for each node and constantly updates the location of the nodes when a change is detected in the coordinates coming from the cloud database
Alerts System	Warns the user when a tracking node goes out of range established
Hide/Show Button	Button to hide and show the operating range on the map application
Modify Center Input	Customization of the center of the operating range
Modify Radius Input	Allows to modify the radius of the range if the user enters a value
Node Device Selector	Allows the user to choose which tracker to modify

4 Web Interface Features

A web interface for object tracking applications requires general and specific features. The general features are the common features of IoT system web interface, the general features involved in the proposed low-cost interface were summarized in Table 2. In addition to these characteristics, specific features are added to the proposed web interface for node tracking applications, these features were summarized in Table 3. All these features were programmed and included on the web page, then they were tested and corrected.

Furthermore, the database functionality was verified, guaranteeing the security of the information of the users and the tracking sensors. In the debugging of the web interface, the features were tested first individually and then together. It is important to notice that more complements could be added to the application to improve tracking and user experience without a significant monetary investment in the software platform.

The web page of the interface must be scalable for several tracking nodes. For adding tracking nodes, the web developer could duplicate the implemented code with all the features with little modifications. Furthermore, this approach permits the customization of each node separately and can be repeated as many times as

necessary to manage even more nodes. The scalability only depends on the amount of data that can be stored and sent in Firebase in its free plan.

5 Main Specific Features

Due to limitations of the length of this paper, only the implementation of the main specific features will be described (i.e., Operating range and Alert system) in this section. They were chosen because little information on this topic has been found in the literature analyzed in this work.

5.1 *Operating Range*

This feature consists in to show the ranges for each tracking node and constantly updates its location in the map of the user interface. It was implemented by the function of JavaScript shown in the Appendix section A.1 Fig. 13. The inputs of the function are the center (CLat, CLong) and the radius (R) of the circle area defined by the user administrator as the operating range of the tracking node. The function must visualize the operating range correctly with any zoom level of the map provided by Mapbox. This visualization is implemented as a polygon (with 65 vertexes to look like a circle) generated by the function (e.g., dimension, center, color, opacity), each polygon vertex was calculated as coordinates (i.e., real latitude and longitude values) avoiding the visualization problem. The first vertex was calculated by

$$Coord_X = \frac{R}{111.32 * \cos(\frac{C_{Lat} * \pi}{180})} \quad (1)$$

$$Coord_Y = \frac{R}{110.57} \quad (2)$$

It is considered that a degree in units of longitude that can take values from -180° to 180° is approximately 110.32 km and due to the irregularities of the earth as it is not completely spherical and the values it can take is from -90° to 90°, one degree of latitude is approximately 110.57 km. After obtaining the coordinates of the first vertex, the coordinates of the other vertices were generated using the following equations:

$$Coord_{xn} = Coord_X \cdot \cos\left(\frac{2 \cdot \pi \cdot n}{V_{num}}\right) + C_{Long} \quad (3)$$

$$Coord_{yn} = Coord_Y \cdot \cos\left(\frac{2 \cdot \pi \cdot n}{V_{num}}\right) + C_{Lat} \quad (4)$$

where Vnum is the total number of vertices in the polygon (i.e., 65) and n is the vertex index (i.e., 0 to Vnum). The function is coded based on (1) and (2) and the explained process, generating the coordinates of each vertex of the polygon following (3) and (4), repeating this process for each vertex.

The overall vertices were stored in an array, which is accessed by Mapbox to draw the operating range. Additional to the process detailed before, the implemented function updates the operating range when it detects a change in the database. The process followed by the code to perform this process is listed below.

1. The tracking node coordinates are stored in the testlong and testlat variables
2. The number of vertices that the polygon will have is established to later obtain the coordinates of the center and radius of the operating range stored in the database in the cloud.
3. The trigonometric functions described in (3) and (4) are used, which is in charge of graphing the polygon according to the radius and center data.
4. The data of the vertices of the polygon vertex coordinates is sent to an array using the push method, which adds these coordinates to the array to be graphed by the map.
5. A polygon of type turf is created passing the points created in the previous step to a new variable, this allows using the turf library. Inside, check if the coordinates of the node are within the operating range.

Finally, the main function of the map is created, which is in charge of graphing on the programmed map and configuring the location of the node, operating ranges and markers for the user according to the established parameters.

5.2 Alert Tracking

The objective of this feature is to alert the user when any tracking node goes out of its operating range, allowing to notify and record the event so that the user can take actions and avoid the loss of the tracked object. It was implemented by the function of JavaScript in Appendix section A.2 Fig. 14. Furthermore, this feature is done by creating a circle with the same operating radius coordinates, but it is created under the “turf” function done in the operating range function described in Sect. 4.1, which has a special feature that allows us to compare whether a group of coordinates are within another group. With this working principle, the “turf.inside” function (i.e., returns a boolean value according to the comparison) is used to verify if the tracking node is within this range, the created function follows the steps described below.

1. The flag of the turf function is verified, where when the node is outside the operating range, the time and date of the event is recorded in the cloud database and the radius variation is started
2. The animation is produced, basically this is done by constantly varying the radius of the outer zone of the node in reduced spaces of time.
3. The process is replicated for the second node of the map, which is registered in another entry in the cloud database.

4. A smaller circle is created than the established operating range also using the turf function, and through the same process of comparing the node with the location of this circle, a warning system is established for the user, where it will be registered in the database of data in the cloud when the node is near the edge of the operating range as a warning.

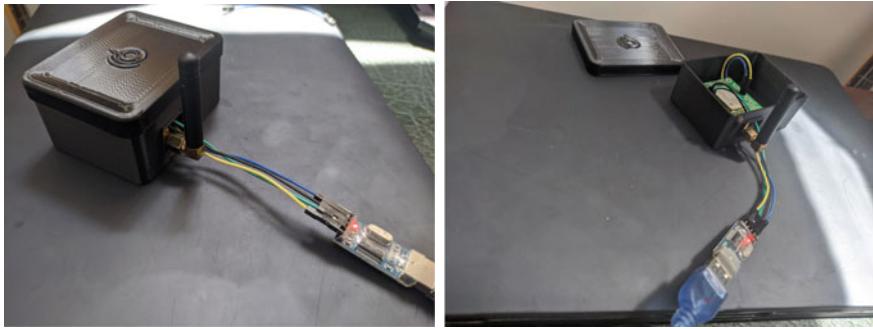
With this value stored in a variable, it is possible to run alert animations on the map and report the exact date of the event in the database. With a similar approach, warnings are added by creating a smaller circle, where if the node goes out of this range but is within the larger range, the warning is sent to the database. These warnings are recorded in the database when a sensor node is very close to the edge of the operating range.

6 Web Interface Results

The operating features were checked using an study case. This study case is based on a low-cost LoRa Network, a cloud server implemented in Google Firebase, and the proposed web interface. The study case architecture is shown in Figs. 1 and 2. To test the web page, first, the code was hosted on a local server to detect in an early stage of the development the errors in the code. Then, the code was hosted on a free web hosting (i.e., 125mb.com) with integrated MySQL database management. This approach allows the integration of the user information directly on the same hosting page.

6.1 Implemented LoRa Network

A low-cost LoRa Network was developed in a star topology with tracking nodes and a gateway, which was based on the Atmega328PB micro-controller. The photographs of a tracking node and the gateway are shown in Figs. 3 and 4. The tracking nodes must send the localization data of the tracking object to the network gateway. These nodes were designed to guarantee low power consumption with low localization error. The LoRa transceiver and the GPS selected to ac hives these design goals were the RN2903 and the Neo 6M, respectively. Furthermore, the LoRa link was configured with the regional parameters of the LoRa Alliance, LoRa WAN US902-928 [20], which are the SF to 9, BW to 125 KHz, CR to 1 and PWR to 20 or 18.5 dBm. Furthermore, the tracking node had a Kensington lock to guarantee easy attaching to the tracking object. Under this setup, the active and idle consumption of the tracking node were 194.5 ± 16.1 mW and 333.3 ± 28.55 uW, respectively. Moreover, the mean localization error was 7 m. The LoRa gateway was connected by serial port to a computer. This computer receives the position data from the transmitter node and through a Python code the coordinates are sent to the Google Cloud database Firebase.



(a) View 1

(b) View 2

Fig. 3 Transmitter node photograph

(a) View 1

(b) View 2

Fig. 4 Receiver node photograph

6.2 General Features Results

In a web application several user types exist, which differ in permissions and visualizations. In particular, for the proposed low-cost web interface, two types of users were defined, there are the administrators and supervisors' information. The administrator can modify all the characteristics of the operating range associated with each tracking node, besides, can view the alert history and create new users by establishing what level of permissions they have.

The supervisor can observe tracking nodes, but only has the possibility to hide and show the operating area and to see the alerts generated if the tracker leaves the operating range established by the administrator. The administrators and supervisors (i.e., email, password, and user role) are stored in the SQL database. The web interface requests this information using PHP commands to validate the users. The options available for the users are shown in Figs. 5 and 6.

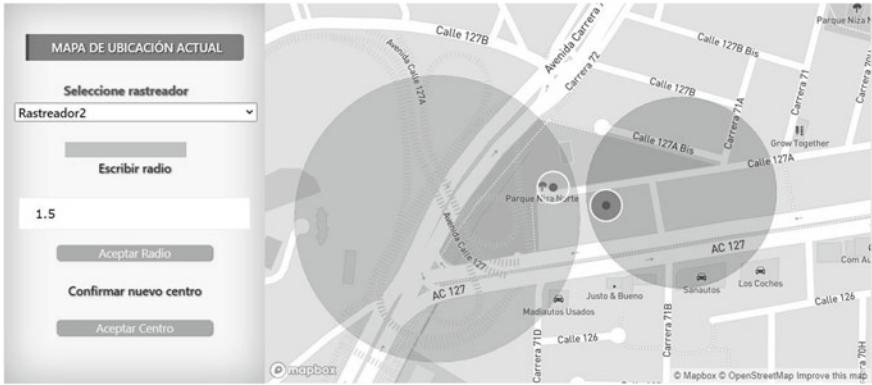


Fig. 5 Map view for the administrator user

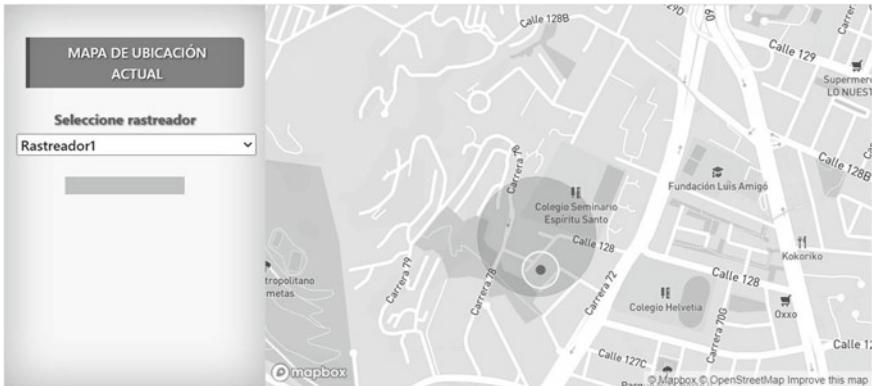


Fig. 6 Map view for the supervisor use

6.3 Specific Features Results

Buttons are added to simplify the interaction with the map of the web interface. They are shown in Fig. 7. Their functionalities allow to hide the operating range, change its center dragging the marker to another location, and change the radius in kilometers through the user prompt. In Fig. 8, the operating range visualization is detailed on the map. The proposed visualization technique based on polygons was verified with several values of radius and centers, measuring that the distance represented on the map is exactly equal to the entered value for an administrator user. Furthermore, the update of this range was checked too.

Additionally, the correct operating of the alert system is checked, observing that the alerts and animations are activated when the tracking node goes outside the established range. Figure 9 shows the correct operating of the palpitation animation when the tracking node is outside the established range.

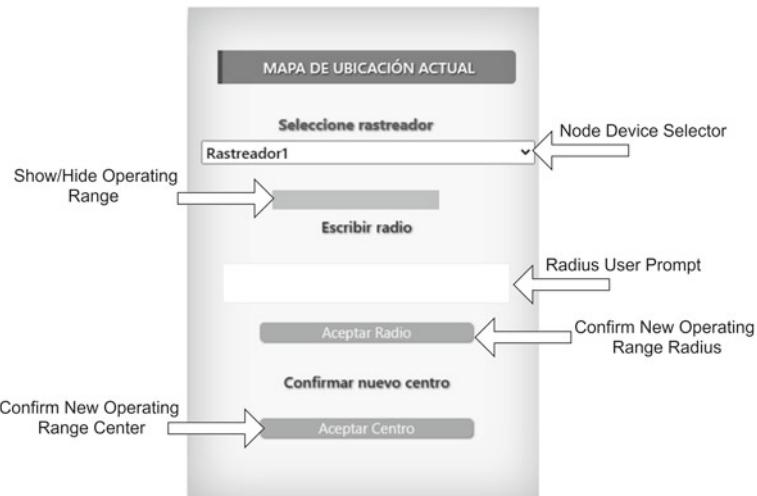


Fig. 7 Map customization menu



Fig. 8 Operating range graphed on the map

6.3.1 Hide/Show Button

This specific function of the system allows the user to hide or show the operating range of the selected node by means of a button, in order to give more comfort to it when observing the map and modifying the parameters of the operating ranges.

This function, being a button for interaction with the user, is programmed in using the PHP programming language and cascading style sheets on the web page in order



Fig. 9 Alert animation on the map

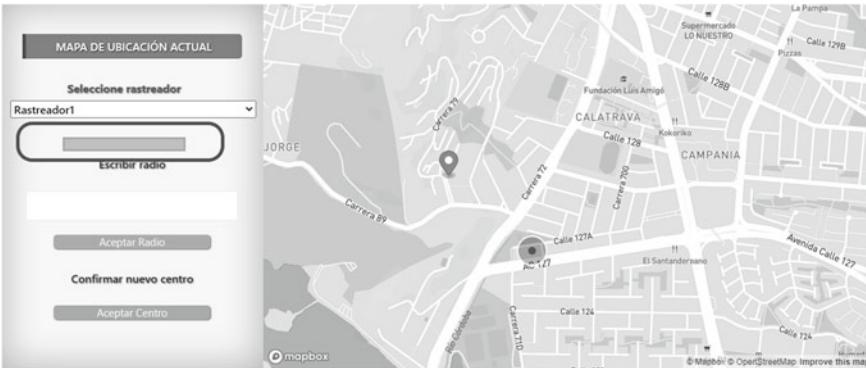


Fig. 10 Hide/Show operating range functionality

to be able to establish its dimensions and other visual characteristics. In addition to this, a function is created within the map code in JavaScript that checks the button state to show or hide the operating range based on the selected node (Fig. 10).

6.3.2 Entry to Modify Center of the Operating Range

This input for the user allows modifying the center of the operating range of a specific node. This is implemented by means of a tracker within the map which can be dragged to the new position of the center and with a confirmation button that accept the new one. Location.

The marker encoding is done within the map code in JavaScript, which creates the marker object with a preset appearance by Mapbox and adds a function that saves the position in the marker coordinates in variables when a change in position is detected original of this. For the confirmation button of the new center, it is done defining in the code destined for the application page in PHP and using Cascading Style Sheets (CSS) to add visual properties, then following the same procedure for the hide / show button, the state of the button is evaluated and as soon as it is pressed, the variables of the position of the marked are reviewed to establish the new center of the operating range (Fig. 11).



Fig. 11 operating range radius change functionality

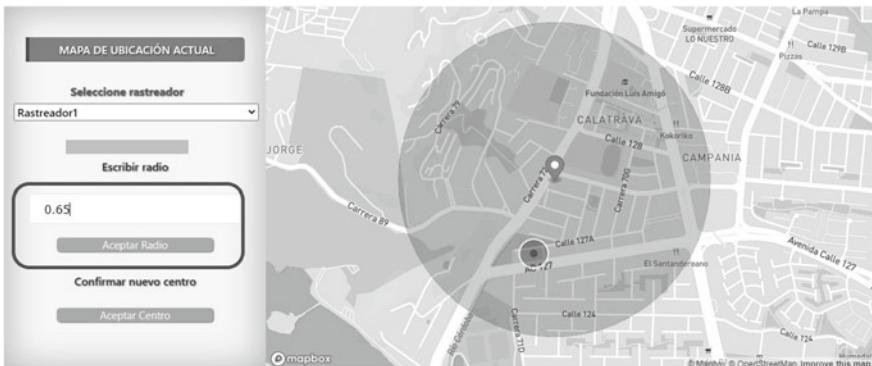


Fig. 12 Operating range center shift functionality

6.3.3 Entry to Modify the Radius of the Operating Range

The specific characteristic to modify the radius also consists of two parts, the first is the text box for the user so that the user can write the numerical value in kilometers of the operating range to modify and the second is a button to accept the entered value in the box. With regard to the programming for this feature, it is carried out in a process similar to the previous functions, where the properties of the objects are set in the sheet of the web page that is programmed in PHP and then the sent data is reviewed in the JavaScript map code to store the values of the objects in variables to modify the value of the radius of the operating range to activate a function that graphs the circumference again with the changes made (Fig. 12).

7 Conclusions

A low-cost web interface for object tracking was proposed, implemented, and tested. All the proposed functionalities were tested in a real LoRa network, and the results showed complete functionality of the developed codes. The more complex codes were appended to this paper to simplify the development of similar functionalities by the readers. The proposed approach helps to understand general object tracking system topology and offers a fast and easy design method to integrate the available technologies. We hope that this document and the support codes motivate engineers to integrate the low-cost cloud platforms in their projects.

Acknowledgements The authors would like to thank the Electronics Department and Electronics laboratory of the Pontificia Universidad Javeriana, for providing the required resources to conduct this study.

```

1 var createGeoJSONCircle = function(center, radiusInKm, points) { // Function to calculate the operating
    range
2   //step 1
3   testlong = Number(dataLong); // You can set input, if not, it has a default value
4   testlat = Number(dataLat); // The coordinates are passed to number
5   //step 2
6   if (!points) points = 64; // Vertices are set by default
7   var coords = {
8     latitude: LatAreaOf, // The coordinates of the operating range are saved
9     longitude: LongAreaOf
10  };
11  var ret = [];
12  var ret2 = [];
13  //step 3
14  Radiokm= Number(RadioKm);
15  var distanceX = Radiokm / (111.320 * Math.cos(coords.latitude * Math.PI / 180));
16  var distanceY = Radiokm / 10.574; // Relation between the distance in Km and the coordinates
17  console.log(km, ret);
18  var theta, x, y;
19  for (var i = 0; i < points; i++) { // The vertices of the polygon are graphed
20    theta = (i / points) * (2 * Math.PI);
21    x = distanceX * Math.cos(theta);
22    y = distanceY * Math.sin(theta);
23    x2 = (distanceX - 0.006) * Math.cos(theta);
24    y2 = (distanceY - 0.006) * Math.sin(theta);
25    //step 4
26    ret.push([coords.longitude + x, coords.latitude + y]); // The coordinates of each point are sent
27  }
28  ret.push(ret[0]);
29  //step 5
30  var polygon = turf.polygon([ret]); // Turf type polygon is created again using the vector obtained in
31  //ret
32  var point = turf.point([dataLong, dataLat]); // Point to evaluate type turf is created
33  var resultado = turf.inside(point, polygon); // Function that checks if it goes out of range
34  console.log(resultado);
35  if (resultado === false) { // If this is active the flag
36    cont = 1;
37    console.log(resultado);
38  } else {
39    cont = 0;
40  }
41  return { // The function returns a string for the polygon to be graphed
42    "type": "geojson",
43    "data": {
44      "type": "FeatureCollection",
45      "features": [
46        {
47          "type": "Feature",
48          "geometry": {
49            "type": "Polygon",
50            "coordinates": [ret]
51          }
52        }
53      ];
54    };
55  };

```

Fig. 13 Operating range code

Appendix

7.1 Operating Range Code

See Fig. 13.

7.2 Alerts System Code

See Fig. 14

```

1 setInterval(() => { // Function that performs the alert animation
2     map.setPaintProperty('rangosf', 'circle-radius', radius);
3     map.setPaintProperty('rangosf2', 'circle-radius', radius2);
4     var Radiof = Number(Radio);
5     var Radiof2 = Number(Radio2);
6
7     //step 1
8     if (cont == 1) // If the flag is up it makes animation
9     {
10         //step 2
11         radius = 3 + radius % 30; // Ratio of change and maximum animation range.
12         // Fires when the first node is found out of range
13         if (AlertaFlag == 1)
14         {
15             hoyFecha();
16             AlertaFlag = 0;
17             var ref = database.ref('Alertas/Rastreador1');
18             var data = {
19                 time: Fecha
20             }
21             ref.push(data); // The value of the date is written to the database
22         }
23     } else {
24         radius = Radiof // Default red zone radius
25         AlertaFlag = 1;
26     }
27     // If the flag is high, register the warning of the first node
28     if ((cont2 == 1) && (cont==0))
29     {
30         if (AlertaFlag2 == 1) // Is activated when out of range
31         {
32             hoyFecha();
33             AlertaFlag2 = 0;
34             var ref = database.ref('Advertencias/Rastreador1');
35             var data = {
36                 time: Fecha
37             }
38             ref.push(data); // The date value is written to the database
39         }
40     } else {
41         AlertaFlag2 = 1;
42     }
43 }, 50);
44

```

Fig. 14 Alerts system code

References

1. Priyadarshi, R., Gupta, B., Anurag, A.: Deployment techniques in wireless sensor networks: a survey, classification, challenges, and future research issues. *J. Supercomput.* (2020)
2. Lai, P., Huang, H., Sheu, M., Wu, C., Le, J., Chen, T.: Bike sensor system design for safety and healthy riding. In: 2018 IEEE International Conference on Consumer Electronics Taiwan (ICCE-TW), pp. 1–2, 1 (2018)
3. Kim, D.H., Park, J.B., Shin, J.H., Kim, J.D.: Design and implementation of object tracking system based on LoRa,” in 2017 International Conference on Informed Network (ICOIN), pp. 463–467 (2017)
4. Zourmand, A., Hing, A.L.K., Hung, C.W., Abdul Rehman, M.: Internet of things (IoT) using LoRa technology. In: 2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS), pp. 324–330 (2019)
5. Samie, F., Bauer, L., Henkel, J.: IoT Technologies for Embedded Computing: A Survey (2016). <https://doi.org/10.1145/2968456.2974004>
6. Croce, D., Garlisi, D., Giuliano, F., Valvo, A.L., Mangione, S., Tinnirello, I.: Performance of LoRa for bike-sharing systems. In: AEIT International Conference on Electrical and Electronic Technologies for Automotive (AEIT AUTOMOTIVE), pp. 1–6 (2019)
7. IEEE Standard for Local and metropolitan area networks—Part 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs). IEEE Std 802.15.4-2011 (Revision of IEEE Std 802.15.4-2006). pp. 1–314 (2011)
8. Prando, L.R., de Lima, E.R., de Moraes, L.S., Hamerschmidt, M. B., Fraindenraich, G.: Experimental performance comparison of emerging low power wide area networking (LPWAN) technologies for IoT. In: 2019 IEEE 5th World Forum on Internet of Things (WF-IoT), pp. 905–908 (2019)
9. Wang, S.-Y., Zou, J.-J., Chen, Y.-R., Hsu, C.-C., Cheng, Y.-H., Chang, C.-H.: Long-Term Performance Studies of a LoRaWAN-Based PM2.5 Application on Campus (2018)
10. Wang, S.-Y. et al.: Performance of LoRa-Based IoT Applications on Campus (2017)
11. Lian, G., Yu, F.: A data platform based on web service for wireless sensor network. In: 4th IEEE International Conference on Information and Computer Technologies, pp. 670–673 (2014)
12. Toomey, C.: Asset Tracking Solution, GitHub (2019). <https://github.com/mapbox/asset-tracking>
13. Samie, F., Bauer, L., Henkel, J.: IoT Technologies for Embedded Computing: A Survey (2016)
14. Nolan, K.E., Guibene, W., Kelly, M.Y.: An evaluation of low power wide area network technologies for the Internet of Things. In: Proceedings of the IEEE International Wireless Communications and Mobile Computing, pp. 439–444 (2016)
15. LoRa Developers: <https://lora-developers.semtech.com/library/tech-papers-and-guides/lora-and-lorawan/>. What are LoRaR and LoRaWANR?
16. LoRa SX1272/73 Datasheet. Semtech, Mar 2015. <http://www.semtech.com/images/datasheet/sx1272.pdf>
17. LoRa SX1276/77/78/79 Datasheet, Rev. 4. Semtech (2015)
18. Elshabrawy, T., Robert, J.: Evaluation of the BER performance of LoRa communication using BICM Decoding. In: 2019 IEEE 9th International Conference on Consumer Electronics (ICCE-Berlin), Berlin, Germany, pp. 162–167 (2016). <https://doi.org/10.1109/ICCE-Berlin47944.2019.8966172>
19. Semtech: AN1200.22 LoRaTM Modulation Basics (2015). <http://www.semtech.com/images/datasheet/an1200.22.pdf>. Semtech
20. LoRaWAN™ 1.0.3 Regional Parameters—LoRa Alliance. https://lora-alliance.org/wp-content/uploads/2020/11/lorawanRegionalParameters_v1.0.3revA_0.pdf

Sentiment Analysis: Choosing the Right Word Embedding for Deep Learning Model



Sarita Bansal Garg and V. V. Subrahmanyam

Abstract Sentiment analysis is a field that helps data analysts to gauge public opinion, observe brand and product reputation, perform nuanced market research, and understand customer experiences. **It is the process by which whether a particular text is positive, negative or neutral can be determined.** This process is a combined application of techniques of natural language processing and artificial intelligence to assign weighted sentiment scores to the entities within a sentence or piece of text. Word embedding or the feature vectors are the vectors that can represent the text into the vector space. Embeddings solve the problem of representing very large non-sparse vectors into a lower-dimensional space. For deep learning models, it is very necessary to input the text data as feature vectors. Various methods of generating feature vectors are explored. The paper mainly focuses on the application of various pre-trained word embedding on deep learning model and compares them through various metrics calculated on them. The results obtained showed that Word2Vec outperformed other word embeddings.

Keywords Artificial Intelligence · Deep learning · Sentiment analysis · Word embedding · Count-Based embedding · Prediction-Based embedding · Word2Vec · GloVe · FastText

1 Introduction

Sentiment analysis is a field that aims at analysing the views or opinions expressed by the users and classifies them for further processing. It is a very big umbrella that can accommodate many applications, like social media monitoring, customer support and feedback, brand monitoring, product analysis, market research and analysis, etc. This

S. B. Garg ()
Maharaja Agrasen Institute of Management Studies, Delhi, India
e-mail: saritabansalgarg.faculty@maims.ac.in

V. V. Subrahmanyam
School of Computer and Information Sciences, IGNOU, New Delhi, India

field is closely associated with natural language processing (NLP) as all the views or opinions with the aim of categorizing them are all expressed in natural languages used by humans which is highly unstructured. So, the first and foremost step in any sentiment analysis task is applying certain NLP techniques, like tokenization, part-of-speech (POS) tagging, parsing, etc. to the data (reviews) before processing and classifying them. After this pre-processing step, any supervised or unsupervised methods can be applied for the classification. In this paper, the supervised method of deep learning was considered for the same. This paper is not focused on deep learning methods but on its pre-requisite for the successful application of the methods.

Deep learning methods are the deep neural networks that work on digital data. As already specified, sentiment analysis considered unstructured text as its input which cannot be feed directly into the system or the model. So, there is a need to pre-process that text and convert it into numeric/digital form or vectors to be more specific. Generating word embedding for the same is one such popular method.

Word embedding refers to a collection of techniques of language modelling and feature learning in the field of natural language processing (NLP). These techniques are applied to words or phrases from the vocabulary to generate vectors of real numbers corresponding to each word/phrase. With the help of word embedding, the meaning of a word in a document can be easily captured along with the syntactical and semantical relation of that word with other words. This paper is an attempt to present the various word embedding methods which can be used to represent the text and choosing the best embedding for the task of sentiment analysis. Due to the limitation of computing resources and time scarcity, only the pre-trained embedding was considered for generating vectors corresponding to the text data. Here, the standard dataset of Stanford Sentiment Treebank was considered for evaluating the different embedding.

The paper is divided into certain sections. The background section discusses different ways for generating feature vectors along with the main two categories of word embedding. How the popular word embedding was developed and generated was also discussed. In the related study section, a similar type of studies was identified and discussed which was the motivation behind the present study. Then the model was proposed for the present study with the experimental setup which was used for the present study. The next section is the result and analysis where the outcomes/results were discussed and analysed to reach for a conclusion and suggested future work.

2 Background

Reference [1] defined feature encoding or feature extraction as the process of converting the text data into vectors such that the various linguistic properties of the text were retained. For NLP-related tasks it is very important to preserve these properties. One of the oldest used methods for feature extraction was bag of words (BOW) method. As the name indicates, this method considers the text or document as simply a bag full of words. This method was only concerned about simply the

presence or absence of words in the document irrespective of its position and order or relation of words with respect to other words in the document. This method results in the very sparse vector representation of text which requires large memory for storage, more computation resources for processing and makes the task of modelling the data all the more challenging. All this gets more complicated with the increase in the size of the corpus. Various modifications are suggested for the same which includes more text pre-processing or instead of considering single word, considering n-gram of words or how many times each word appears or the frequency for each word or term frequency-inverse document frequency (TF-IDF) method. But the major disadvantages of using these representations were that they were in direct proportion to the vocabulary size, did not capture word's position and semantics in text, co-occurrences of words in the same / different documents and consider them all to be independent, etc. That is why, these representations were only useful as a lexical level feature. For the task of sentiment analysis, it is very necessary to consider the semantics of the words to accurately and fairly assess the sentiments of the reviews.

Another method most popularly used for vector representation is one-hot encoding. This method is said to be the traditional method of vectorization for NLP which is used to represent the categorical data. Each word/token of the text is considered as an independent category. Here, the length of vectors equals the total number of unique words in the vocabulary/corpus considered for the dataset with one binary vector generated for every word in the review or dataset. The vectors or matrix generated as a whole is said to be very sparse and high-dimensional with only one bit set as 1 for the index position of word in the vocabulary and all other bits are set as 0. Again, this representation also is not positional and does not consider the co-occurrence of words. So, one-hot representation is not a powerful and useful method of representation of text. But this representation is a preferred choice for representing labels in case of supervised data as labels are limited and can be represented as equidistant categories of labels independent of each other.

The search is for a method that gives dense and low-dimensional vectors for representation. These vectors will be highly computational and have generalization power. Word embedding is one such method for learned representation for text. It is a name collectively given to a set of feature learning techniques and language modelling methods in NLP. Here, words with similar meaning got similar representation in a predefined vector space with similar vectors. A single real-valued vector was generated and used to represent and map every individual word in the vocabulary. The values of these vectors are learned through a process similar to a neural network. Because of this, the technique is associated with the field of deep learning.

Depending upon the process of derivation of word vectors, embedding models are divided into two main categories. Models which derive vectors based on neural network language model are known as prediction-based models and those which are matrix-based are known as count-based models. The popular Word2Vec word embedding model belongs to the first class, whereas GloVe belongs to the latter [2]. Both types of models depend on the same assumption that words used in similar contexts have the same meaning and consequently they must be represented with

similar vectors in the vector space. This concept was known as the distributional hypothesis and was given by [3], along with others.

2.1 Prediction-Based Models

Neural Net-based large-scale language model was first developed by [4]. For the first time, to develop this language model, they integrated the vector space model (VSM) with the language model which was a statistical model of language usage. They want to reframe the problem and state it as a problem of unsupervised learning. This neural network language model produced embedding as a by-product. Here, raw word vectors were first projected onto an embedding layer and then they were being passed into the remaining layers of the network. Amid various other reasons, for language models, this helped in easing the effect of the curse of generalization and dimensionality. But the major problem with the proposed model is the computation cost induced due to the normalization factor required by the SoftMax output layers.

An improvised model was suggested in [5], where using importance sampling [6], gradients were estimated on randomly selected examples from the vocabulary having a supplementary distribution (e.g. old n-gram language models). In terms of perplexity, in training time, as compared to the previous models, they reported gains of a factor of 19. Reference [7] further modified the model by replacing and implementing the binary tree approach which was more efficient than implemented full SoftMax prediction. In the modified model, decisions need to be made only at each node of the binary tree which ultimately led to the target word. At a slightly lower score of perplexity, during training time a speedup of over 3 times and 100 times during testing was reported.

In [8] the model was tested with the log-bilinear model, which was a feedforward neural network with a linear, single hidden layer and this simple model is much faster and outscore the model proposed in [4]. As suggested in [7], using hierarchical SoftMax the authors of [9] trained the log-bilinear model. The only difference is instead of using the word tree from external sources (WordNet), they learned it during the process itself which results in a 200 times faster model.

In all the models discussed so far, word embedding was just the byproduct, and the main objective was to develop some language models. It was the study in [10] that focused solely on generating word embedding. For this, they suggested two improvements over the previous models. They used the word's full contexts, i.e., both left and right contexts to predict the centre word along with expanding the dataset with false or negative examples. References [11, 12] contributed majorly in developing and deriving the prediction-based word embedding. A two-step method for bootstrapping a neural network language model was suggested by [11] which form the basis for future models. Here, using a single word as context, the model was trained and the initial embeddings were generated that were used in the second step to train the full model with a larger context. Recurrent neural network (RNNs) was used in [12] which instead of deciding beforehand that on either side how many

words need to be used as context, helped the model to remember arbitrarily long contexts. These modifications helped in reducing the training time of the models.

It was Mikolov, who through [13, 14] and [15] gave a major breakthrough in the field of prediction-based word embedding and drew the attention of NLP researchers towards this. His efforts are popularly known as Word2Vec embedding which is one of the most used embeddings. Taking [12] as a base for their research, [13] discovered the syntactic and semantic regularities in the data, and several common associations such as singular–plural, male–female, family relations like parents–children (king–prince) etc. actually correspond to arithmetical operations which can be performed on word vectors. Reference [14] introduces two new log-linear models of Word2Vec, namely continuous bag of words (CBOW) and Skip-Gram (SG) based on [11] methodology, where CBOW predicts the target centre word given the adjacent context words (both left and right), SG predicts the opposite, i.e., predicting surrounding target words when the centre context word was given. Instead of subsampling of frequent words and hierarchical SoftMax, [15] improvised the model by including the negative sampling. This accelerated the training of the model and helped in reducing the amount of noise due to excessively frequent words. With faster training time, Skip-Gram with negative sampling (SGNS) is the best performing variant of Word2Vec.

An improvement was suggested by [16] over the [15] SGNS to learn n-gram embeddings. These embeddings can be used to compose the words later on if required. There are languages, like Turkish, Finnish etc., which rely heavily on morphological and compositional word building and some information is encoded in the word parts themselves. So, this rationale helped in generalizing the unseen words. Embedding developed here is known as FastText and it reports better results than SGNS.

2.2 Count-Based Models

These are another category of models which can be employed for producing word embedding. These models represent word embedding as word-context matrices [17] which were often generated by globally grasping the counts of co-occurrence for similar context words in a corpus. Often, [17] represented these counts by word-context matrices. Reference [18] was the earliest appropriate example in the field of information retrieval where latent semantic analysis (LSA) was introduced to produce word-context matrices. Here, the term-document matrix was applied by singular value decomposition (SVD). By looking across the rows, word vectors can be retrieved from the factorized matrix. The hyperspace analogue to language (HAL) method was introduced by [19]. With a context window size of 8, the co-occurrence count between the target word and each context word was calculated by HAL which was inversely proportional to the distance between that context word and the target word. In spite of the good results produced, the main problem with HAL method was that very common words like ‘a’, ‘an’, ‘the’ contribute disproportionality to all co-occurrence words. To overcome this problem, [20] introduced COALS method. In

this method, conditional co-occurrence was considered instead of frequency difference in words which were factored out by incorporating the normalization strategies. Using SVD factorized variant, it reports better results than [18, 19].

The low-rank multi-view learning (LR-MVL) method was introduced by [21]. It was an iterative algorithm where canonical correlation analysis (CCA) was used to derive the embeddings [22] between the right and left contexts of the given word. Over various NLP tasks, this method reported better results over other factorization methods as well as neural embeddings. A modified version of principal component analysis, i.e., Hellinger PCA was applied to the word-context matrix by [22]. This reported even better results than [21] and neural embeddings like [9, 10].

Reference [23] introduced GloVe model. It was a log-linear model. This model was based on the insight that instead of raw counts of word-context co-occurrence, their ratios can better convey the word meaning. This model was trained to encode semantic relationships between words as vector offsets in the learned vector space. As compared to other prediction-based and count-based models [15], this method reported better results for tasks such as named entity recognition (NER) and word analogy.

3 Related Study

Reference [24] evaluated the impact of word embeddings in sentiment analysis. They explored the effect of five factors on the performance of a sentiment analysis classification system. Those five factors were the size of the corpus, text domain, learning method, method to combine the word vectors and identification of the most important words in terms of POS. They tested using two types of word embeddings, i.e., Word2Vec and FastText. They performed several classification experiments to study the effect of a sentiment classifier using different datasets and methods. Results showed that word embeddings gave very positive results for word-based word embeddings based on large size corpus. This confirmed our approach to explore the effect of using different word embeddings for our task of sentiment classification.

In [25], the authors trained a recursive deep neural network (LSTM) using GloVe pre-trained word embedding on Sentiment TreeBank. The classification accuracy was calculated with different variations of word feature vectors and model architectures. For binary classification, the test accuracy of 0.847 was reported which was the highest among all the models tested. The model was trained for 100 epochs. Here, the study was limited to GloVe pre-trained embedding and comparison with other word embedding was not discussed.

Word2Vec pre-trained word embedding was explored by [26]. In their experiment, the authors explored the classification process of the dataset as a convolutional neural network and trained the model for 25 epochs. They tested the Stanford Sentiment dataset both for binary classification and five-class classification. For five-class classification model, they obtained below-average classification accuracy but for binary classification, the accuracy obtained was moderate (around 84%) but still low. Again,

the concentration is on a single-word embedding. In previous study similar accuracy was obtained for GloVe embedding for 100 epochs, but here it was obtained in just 25 epochs.

Reference [27] experimented with the word embedding model based on Word2Vec along with LSTM method for short text sentiment classification. They tested how combining word embeddings with deep neural networks increases the efficiency of classifiers and gave better results as compared to other classifiers. They compared their approach with Naïve Bayes (NB) and extreme learning machine (ELMs). For the experimentation purpose the authors used three datasets. First was English movie reviews from IMDB, the second review about Chinese movies from Douban and third was Chinese posts in the PTT discussion forum. They concluded that using word embedding models, the sentiments of short texts can be effectively classified, and given more number of training data, deep learning methods like LSTM show better performance of sentiment classification.

Reference [28] focused on a similar problem. But instead of using pre-trained word embedding they derived the word embedding and experimented with the different versions of Word2Vec, Glove and FastText. The models considered were independent models of LSTM, CNN and FastText and tested only for five epochs. For Stanford Sentiment Treebank, they reported average classification accuracy of around 67% across all the samples.

Focusing on the Stanford Sentiment Tree Bank, the classification accuracy obtained so far was the motivation to experiment with the pre-trained word embedding. There is a lot of discrepancies in terms of the number of epochs also, where one study is training the model for 100 epochs, while other does it for only five epochs. It is a known fact that in deep neural network a lot depends on the model architecture and hyper-parameters but one cannot ignore the input feed into the model. Obtaining the right feature vectors also plays a crucial role. So, exploring and testing different feature vectors obtained through pre-trained word embedding and increasing the test accuracy for the binary sentiment analysis and classification are the focus of this paper.

4 Proposed Model

For similar words, different vector representations were generated by the different word embedding models. A good representation should exhibit few important properties. The same word can be used in different contexts. Word embedding models must point out this difference in contexts and these distinctions must be encoded into meaningful representations in the word subspace [9]. Next, the model should be able to distinguish the meaning of a word from its context and find the relevant embedding representing the syntactic and semantic property of that word. Also, the results of a word embedding model should be reliable [23] and the geometry of an embedding space should have a good spread with frequent and rare words [29]. So, embedding can be evaluated using either intrinsic evaluation methods [30] or extrinsic evaluation

Table 1 Details of pre-trained word embedding used

Embedding Name	Developed by	Dataset	Dimensions of vectors considered
Word2Vec	Google	Google News Dataset	300
GloVe	Stanford	Wikipedia 2014	300
FastText	Facebook AI	Wikipedia 2017, statmt.org news dataset and UMBC web base corpus	300

methods [31]. Where the intrinsic evaluators are general and include word analogy, word similarity, outlier detection and concept categorization, extrinsic evaluators are more task-specific such as POS tagging, named entity recognition or sentiment analysis. Since the focus of the present study is sentiment analysis, so extrinsic evaluator (sentiment analysis) was considered, as the results of the analysis will itself display how appropriately the word embedding model had represented the word vectors.

For experiment purposes, the three most popular unsupervised pre-trained word embedding models were considered: GloVe, Word2Vec and FastText (Table 1). Pre-trained word embedding is the embedding learned in one task and stored and later on used for solving another similar task. The purpose behind using the pre-trained word embedding is the use of very large datasets (which can accurately capture the syntactic and semantic meaning of words) by the developers in developing these embedding as with our limited computing resources (for a large number of trainable parameters) and the specific sparse dataset it would be difficult for us to develop embedding for our dataset and also, using the pre-trained embedding will give a pinch of generality in the experiment. Using pre-trained word embedding is also known as a form of transfer learning where we are transferring the embedding from one task to another.

4.1 Experimental Setup

For the experiment purpose, the Stanford Sentiment Treebank dataset was used. This dataset consists of 11,855 labelled movie reviews from Rotten Tomatoes. The dataset is further split into 8544, 1101 and 2210 reviews for train, dev and test sets, respectively. The dataset is available in both tree form and phrases. Reviews are labelled on five-point (0–4) scale with very positive, positive, neutral, negative and very negative labels. For the ease of experiment, initially the problem of sentiment classification is considered as a binary classification and the labels are converted from five-point scale to two-point where labels greater than or equal to 2 are considered as positive and less than 2 as negative. Data is pre-processed using standard nltk and nlp-preprocessor libraries. Word embedding matrix is generated using the pre-trained word embedding.

To give fair chance to all the word embedding, the pre-trained models are tested on the same deep neural network. The focus of the experiment was to develop a binary classifier that gives probability about whether the review is a positive review or a negative review. The model developed was a small RNN models with embedding layer as the first layer of the model using feature/word vectors derived from the embedding matrix as input and five other layers consisting of LSTM, two dense layers, flatten layer along with a dropout layer of 50% to avoid overfitting. The whole experiment was conducted on Google Colab. While training the model it was observed that the model tends to overfit the data. So, to avoid overfitting, we apply L1_L2 regularization technique. Models are trained for different number of epochs to find out the optimum number of epochs. Using loss metric and plot, it is observed that model gives the best results in 10 epochs. For a large number of epochs, lot of variation is observed in the loss. So, the number of epochs is fixed to be 10. Various metrics were calculated for the trained models, like confusion matrix, accuracy, precision, recall, f-score along with the accuracy and loss curves to evaluate the three models.

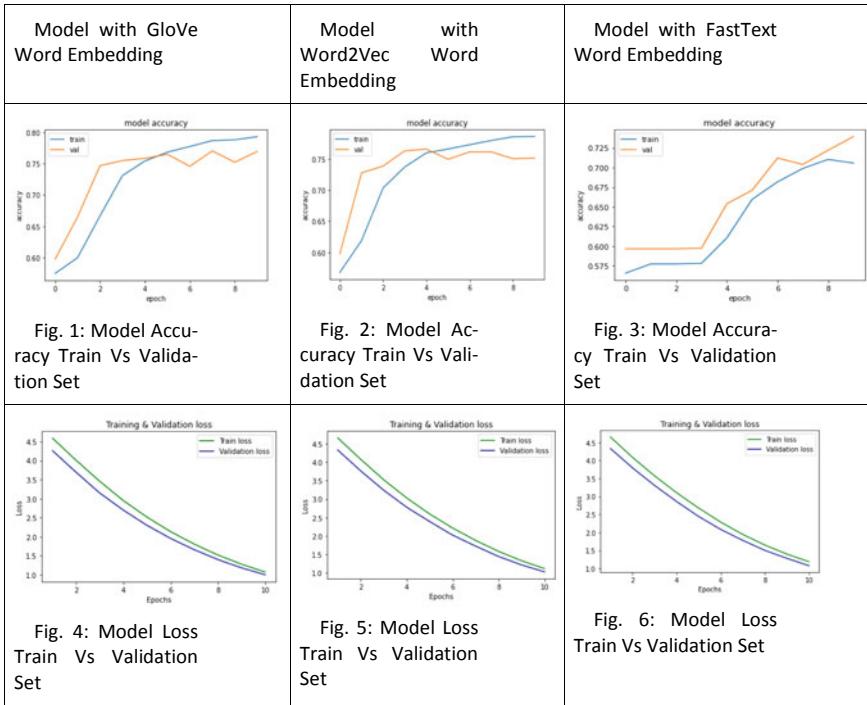
5 Results and Analysis

Table 2 gives the different measures for the experiment conducted for all three types of word embeddings. Also, the learning curve showing the accuracy and loss for all three models is displayed in Table 3.

Looking at the statistics obtained by performing the experiment, it is summarized that model gives very competitive results for both Word2Vec and GloVe-based word embedding. As both Word2Vec (0.7787) and GloVe (0.7688) gave similar and approximately equal accuracy, we need to consider other metrics. So, precision and recall score for both the embeddings were compared. Precision scores for both the GloVe (0.7231) and Word2Vec (0.7071) are almost similar but shows a difference of almost 8% in their recall score (0.7096 for GloVe and 0.7888 for Word2Vec). This implies where both GloVe and Word2Vec were predicting the same number of positive sentiments out of all the predicted positive sentiments, Word2Vec was predicting a greater number of positive sentiments out of actual positive sentiments.

Table 2 Results of sentiment analysis using a different word embedding

	GloVe		World2Vec		FastText	
Confusion_matrix	1054	247	1004	297	1096	205
	264	645	192	717	378	531
Accuracy	0.7688		0.7787		0.7362	
Precision	0.7231		0.7071		0.7215	
Recall	0.7.96		0.7888		0.5842	
F-score	0.7163		0.7457		0.6456	

Table 3 Plots for model accuracy and loss using a different word embedding

Also, in terms of F-measure, the verdict goes in the favour of Word2Vec (0.7457) as compared to GloVe (0.7163).

Next, the performance and optimizing learning curves were also considered for all three models. Loss curves for all three models are similar and optimizing the loss very effectively. Accuracy curves indicate that models are learning the parameters and is effective more for Word2Vec. So, statistical results go in favour of Word2Vec word embedding.

6 Conclusion and Future Work

Using pre-trained word embedding was a very convenient way of generating the word vectors required to apply deep learning models. These embeddings were developed on the text corpus containing millions of words with high-end computing resources. So, instead of generating the embeddings from scratch using limited computing resources, the use of pre-trained and generated embedding was preferred. Out of all available pre-trained word embedding, here the concentration was on the three most popular word embedding models and tested them for sentiment analysis on movie

reviews from Stanford Sentiment Treebank. Experiments showed that Word2Vec gave very promising results for the problem.

The task of sentiment analysis is incomplete without considering negations as the occurrence of negation terms both implicit and explicit reverses the polarity of the reviews. So, when the decision on which word embedding to use is taken, now the concentration is on the model architecture to obtain the best optimum results along with negation handling. Even though the Word2Vec took care of the input weights for the model, various model architectures from recurrent to recursive neural networks need to explored and tested upon.

References

1. Goldberg, Y.: Neural network methods for natural language processing. *Synth. Lect. Hum. Lang. Technol.* **10**(1), 1–309 (2017)
2. Almeida, F., Xexéo, G.: Word Embeddings: A Survey (2019). [arXiv:1901.09069](https://arxiv.org/abs/1901.09069)
3. Harris, Z.S.: Distributional structure. *Word* **10**(2–3), 146–162 (1954)
4. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. *J. Mach. Learn. Res.* **3**, 1137–1155 (2003)
5. Bengio, Y., Senécal, J.S.: Quick training of probabilistic neural nets by importance sampling. In: AISTATS, pp. 1–9, Jan 2003
6. Doucet, A., De Freitas, N., Gordon, N.: An introduction to sequential Monte Carlo methods. In: Sequential Monte Carlo Methods in Practice, pp. 3–14. Springer, New York, NY (2001)
7. Morin, F., Bengio, Y.: Hierarchical probabilistic neural network language model. In: Aistats, vol. 5, pp. 246–252, Jan 2005
8. Mnih, A., Hinton, G.: Three new graphical models for statistical language modelling. In: Proceedings of the 24th International Conference on Machine learning, pp. 641–648, June 2007
9. Mnih, A., & Hinton, G. E.: A scalable hierarchical distributed language model. In: Advances in Neural Information Processing Systems, pp. 1081–1088 (2009)
10. Collobert, R., Weston, J.: A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning, pp. 160–167, July 2008
11. Mikolov, T., Kopecky, J., Burget, L., Glembek, O.: Neural network based language models for highly inflective languages. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4725–4728. IEEE (2009)
12. Mikolov, T., Karafiat, M., Burget, L.: Jan Černocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In: Eleventh Annual Conference of the International Speech Communication Association, pp. 1045–1048 (2010)
13. Mikolov, T., Yih, W. T., Zweig, G.: Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies (pp. 746–751), June 2013
14. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space (2013). [arXiv:1301.3781](https://arxiv.org/abs/1301.3781)
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).
16. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of Tricks for Efficient Text Classification (2016). [arXiv:1607.01759](https://arxiv.org/abs/1607.01759).
17. Turney, P.D., Pantel, P.: From frequency to meaning: Vector space models of semantics. *J. Artif. Intell. Res.* **37**, 141–188 (2010)

18. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **41**(6), 391–407 (1990)
19. Lund, K., Burgess, C.: Producing high-dimensional semantic spaces from lexical co-occurrence. *Behav. Res. Methods Instrum. Comput.* **28**(2), 203–208 (1996)
20. Rohde, D.L., Gonnerman, L.M., Plaut, D.C.: An improved model of semantic similarity based on lexical co-occurrence. *Commun. ACM* **8**(627–633), 116 (2006)
21. Dhillon, P., Foster, D.P., Ungar, L.H.: Multi-view learning of word embeddings via cca. In: *Advances in Neural Information Processing Systems*, pp. 199–207
22. Lebret, R., Collobert, R.: Word Emddeddings Through Hellinger PCA. [arXiv:1312.5542](https://arxiv.org/abs/1312.5542)
23. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Oct 2014
24. Petrolito, R., Dell'Orletta, F.: Word embeddings in sentiment analysis. In: *CLiC-it*, Dec 2018
25. Hong, J., Fang, M.: Sentiment analysis with deeply learned distributed representations of variable length texts. Stanford University Report, pp. 1–9 (2015)
26. Mandelbaum, A., Shalev, A.: Word Embeddings and Their use in Sentence Classification Tasks (2016). [arXiv:1610.08229](https://arxiv.org/abs/1610.08229)
27. Wang, J.H., Liu, T.W., Luo, X., Wang, L.: An LSTM approach to short text sentiment classification with word embeddings. In: *Proceedings of the 30th Conference on Computational Linguistics and Speech Processing (ROCLING 2018)*, pp. 214–223, Oct 2018
28. Wang, B., Wang, A., Chen, F., Wang, Y., Kuo, C.C.J.: Evaluating word embedding models: methods and experimental results. *APSIPA Transactions on Signal and Information Processing*, vol. 8 (2019)
29. Hellrich, J., Hahn, U.: Don't get fooled by word embeddings-better watch their neighborhood. In: *DH*, Aug 2017
30. Gladkova, A., Drozd, A.: Intrinsic evaluations of word embeddings: What can we do better?. In: *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 36–42, Aug 2016
31. Schnabel, T., Labutov, I., Mimno, D., Joachims, T.: Evaluation methods for unsupervised word embeddings. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 298–307, Sept 2015

Detection and Localization of Unmanned Aerial Vehicles Based on Radar Technology



Sally M. Iddis, Takwa Dawdi, Qassim Nasir, Manar Abu Talib, and Yara Omran

Abstract Ever since unmanned aerial vehicles (UAV), also known as drones, became available for civilian use, their risks increased. Along with that, the need to control and monitor them increased rapidly. In the literature, many researchers explored different techniques and developed genuine and hybrid systems to counter UAVs. These systems are not only found in academia, but different commercial options are also available. In this paper, the detection and localization using radar systems method are addressed. This method is one way to thwart drones. Other approaches are briefed in the background of this research study. Different radar technologies exist, and consequently different approaches to apply UAV detection and localization using radars. These approaches that are addressed in academia and the industry are surveyed in this research paper. Furthermore, the different possible radar categories are detailed to help the reader understand radar technology and thus its applications. In addition to commonly used localization algorithms for UAVs, the software and hardware implementation methods are also surveyed in this paper.

Keywords Radar · Unmanned aerial vehicle · Unmanned aerial vehicle localization · UAV · UAS

1 Introduction

Over the past decade, unmanned aerial vehicle (UAV) technologies have advanced rapidly and gained worldwide interest for their numerous applications across a wide range of fields. The deployment of amateur UAVs, however, has made them a possible accessory to crime and security threats, creating the need for drone detection and

S. M. Iddis (✉) · T. Dawdi · Q. Nasir · Y. Omran
Department of Electrical Engineering, University of Sharjah, Sharjah, UAE
e-mail: sidhis@sharjah.ac.ae

M. A. Talib
Department of Computer Science, University of Sharjah, Sharjah, UAE

control [1–3]. Drone incidents are numerous, and vary in risk. Thus, it is crucial for security to enhance UAV localization and detection methodologies [4].

In October 2020, a drone intrudes on restricted airspace over a red salmon fire in California, USA. This unauthorized usage of the drone hindered fire crews' efforts against the fire [4]. UAV detection technology can be considered as relatively new; however, it has already been explored in both the industry and academia. The main purpose of this survey paper is to provide a review of the currently available techniques for detecting and localizing unauthorized UAVs that could pose a threat to society [1]. This research pays special attention toward radar-based technologies among a broad range of gathered references, as well as provides a useful summary of the different methodologies that could be used to tackle the drone detection problem using radar.

Radar detection and localization schemes differ in many aspects, such as the type of radar used, the band of operation, the radar topology, combining different schemes to increase efficiency, and finally to design radars using software defined radios (SDR) [5].

Although radars were first developed for military use, they have evolved to become crucial in various applications, such as civilian or air traffic control. Regardless of the broad application range of radars they mainly have the following functions [5]:

1. Search environment (surveillance)
2. Target detection
3. Target position measurement (localization) and tracking
4. Measurement of target characteristics.

Target characteristics may include target size, shape, components, moving parts (type of propellers) and material composition [5, 6]. Target size is generally deduced from the magnitude of the return signal. As for the shape and components, they are a function of the angle of the returned signal. Finally, the moving parts and the target's material composition can be inferred from the modulation of the returned signal [5].

The survey work done in this paper covers the existing radar technologies, research and market work done in this area. This is motivated by the paramount significance of developing surveillance systems that alarm any intrusion of this kind. Since this technology is still maturing, the regulatory measures by governments are required to stay up-to-date. This can be done through detection and localization, which can be further followed by means of control, jamming, capture or forensic analysis, depending on the situation.

Some research papers in the literature have discussed UAV detection and localization, but none of them focused on all radar-based techniques for UAV detection and localization. The related survey papers are listed in Table 1.

For instance, in [1] the authors' main aim was to answer the following questions: what are the techniques dependent on ambient radio frequency? and what are common techniques to localize UAVs?

The authors answered these questions by reviewing over 100 references, covering techniques, including radar, RF techniques, acoustic sensors and computer vision-based techniques.

Table 1 Related survey papers comparison

Paper title	Year	Covered all types of radars	Detection by radar technologies	Localization by radar technologies	Number of radar localization and detection references
Detection, Localization, and Tracking of Unauthorized UAS and Jammers [1]	2017	✓	✓	✓	53
A Study on the Methods and Technologies Used for Detection, Localization, and Tracking of LSS UASs [2]	2018		✓		8
Detection and Classification of Multirotor Drones in Radar Sensor Networks: A Review [3]	2020		✓		39
Work in this paper	2020	✓	✓	✓	70

Similarly, in [2] the research covers methods and technologies used for detection, localization and tracking of LSS (low, slow and small) UASs, reviewing 30 references listing the techniques of radar, visual-based methods and acoustics.

Another complementing work is in [3], where the research aims on FMCW radar sensors, along with the different stages this radar type can provide, which are detection, tracking and classification.

In this research paper the following research questions are addressed:

1. What are radars?
2. How were radars used to counter UAVs in the literature?
3. What is the future of radar techniques in countering UAVs?
4. Why is localization important, and what algorithms are used to implement it?

This research paper used a number of the references cited in [1–3], and enlarged the research to include more references, focusing on the earlier mentioned research questions.

The structure of this paper is as follows: Sect. 2 covers a background on existing technologies to detect and localize UAVs. Section 3 lays out radar types and how they can be categorized. Section 4 presents radars frequency bands of operation. In Sect. 4, the most common localization algorithms are summarized. As for Sects. 5 and 6, the software tools used for radar simulation and possible hardware tools for

radar implementation are reviewed, respectively. Possible advancements for radar localization drawbacks are discussed in Sect. 7. Finally, conclusions are presented in Sect. 8.

2 Background

Existing technologies used for UAV detection can be categorized based on the technology of the detection method. This leads to four categories, namely acoustic, visual, radio frequency (RF) and radar [1–3].

Acoustic methods allow for the identification and classification of UAVs through their acoustic signals, usually involving the use of microphone arrays [7]. Using multiple acoustic sensors and beamforming technologies, researchers have managed to detect UAVs in the range 20–600 m [7]. Busset et al. implemented UAV detection via a beamforming algorithm that selectively extracts sounds coming from each detected sound source [8]. The acoustic signatures can then be analyzed and used to classify the type of UAV while a tracking algorithm is able to follow the sound source. While acoustic signals provide good grounds for UAV detection, limitations such as background noise stand in the way of a seamless solution [8].

With the advent of computer vision technologies, such as artificial intelligence (AI) and machine learning (ML), optical UAV detection has seen a rise in popularity most recently. Unlu et al. used two cameras and a rotating turret to detect and track UAVs in their research [9]. The team tested three deep learning models, namely YOLOV3, Haar and GMM. Each of these was trained on an NVIDIA GPU and proved to be effective for UAVs within a range of 1 km. While this method is popular, visual-based detection gives rise to many limitations when exposed to harsh weather conditions or darkness [10].

Another method of UAV detection uses RF signals. RF UAV detection is based on capturing the RF signals from both the UAV and its controller, if any. This will usually involve an RF receiver to capture the communication and processing hardware to classify and analyze the signals. Using RF-based detection, Nguyen et al. localized not only the UAV but also its controller [11]. Similarly, another team expanded on the use of passive and active RF methods to detect UAVs [12]. The limitation of such methods is that it requires a communication link to be present. Pre-programmed UAVs cannot usually be detected this way.

One of the more traditional and reliable methods of UAV detection would be detection by radar. Yadav et al. refer to an object detection technique that relies on the reflection of electromagnetic signals in the UHF and microwave ranges [13]. In another study [14], M. Hill examines the efficiency of combining software defined radio (SDR) and GNU Radio to produce a generic radar transmitter. The transmitter manages to generate CW, FM waveforms and pulsed waveforms. Another inspection of the topic was done by Patton L, who develops a low-complexity stochastic gradient algorithm that improves radar performance in band-limited interference by improving signal to interference plus noise ratio (SINR) [15].

Table 2 Summary of the four UAV detection technologies

Type of UAV detection	Advantages	Disadvantages
RF	– Relatively low cost [8]	– Unable to detect autonomous UAVs
Acoustic	– Can provide drone direction [1] – Can be assisted with AI [18]	– Noisy backgrounds can impact results – Short range (up to 600 m) [1]
Vision-based	– AI makes this a powerful tool [9]	– Visual limitations such as fog make this method unreliable
Radar	– Relatively long range [1] – Constant tracking	– High false alarm rates – Difficult to detect smaller UAVs [16]

While radar is one of the more typical technologies used to detect UAVs, its drawbacks include high false alarm rates due to the detection of small animals and birds, as well as difficulty in the detection of smaller drones with smaller radar cross-section (RCS) values [16, 17]. Nonetheless, drone detection using radar is still increasingly popular and the implementation of which will be the focus of this survey paper. Table 2 summarizes the detection techniques mentioned above.

3 Radar Types

Radar can be categorized upon the technology utilized, the topology implemented or their band of operation [19–23].

3.1 Technology-Based Radar Categories

Figure 1 categorizes the different radar types based on radar technology and their mode of operation. The main two categories are primary radars and secondary radars.

Primary Radar

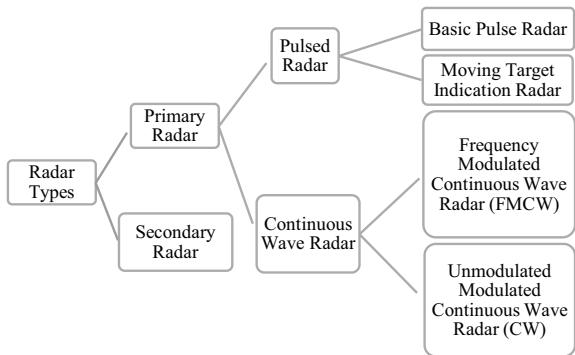
This system transmits a signal, continuous or pulsed. The system listens to the echo signal from targets hit by its own signal [19].

Pulsed Radar

Pulsed radars broadcast repetitive signals. The receiver listens to the echo for the remaining time of the pulse repetition period (PRP) [24]. This category includes basic pulse radar (BPR) and moving target indication radar (MTI) [25].

MTI uses the Doppler effect to discriminate between stationary and moving targets. As the system sends out recurring signals, the change of phase is analyzed

Fig. 1 Operation-based radar types



between two consecutive echoes. That is, if the echo indicates a change of position, then the reflecting object is moving. Otherwise, if the position is maintained, then the object is stationary [25].

Continuous-Wave (CW) Radar

Continuous-wave radar not only continuously transmits a signal but also simultaneously and constantly awaits the echo signal reflected by the target [26]. This signal is then processed to deduce information about the target [25]. This radar is further subcategorized into unmodulated continuous wave (CW) radar, and frequency modulated continuous wave (FMCW) radar [19].

Regular CW radars transmit signals with constant amplitude and frequency. The received signal's Doppler frequency shift, known as the Doppler effect, is analyzed to detect the target's speed [27]. FMCW radars are regular CW radars but transmit FM signals. The received echo is mixed with a portion of the transmitted signal to produce a beat signal at a frequency f_b . This allows the radar to not only determine the speed of the moving target using Doppler frequency shift but also measure the distance to that moving target [28]. Table 3 illustrates the key differences between the different primary radar technologies.

Secondary Radar

Unlike primary radar systems, secondary radars would have their *interrogator* to send out a signal and awaits a signal from the target's *transponder* (transmitter, responder). The transponder's signal would provide details about the target such as altitude, identification and onboard problems [19]. Therefore, secondary radar would only be useful when the target is cooperative and nonmalicious, whereas primary radars are useful for all kinds of targets. Furthermore, the reviewed literature summarizes the most used types of radar for UAV detection and localization. FMCW was used in [14, 33–41]. Adding to the aforementioned advantages of FMCW, it can also be implemented using SDR [42–44], as it is considered relatively not complex, is highly programmable and offers higher bandwidth compared to CW radar [45].

Table 3 Primary radar technologies pros and cons summary

Radar technology	Pros	Cons
MTI pulsed radar	The use of PRF avoids range ambiguities [29]	Complex to achieve low velocity ambiguity [25]
Unmodulated CW radar	Accurate in target velocity acquisition and detection at low altitudes [30]	<ul style="list-style-type: none"> - Lack of modulation hinders range calculation [25] - Spillover, the direct leakage of the transmitter and its accompanying noise into the receiver [25]
FMCW Radar	<ul style="list-style-type: none"> - Low power - Simple solid-state transmitters - Resistance to interception - Good range resolution [31] - High average power leading to small RCS [32] 	<ul style="list-style-type: none"> - Reduced range compared to pulse radar [32]

3.2 Topology-Based Radar Categories

The basic concept of a radar system comprises a transmitter and a receiver. However, the number and location dispersion of transmission and receiving stations creates different topologies, upon which radars can be classified [21, 22].

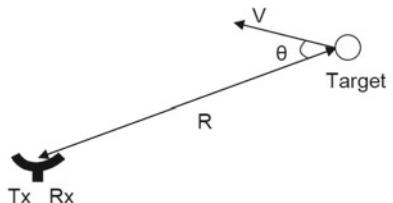
Monostatic Radars

Monostatic radars are the simplest topology. They consist of one transmitter and one receiver co-located and might even share one antenna working in full-duplex mode [21, 22]. This single antenna is used for both transmission and receiving. Figure 2 illustrates this topology [22].

A key difference between different radar topologies is the design equations. In a monostatic system, the Doppler shift f_d and radar range R would be calculated using (1) and (2), respectively.

$$f_d = \frac{2v \cos\theta}{\lambda} \quad (1)$$

Fig. 2 Monostatic radar topology [22]



where v is the target's speed, θ is the angle between the target and the line R, where R is the range between the target and receiver. Finally, λ is the wavelength of the transmitted radar signal [22].

$$R^2 = \sqrt{\frac{P_t G_t P_r G_r \partial \lambda^2}{4\pi^3 k T_s B (\text{SNR}) L}}, R = \frac{1}{2} c \tau, R_t = R_r = R \quad (2)$$

P_t and P_r are the transmitted and received power; G_t and G_r are the transmitted and received gains. ∂ is the RCS of the target, and λ is the wavelength of the transmitted radar signal. k is Boltzmann's constant, T_s is the receiving system noise temperature, B is the noise bandwidth, and L is the system loss ($L > 1$) [22].

Bistatic Radars

Unlike monostatic radars, bistatic radars have one transmitter and one receiver [21], but they are separated in location as demonstrated by Fig. 3.

Equations (3) and (4) depict the range and Doppler shift calculations in a bistatic radar system.

$$R_t + R_r = \frac{1}{2} c \tau \quad (3)$$

R_t and R_r are the transmitter to target range and the target to receiver range, respectively. c is the speed of light in free-space, and τ is the total time delay between transmitted and received signals [22].

$$f_d = \frac{2v \cos \varphi \cos(\beta/2)}{\lambda} \quad (4)$$

where φ is the angle between target velocity and bistatic bisector, positive clockwise. β is the bistatic angle between the transmitter and receiver with the target as the vertex [22].

Fig. 3 Bistatic radar topology [22]

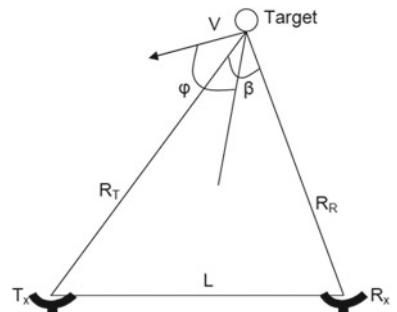
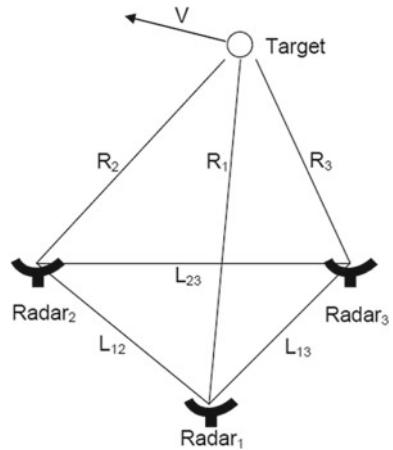


Fig. 4 General multistatic radar topology [22]



Netted (Multistatic) Radars

Netted or multistatic radars are like bistatic radars, except for the number of transmitters and receivers [22, 46]. They can be composed of multiple transmitters and one receiver, or even multiple transmitters and multiple receivers spatially separated, as Fig. 4 depicts.

As multistatic radars consist of multiple nodes, these nodes can be multiple monostatic nodes, or multiple bistatic nodes, or a mixture of the two. The range and Doppler shift (range to target) equations differ accordingly [46]. In his thesis, Teng [22] explicitly studies different cases, whereas [21] Ivashko studies the accuracy of different node architectures as well as their localization performance.

For these different node architectures, the performance of the radar network is affected notably. As demonstrated in [21], nodes from 2 to 19 of transmit–receive channels are randomly simulated, creating over 10^3 different radar systems. Then these different systems are compared for their target localization potential accuracy, using statistical averaging of the localization error, σ_p .

3.3 Operation Band-Based Radar Categories

The available operation frequency bands for radar include a large section of the electromagnetic spectrum from 3 Hz to 3×10^{12} Hz; after which comes the band used for LiDAR under the visible light spectrum [14]. Frequency bands are divided based on their different physical properties. As Fig. 5 illustrates, there are mainly 10 radar subranges or bands. Due to the differences in their physical properties, each of these bands is best used according to the application and its restrictions. Table 4 summarizes the key pros and cons of these bands. The nominal frequency ranges

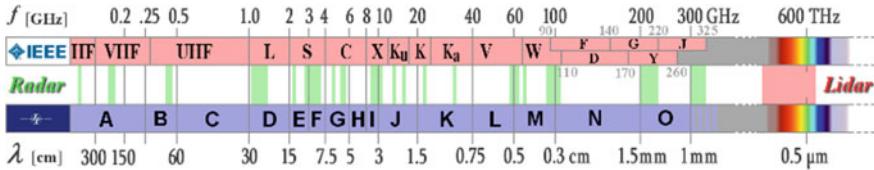


Fig. 5 Available radar bands [17]

in this table are based on the standardized radar bands found in IEEE Std 521-2002 [23].

3.4 Modern Radar Systems

Modern radar systems were developed mainly to overcome the limitations of the lateral mentioned technologies through combining their best features. These limitations are discussed in Sect. 8.

Adaptive Radars

Radar face many performance-hindering challenges; one of which is electromagnetic interference. EM interference can be externally intentional to jam the radar, or internal, in which case it is called clutter [51]. An adaptive radar receives the echo signal, analyzes if a target is present, applies certain algorithms to adaptively cancel interference, and then estimates the location and speed of the target [51, 52].

Cognitive Radar

Cognitive radar mainly differs from *adaptive radar* with the active feedback from the receiver to the transmitter. This feedback allows the radar to pass adaptivity information from the receiver to the transmitter so that transmission parameters can be changed optimally [53]. Transmission parameters include on-the-fly transmit frequency, channel, waveform shape, time-on-target, pulse repetition frequency (PRF), power, number of pulses and polarization [52]. Another key difference is intelligence. A cognitive radar must decide how and what parameters to change before passing the adapted information to the transmit level. This decision process can be applied using the *Bayesian* approach, or using machine learning techniques [52].

Opportunistic Radars

Other than the promising spectrum of applications for the recently implemented 5G networks is that their base stations can be utilized as radar ground stations to detect unauthorized UAVs, as shown in Fig. 6. Exploiting mmWave band offers high bandwidths of up to 2 GHz. Another benefit provided is applying beamforming

Table 4 Summary of key pros and cons of radar bands

Radar band	Nominal frequency range	Pros	Cons
HF and VHF	0.003–0.3 GHz	– Also called over the horizon Radars (OTR), due to their extremely large detection range [19]	– Extremely large antennas. Several kilometers [19] – Accuracy is compromised due to the band occupation with communication radio services [19, 47]
UHF	0.3–1 GHz	– Weather resilience allows very long range – Low frequency offers jamming resiliency [47]	– Limited available bandwidth, due to the band use by other electronic devices [47]
L-band	1–2 GHz	– Long-range air surveillance, up to 400 km [19, 48] – Low enables radiation with very high power [19]	– Relatively large, slowly rotating antennas [19]
S-band	2–4 GHz	– Smaller antenna [19]	– Shorter range. Up to 100 km [19] – Requires higher pulse power to achieve long ranges
C-band	4–8 GHz	– Largely influenced by weather [19]	– The range is short to medium [19]
X-band and Ku-band	8–12 GHz and 12–18 GHz	– Higher resolution [19, 38] – Extremely short pulses of a few nanoseconds, an excellent range resolution [19]	– High atmospheric attenuation [25] – Large ranges can no longer be achieved [19, 38]
K and Ka- band	18–27 GHz and 27–40 GHz	– Accuracy and range resolution increase. Best suited as pulsed when targeting aircrafts [19]	– Atmospheric attenuation increases [25]
V-band	40–75 GHz	– High bandwidth availability [19]	– Strong atmospheric attenuation. Limited range to 10 s m [19]
W-band	75–110 GHz	– Higher range resolution [19, 35]	– Atmospheric attenuation [19]
mm-band	100–300 GHz	– Robust against tough weather conditions [1, 49] – Precise targeting. Ideal for military applications [49]	– More expensive [49] – Scarcity of development kits with this band [49]

(continued)

Table 4 (continued)

Radar band	Nominal frequency range	Pros	Cons
Submillimeter (Submm) band	300–3000 Hz	<ul style="list-style-type: none"> – Ultrahigh-range resolution radars using undemanding back-end RF hardware [50] – Softer geometrical fading effects that stealth technology relies on for concealment [50] 	<ul style="list-style-type: none"> – Small market for THz components [50] – Strong atmospheric attenuation [50]

Fig. 6 5G Base stations as UAV radars [55]

high-gain mmWave antennas which increase UAV detection accuracy due to less interference with signals from low-gain directions.

In addition, using 5G network phased antenna arrays to steer the transmitted beams allows the assessment of the target's threat level more intelligently. A full system design on this novel technology is discussed in [54]. In [55], localization is achieved using the narrow steerable beam. This enables the base station to function as an mmWave radar system, detecting and identifying UAVs in urban environments.

Following the IEEE 802.11ad standard, Grossi et al. [56] prove the feasibility of using the sequence of directional transmissions performed during the ISS or RSS phase to opportunistically detect any objects in the environment. The principle is during the sector level sweep, if this transmission is echoed by an object, this echo would be detected. Not only does Grossi's proposed system [56] offer target detection, but it offers the Cramér-Rao bound (CRB) as a benchmark for the proposed estimators and a numerical study to assess the detection and estimation performance of the proposed solutions, which enables the system to estimate the object's position, radial velocity and backscattered signal amplitude.

SDR Radars

SDR radars have been investigated and experimented in academia frequently. Every work tackled different issues utilizing SDR radars, where those SDRs were implemented and simulated in different approaches. For example, in [42] the authors simulate and implement FMCW radar sensor using multiple SDR platforms and compare their performances. Their main application was to detect small UAVs. Another case [48] depicts the development and experimentation of a small-scale digital array radar (DAR) prototype, to detect a slow-moving micro-UAV target at a relatively long range. In this work, the authors utilized NI-PXIe SDR. In [57] an SDR radar was simulated and designed using LABVIEW software and implemented using USRP SDR; however, the authors do not specify their target profile.

Moreover, the authors of [40] discuss the implementation of FMCW surveillance radar for drone detection on SDR using GNU Radio and the USRP B210. However, in their conclusion they fail to detect drones, and their system only works for large targets such as cars and human beings; after which they analyze the possible improvements to achieve drone detection.

Another approach is used in [43], where the radar is designed using GNU Radio software; conversely MATLAB and Simulink were used to implement the logic blocks to process the received data and calculate the range of the target.

In [41], the authors design an FMCW radar using GNU Radio and USRP N210, however, their design lacks the target profile, as well as detection of multitarget scenarios.

4 Radar Bands of Operation Used for UAV Detection and Localization

While the available radar bands are many, some bands are more preferred than others for UAV detection and localization. In general, as the frequency increases, the antenna size decreases, and vice versa. The same relationship can be deduced between the frequency and the range of detection. This indicates that as the frequency decreases, the radar range increases. Therefore, the choice of band of operation for drone detection is inferred by trading off the range of detection, antenna size, RCS, bandwidth needed and available hardware. USRP and GNU-based radar would be limited to lower bands, up to about 6 GHz, as seen in [14, 15, 40, 41], as compared to other SDR kits reviewed, as seen in [1, 9].

UAV detecting and localizing radars can also be designed and built to meet a certain operating frequency, as a unique hardware, as seen in [38, 39] where the X-band was used. Another radar designed and built is depicted in [33], experimenting the W-band. UAV detection using radar technology is not only limited to the academic field, as it has been also explored commercially in the industry. Making use of mainly X-band and S-band radar systems, several of these industrial organizations and their respective product names and bands used have been summarized in Table 5.

Table 5 UAV detection using radar in the industry

Company name	Product name	Frequency
Aselsan corporation	IHTAR [58]	Ku-Band
BATS	UAV Guard [59]	X Band
Blighter/chess dynamics/enterprise control systems	AUDS Anti UAV Defense System [60]	Ku-Band
UAVShield	RadarOne/UAV Sentry [61]	X Band
IACIT	UAV BLOCKER [62]	HF band
JCPX development/DSNA services/aveillant	UWAS [63]	L Band
DeTect	HARRIER UAV Surveillance Radar [64]	S or X Band
SRC	Silent Archer [65]	S Band
Advanced protection systems	ctrl + sky [66]	X Band

5 Target Localization Algorithms

Targets can be localized via various methods. These methods depend on the technology used to detect these targets. Authors in [33, 41, 46, 67, 68] used FMCW radar. In [69], the authors used vision-based detection and localization RSS-based RLS algorithm. In [11], Nguyen et al. studied RF-based detection technique and angle of arrival (AoA) triangulation algorithm for localization. In [8], the authors superimpose acoustic imaging with visual imaging.

In [44], Sit et al. implement a MIMO OFDM radar using USRP X310 SDR. Their setup consists of four transmitters and four receivers operating at 4.05 GHz. The azimuth direction of arrival (DOA) is estimated using the Fourier beamforming method with Frobenius product of the processed radar matrices with the beamforming matrix, resulting in a DOA matrix. A similar localization approach is seen in [35], where the authors also use multiple receiver antennas. Four receiver channels are used to determine the azimuth and elevation angles, after that mono-pulse processing is done to the data coming from those receiver antennas, which enable localization.

Another example is demonstrated in [70]. The authors utilize laser scanner (LIDAR), and inertial data is given by a low-cost inertial measurement unit (IMU) to detect and classify the target. After which localization is carried through the iterative closest point (ICP) algorithm and LIDAR/Inertial Odometry (L/I-O) algorithm. Table 6 summarizes the most used localization algorithms accompanying each detection technology.

Table 6 Localization algorithms summary

Technology used	Localization mechanism
Radar (FMCW)	FFT-based method depending on beat frequency with triangulation and DOA algorithms
RF-based	AoA triangulation
Acoustic/Audio-based	Acoustic imaging algorithm which (sound power level)
Visual-based	RSS-based, RLS algorithm
LiDAR	Hybrid LIDAR/Inertial Odometry algorithm (L/I-O). Iterative closest point (ICP) algorithm

6 Software Tools for Radar Simulation

Owing to the challenges that arise when implementing radar systems for the detection of UAVs, most applications prefer to simulate before implementation, as seen in [8, 10, 42, 70–76]. Based on the most recent research in UAV detection, the most popular simulation platforms are expanded upon below.

6.1 MATLAB

MATLAB is a programming platform containing invaluable tools for science and research. Simulink is a MATLAB-based graphical programming environment used for modeling and simulating all types of systems [77]. This tool is used for countless applications in UAV control. One example involves the use of MATLAB to simulate a UAV jamming system [17]. These communications signals in MATLAB used eight channels that were used to find the optimal jamming frequency. In another application, realistic UAV flight paths were simulated to assist the development of a LiDAR UAV localization system [10]. Similarly, a different LiDAR application saw the use of MATLAB to assess the difference between a proposed algorithm and the traditional algorithm used for line fitting in the steps toward UAV localization in complex environments [70]. MATLAB can also be used to simulate FMCW signals, allowing various types to be compared and assessed within a short timeframe as found by [78] in their FMCW systems overview.

6.2 GNU Radio

GNU Radio is a well-known open-source software used to allow software defined radios to be simulated and implemented on RF hardware and to create simulation environments for RF applications [79]. Its popularity arose due to its easy access and compatibility with all USRP devices and most RF hardware devices. The essential

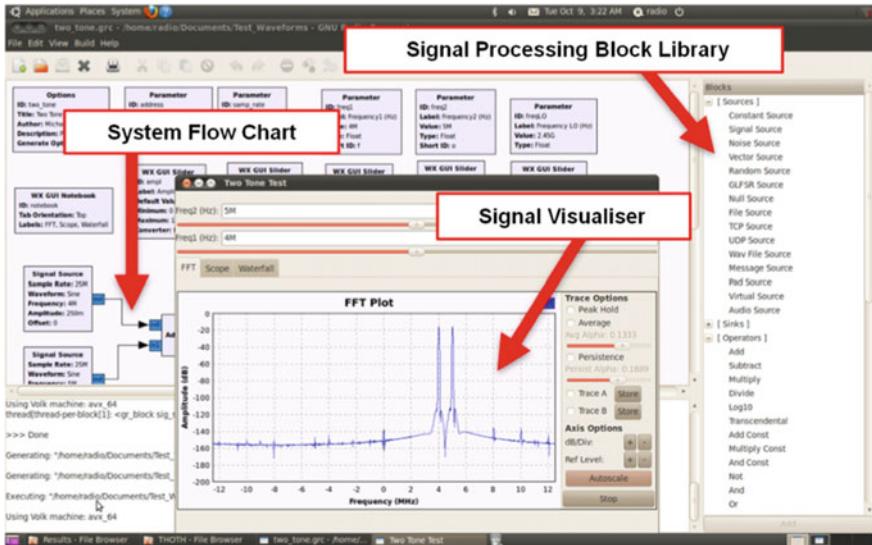


Fig. 7 GNU Radio Companion screenshot showing the main features [79]

building blocks of the platform are written in C++, while Python is used to create user-tools and generate graphs.

GNU Radio Companion is an environment consisting of a complete toolkit of signal processing blocks that allow GNU Radio to be used as a simulation tool. This software was used in [79] for this purpose, as seen in Fig. 7. Similar to MathWorks's Simulink, this features a signal processing block library and the ability to visualize signals in the form of graphs.

Before hardware implementation, L. Patton used GNU Radio as a proof-of-concept for the design of a waveform optimization algorithm capable of improving interference plus noise ratio in the presence of a band-limited interferer [15]. Similarly, another application used GNU Radio blocks with USRP N210 to simulate the delay caused by radar target reflection [41].

6.3 Other Supporting Software Tools

Another radar-assisting software would be LabView Communications Design Suite. One project used this software project panel to simulate the target and Doppler frequency [57]. Another research endeavor used CST Microwave Studio, a 3D electromagnetic (EM) simulation software used for mainly EM analysis [80]. Here the platform was used to see the effect of a nonlinear VCO tuning curve [71]. The software was then used to graph the simulated range profiles of UAV point targets for both linear and nonlinear VCO. In addition, various ray-tracing software can also be

used in UAV detection to simulate different types of environments and scenarios [1]. Examples of these types of software include Winprop [81] and Wireless InSite [82].

7 Implementation Tools for SDR Radars

With the introduction of software defined radio (SDR), wireless communication implementation has become far more flexible and applicable. There are numerous SDRs available for radar implementation, differing in terms of software, operating frequencies, FPGAs, microprocessors and price. Table 7 compares the different USRP devices for UAV detection.

7.1 Universal Software Radio Peripheral (USRP)

Universal software radio peripheral devices are tunable Ettus Research SDR transceivers that can be used for designing, prototyping and implementing radio communication systems [83]. An implementation of FMCW radar for target detection used the USRP network series N210 along with log-periodic antennas to achieve transmission and reception of 1 GHz [41].

Combining the use of GNU Radio and the USRP Bus Series B210, another research endeavor was able to develop an FMCW radar toolkit with a center frequency of 5.5 GHz [40]. While the B210 is not as powerful or efficient as the N210 for radar research, it is an effective low-cost option for proof-of-concept radar research. High-performance USRP modules as the X Series X310 have also been used for multi input multi output (MIMO) UAV radar applications [44, 84]. Implementing a 4×4 MIMO setup at 4.05 GHz operating frequency, one such application is able to successfully detect a DJI Phantom 3 UAV, recommending higher operating frequencies for future development.

Table 7 A comparison between USRP devices used for UAV radar detection [81]

	USRP N210	USRP B210	USRP X310
FPGA	Spartan3	Spartan6	Kintex 7
Logic cells	53 k	147 k	406 k
Clock rate	100 MHz	56 MHz	200 MHz
Streaming bandwidth per channel (16-bit)	25 MS/s	61.44 MS/s	200/s

7.2 Ancortek

Providing SDR kits with frequency limits of up to 26 GHz, Ancortek is another provider of embedded units that can be used for radar applications [85]. One application of Ancortek technologies is the 2400AD2 SDR kit, operating in the k-band and allowing interferometric radar measurements, making it an excellent candidate for target detection [1]. The Ancortek 580A2 radar, designed for use around 5.8 GHz, was implemented for another UAV detection system that explored the analysis of micro-Doppler signatures [86].

8 Advancements of Localization with Radar

Although radars are one of the leading technologies used for detection and localization, yet it faces different limitations in each application. Our focus in this research paper is on the challenges faced by radars used to detect and localize small targets. Table 8 summarizes the frequent limitations and developments and their impact on the radar and their possible advancements.

Table 8 Summary of radar limitations and possible advancements

Limitations	Impact on radar	Advancements
<ul style="list-style-type: none"> – Small UAVs [16, 90, 91]. – Blades/propeller size [92]. – Nonmetallic hulls [16, 40, 92] 	<ul style="list-style-type: none"> – Small RCS values 	<ul style="list-style-type: none"> – 2D and 3D digital beamforming to generate radar images – Micro Doppler identification
<ul style="list-style-type: none"> – Animals, birds and insects [16, 90]. 	<ul style="list-style-type: none"> – False alarms – High Doppler resolution 	<ul style="list-style-type: none"> – Implementing unwanted targets filters
<ul style="list-style-type: none"> – Flying close to the ground [16, 90]. – Urban environment [16, 90]. 	<ul style="list-style-type: none"> – High clutter 	<ul style="list-style-type: none"> – Highly adaptive active radar systems – Cognitive algorithms
<ul style="list-style-type: none"> – High-performance UAVs. [16, 36, 90] 	<ul style="list-style-type: none"> – High power – Used by authorities and governmental institutions – Health consequences – Expensive 	<ul style="list-style-type: none"> – Low-cost, low-power systems, FMCW radar
<ul style="list-style-type: none"> – Terrain masking effects [16] 	<ul style="list-style-type: none"> – Targets are not observable – No line of sight 	<ul style="list-style-type: none"> – MIMO radars – Active multi-static radar systems
<ul style="list-style-type: none"> – Design restrictions [35, 37, 39, 40, 76] 	<ul style="list-style-type: none"> – Imperfect results 	<ul style="list-style-type: none"> – Upgrading the design to improve the performance and accuracy – Additional stages in the processing stage

Localization of UAVs can be challenging due to their small cross-sectional area, slow movement, and low altitude flight. High levels of *clutter* and *Doppler shift* are common challenges faced by radars, therefore combining different radar technologies is highly recommended as Güvenç et al. indicate in their paper [16], and reinforced through Multerer et al. research [37]. Syeda et al. [87] combine FMCW and MIMO radars, resulting in low-power, low-cost, and commercially available radar.

Most recent approaches to improve radar performance rely on sensing the environment and learning from it by cognitive and adaptive methods as discussed in Sect. 3.4. References [53, 88, 89] show how cognitive radar enhances high clutter from flying close to the ground and urban environments. Multerer et al. represented a 3D beamforming radar image, helping in target discrimination and preventing false alarms by eliminating *Doppler shift* and increasing *Doppler resolution* [36]. This radar resembles *holographic radars*, which differ in having high output power. Thus, they are only used by authorized users and are expensive, for example, the *Gamekeeper-16U* [20, 90].

9 Conclusion

In this research paper, we have studied in-depth, analyzed and presented detailed literature on the topic of drone detection and localization. In our survey paper, we have emphasized on the use of radar technology to localize unauthorized UAVs. Traditional radar types have been explored in addition to more modern types emerging in research such as 5G cognitive radars and SDR radars. The different radar bands used for UAV detection have been expanded upon and the most popular and effective bands have been identified. Summaries of various radar-supporting software platforms used in the literature, such as GNU Radio, MATLAB and LabVIEW, were documented and their uses were explained. We also touched upon the future advancements of radar, as well as their limitations in Sect. 8.

Acknowledgments We would like to express our sincere gratitude to the General Civil Aviation Authority (GCAA) in the UAE for establishing the Aerospace Center of Excellence and conducting this research study. We also thank our supervisors and colleagues from OpenUAE Research and Development Group at University of Sharjah, who provided insight and expertise that greatly assisted the research.

References

1. Güvenç, I., Ozdemir, O., Yapıcı, Y., Mehrpouyan, H., Matolak, D.: Detection, localization, and tracking of unauthorized UAS and Jammers. In: AIAA/IEEE Digital Avionics Systems Conference—Proceedings, Nov. 2017, vol. 2017-Sept. <https://doi.org/10.1109/dasc.2017.8102043>

2. Mototolea, D.: A study on the methods and technologies used for detection, localization, and tracking of LSS UASs. *J. Mil. Technol.* **1**(2), 11–16 (2018). <https://doi.org/10.32754/jmt.2018.2.02>
3. Coluccia, A., Parisi, G., Fascista, A.: Detection and classification of multirotor drones in radar sensor networks: A review. *Sensors (Switzerland)*, vol. 20, no. 15. MDPI AG, pp. 1–22, 01 Aug. 2020, <https://doi.org/10.3390/s20154172>
4. Drone Intrudes on Restricted Airspace Over Red Salmon FireNews Blog. <https://www.northcoastjournal.com/NewsBlog/archives/2020/10/18/drone-intrudes-on-restricted-airspace-over-red-salmon-fire>. Accessed Nov. 12 2020
5. David, J.: Radar Fundamentals. Accessed 12 Nov. 2020. <http://www.nps.navy.mil/faculty/jenn>
6. Curry, G.R.: Radar Measurement and Tracking, pp. 165–193 (2005). Accessed 12 Nov. 2020. https://www.mendeley.com/catalogue/8dc6e5ae-ffc6-3e4e-8818-7bb7bc416925/?utm_source=desktop&utm_medium=1.19.4&utm_campaign=open_catalog&userDocumentId=%7Bc0b47b2e-8681-44e9-b8e3-0c2e75c6fd13%7D
7. Hauzenberger, L., Ohlsson, E.H., Swartling, M.: Drone Detection using Audio Analysis (2015). Accessed 12 Nov. 2020. <http://lup.lub.lu.se/student-papers/record/7362609>
8. Busset, J. et al.: Detection and tracking of drones using advanced acoustic cameras. In: Unmanned/Unattended Sensors and Sensor Networks XI; and Advanced Free-Space Optical Communication Techniques and Applications, Oct. 2015, vol. 9647, p. 96470F (2015). <https://doi.org/10.1117/12.2194309>
9. Unlu, E., Zenou, E., Riviere, N., Dupouy, P.E.: Deep learning-based strategies for the detection and tracking of drones using several cameras. *IPSJ Trans. Comput. Vis. Appl.* **11**(1), Dec. 2019. <https://doi.org/10.1186/s41074-019-0059-x>
10. Deilamsalehy, H., Havens, T.C.: Sensor fused three-dimensional localization using IMU, camera and LiDAR, Jan 2017. <https://doi.org/10.1109/icsens.2016.7808523>
11. Nguyen, P. et al.: Towards RF-based Localization of a Drone and Its Controller (2019). <https://doi.org/10.1145/3325421.3329766>
12. Nguyen, P., Ravindranathan, M., Nguyen, A., Han, R., Vu, T.: Investigating cost-effective RF-based detection of drones. In: DroNet 2016 - Proceedings of the 2nd Workshop on Micro Aerial Vehicle Networks, Systems, and Applications for Civilian Use, co-located with MobiSys 2016, June 2016, pp. 17–22 (2016). <https://doi.org/10.1145/2935620.2935632>
13. Singh Yadav, A., Kumar, S.: A Review Paper on Radar System, pp. 2349–6010 (2016)
14. Hill, M.M.: Developing a Generic Software-Defined Radar Transmitter using GNU Radio (2012)
15. Patton, L.K.: A GNU Radio Based Software-Defined Radar, 12 Nov. 2020. https://corescholar.libraries.wright.edu/etd_all, https://corescholar.libraries.wright.edu/etd_all/91
16. Wellig, P. et al.: Radar systems and challenges for C-UAV. In: Proceedings International Radar Symposium, Aug. 2018, vol. 2018-June. <https://doi.org/10.23919/irs.2018.8448071>
17. Zhang, Y., Cruz, J.R., Dyer, J.W.: Effectiveness of Electronic Counter Measures (ECM) on Small Unmanned Aerial Systems (Suas): Analysis and Preliminary Tests a Thesis Approved For The School Of Electrical and Computer Engineering (2018). Accessed 12 Nov. 2020. <https://shareok.org/handle/11244/299802>
18. Ezuma, M., Erden, F., Anjinappa, C.K., Ozdemir, O., Guvenc, I.: Micro-UAV Detection and Classification from RF Fingerprints Using Machine Learning Techniques. In: IEEE Aerospace Conference Proceedings, Mar. 2019, vol. 2019-Mar. <https://doi.org/10.1109/aero.2019.8741970>
19. Radartutorial (2020). <https://www.radartutorial.eu/index.en.html>. Accessed 12 Nov. 2020
20. Jahangir, M., Baker, C.: Persistence surveillance of difficult to detect micro-drones with L-band 3-D holographic radarTM, Oct. 2017. <https://doi.org/10.1109/radar.2016.8059282>
21. Ivashko, I.: Radar Networks Performance Analysis and Topology Optimization (2016). <https://doi.org/10.4233/uuid:1a6dab8e-ebbd-41a1-bd5e-866a9050fc68>
22. Teng, Y.: Fundamental Aspects of Netted Radar Performance (2010)
23. 521-1984-521-1984-IEEE Standard Letter Designations for Radar-Frequency Bands-IEEE Standard. Accessed 12 Nov. 2020. <https://ieeexplore.ieee.org/document/29086>

24. US7002511B1—Millimeter wave pulsed radar system—Google Patents
25. Curry, G.R.: Radar Essentials: A Concise Handbook for Radar Design and Performance Analysis. Institution of Engineering and Technology (2012)
26. Chapter 14: CW and FM Radar|Engineering360
27. Continuous wave ranging radar, 12 Oct. 1971
28. Brooker Graham, M.: Understanding Millimetre Wave FMCW Radars. Accessed 12 Nov. 2020. https://www.researchgate.net/publication/228979037_Understanding_millimetre_wave_FMCW_radars
29. RF Wireless Vendors and Resources|RF Wireless World (2020). <https://www.rfwireless-world.com/> Accessed 12 Nov. 2020
30. 4,176,351 1. Method of Operating A Continuous-Wave Radar, 18 Aug. 1978
31. Zhang, D., Kurata, M., Inaba, T.: FMCW radar for small displacement detection of vital signal using projection matrix method. *Int. J. Antennas Propag.* **2013** (2013). <https://doi.org/10.1155/2013/571986>
32. Moore, E.G., Rutherford, M.J., Valavanis, K.P.: Radar Detection, Tracking and Identification for UAV Sense and Avoid Applications (2019)
33. Caris, M., Johannes, W., Stanko, S., Pohl, N.: Millimeter wave radar for perimeter surveillance and detection of MAVs (Micro Aerial Vehicles). In: Proceedings International Radar Symposium, Aug. 2015, vol. 2015-August, pp. 284–287. <https://doi.org/10.1109/irs.2015.7226314>
34. Drozdzowicz, J. et al.: 35 GHz FMCW drone detection system. In: Proceedings International Radar Symposium, June 2016, vol. 2016-June. <https://doi.org/10.1109/irs.2016.7497351>
35. Caris, M., Johannes, W., Sieger, S., Port, V., Stanko, S.: Detection of small UAS with W-band radar, Aug. 2017. <https://doi.org/10.23919/irs.2017.8008143>
36. Multerer, T. et al.: Low-cost jamming system against small drones using a 3D MIMO radar based tracking. In: European Microwave Week 2017: “A Prime Year for a Prime Event”, EuMW 2017—Conference Proceedings; 14th European Microwave Conference, EURAD 2017, June 2017, vol. 2018-January, pp. 299–302. <https://doi.org/10.23919/eurad.2017.8249206>
37. Laučys, A., et al.: Investigation of detection possibility of uavs using low cost marine radar. *Aviation* **23**(2), 48–53 (2019). <https://doi.org/10.3846/aviation.2019.10320>
38. Moses, A., Rutherford, M.J., Valavanis, K.P.: Radar-based detection and identification for miniature air vehicles. In: Proceedings of the IEEE International Conference on Control Applications, pp. 933–940 (2011). <https://doi.org/10.1109/cca.2011.6044363>
39. Itcia, E., Wasselin, J.-P., Mazuel, S., Otten, M., Huizing, A.: FMCW radar for the sense function of sense and avoid systems onboard UAVs. In: Emerging Technologies in Security and Defence; and Quantum Security II; and Unmanned Sensor Systems X, Oct. 2013, vol. 8899, p. 889914 (2013). <https://doi.org/10.1117/12.2028518>
40. Mathumo, T.W., Swart, T.G., Focke, R.W.: Implementation of a GNU radio and python FMCW radar toolkit. In: 2017 IEEE AFRICON: Science, Technology and Innovation for Africa, AFRICON 2017, Nov. 2017, pp. 585–590 (2017). <https://doi.org/10.1109/afrcon.2017.8095547>
41. Sundaresan, S., Anjana, C., Zacharia, T., Gandhiraj, R.: Real time implementation of FMCW radar for target detection using GNU radio and USRP. In: 2015 International Conference on Communication and Signal Processing, ICCSP 2015, Nov. 2015, pp. 1530–1534 (2015). <https://doi.org/10.1109/iccsp.2015.7322772>
42. Jiaxi zhu Norman, B.: Low-cost, Software Defined FMCW Radar for Observations of Drones (2017)
43. Fernandes, V.N.: Implementation of a RADAR System using MATLAB and the USRP. California State University—Northridge, 2012. <http://oatd.org/oatd/record?record=handle%5C%3A10211.2%5C%2F1039>. Accessed 12 Nov. 2020
44. Sit, Y.L., Nuss, B., Basak, S., Orzol, M., Wiesbeck, W., Zwick, T.: Real-time 2D + velocity localization measurement of a simultaneous-transmit OFDM MIMO Radar using Software Defined Radios. In: 2016 13th European Radar Conference EuRAD 2016, no. May 2017, pp. 21–24 (2016)

45. Jankiraman, M.: Design of Multi-Frequency CW Radars. Institution of Engineering and Technology (2007)
46. Sandström, S.-E., Akeab, I.K.: A study of some FMCW radar algorithms for target location at low frequencies. *Radio Sci.* **51**(10), 1676–1685 (2016). <https://doi.org/10.1002/2016RS005974>
47. Kuschel, H.: VHF/UHF radar. Part 2: operational aspects and applications. *Electron. Commun. Eng. J.* **14**(3), 101–111 (2002). <https://doi.org/10.1049/ecej:20020302>
48. Aldowesh, A., Alnuaim, T., Alzogaby, A.: Slow-Moving Micro-UAV Detection with A Small Scale Digital Array Radar, Apr. 2019. <https://doi.org/10.1109/radar.2019.8835567>
49. An Overview of Antenna Systems in Millimeter-Wave Radar Applications—Microwave Product Digest (2020). <https://www.mpdigest.com/2017/09/22/an-overview-of-antenna-systems-in-millimeter-wave-radar-applications/>. Accessed 12 Nov. 2020
50. Cooper, K.B., Chattopadhyay, G.: Submillimeter-wave radar. *IEEE Microw. Mag.* **15**(7), 51–67 (2014). <https://doi.org/10.1109/MMM.2014.2356092>
51. Smith, S.T.: Adaptive radar. In: Wiley Encyclopedia of Electrical and Electronics Engineering. Wiley (1999)
52. Greco, M.S., Gini, F., Stinco, P., Bell, K.: Cognitive radars: a reality?, Feb. 2018. <http://arxiv.org/abs/1803.01000>. Accessed 12 Nov. 2020
53. Oechslin, R., Wellig, P., Hinrichsen, S., Wieland, S., Aulenbacher, U., Rech, K.: Cognitive radar parameter optimization in a congested spectrum environment. In: 2018 IEEE Radar Conference, RadarConf 2018, June 2018, pp. 218–223 (2018). <https://doi.org/10.1109/radar.2018.8378560>
54. Solomitckii, D., Gapeyenko, M., Semkin, V., Andreev, S., Koucheryavy, Y.: Technologies for efficient amateur drone detection in 5G Millimeter-Wave cellular infrastructure. *IEEE Commun. Mag.* **56**(1), 43–50 (2018). <https://doi.org/10.1109/MCOM.2017.1700450>
55. Zhao, J., Fu, X., Yang, Z., Xu, F., Rong, B.: Radar-Assisted UAV detection and identification based on 5G in the internet of things. *Wirel. Commun. Mob. Comput.* **2019** (2019). <https://doi.org/10.1155/2019/2850263>
56. Grossi, E., Lops, M., Venturino, L., Zappone, A.: Opportunistic Radar in IEEE 802.11ad Vehicular Networks, Nov. 2017, pp. 1–5 (2017). <https://doi.org/10.1109/vtcspring.2017.8108446>
57. Maggiora, R., Ruka, R.: Politecnico di Torino Implementation of a complete radar system on the NI USRP-2944R software defined radio platform Supervisor: Candidate
58. IHTAR Anti-Drone System|ASELSAN (2020). <https://www.aseisan.com.tr/en/capabilities/air-and-missile-defense-systems/air-and-missile-defense-systems/ihtar-antidrone-system>. Accessed 12 Nov. 2020
59. Tactical Air Defense Radars—Belgian Advanced Technology Systems|Securing your homeland (2020). <https://www.bats.be/solutions/radars/tactical-air-defense-radar>. Accessed 12 Nov. 2020
60. AUDS Anti-UAV Defence System|Counter-UASIC-UASI Blighter (2020). <https://www.blighter.com/products/auds-anti-uav-defence-system/>. Accessed 12 Nov. 2020
61. DroneShield Rolls Out Revolutionary Compact Radar|DroneShield (2020). <https://www.droneshield.com/press-releases-content/2018/4/12/droneshield-rolls-out-revolutionary-compact-radar>. Accessed 12 Nov. 2020
62. Home|IACIT (2020). <https://www.iacit.com.br/en/>. Accessed 12 Nov. 2020
63. Introducing UWAS, the UAV Watch—Aveillant (2020). <https://www.aveillant.com/introducing-uwas-the-uav-watch/>. Accessed 12 Nov. 2020
64. Security & Surveillance Radars|DeTect, Inc. (2020). <https://detect-inc.com/security-surveillance-radars/>. Accessed 12 Nov. 2020
65. Products—Redefining Possible with Advanced Technologies and Systems|SRC, Inc (2020). <https://www.srccinc.com/products/>. Accessed 12 Nov. 2020
66. Advanced Protection Systems Inc. (2020). <https://apsystems.tech/>. Accessed 12 Nov. 2020
67. Pan, X., Xiang, C., Liu, S., Yan, S.: Low-Complexity time-domain ranging algorithm with FMCW sensors. *Sensors* **19**(14), 3176 (2019). <https://doi.org/10.3390/s19143176>

68. López Martínez, C., Vidal Morera, M.: Simulation of FMCW radar systems based on software defined radio (2016). <https://upcommons.upc.edu/handle/2117/96865>. Accessed 12 Nov. 2020
69. Hassan, A., Gelman, I., Loftus, J.: Adversary UAV localization with software defined radio, Apr 2019. <https://digitalcommons.wpi.edu/mqp-all/7115>. Accessed 12 Nov. 2020
70. Opromolla, R., Fasano, G., Rufino, G., Grassi, M., Savvaris, A.: LIDAR-inertial integration for UAV localization and mapping in complex environments. In: 2016 International Conference on Unmanned Aircraft Systems, ICUAS 2016, June 2016, pp. 649–656 (2016). <https://doi.org/10.1109/icuas.2016.7502580>
71. Peng, Z.: Development of portable radar systems for short-range localization and life tracking (2018). Accessed 12 Nov. 2020. <https://ttu-ir.tdl.org/handle/2346/82111>
72. Husodo, A.Y., Jati, G., Alfiandy, N., Jatmiko, W.: Intruder drone localization based on 2D image and area expansion principle for supporting military defence system. In: 2019 IEEE International Conference on Communication, Networks and Satellite, Commetsat 2019—Proceedings, Aug. 2019, pp. 35–40. <https://doi.org/10.1109/commetsat.2019.8844103>
73. Ding, G., Wu, Q., Zhang, L., Lin, Y., Tsiftsis, T.A., Yao, Y.D.: An amateur drone surveillance system based on the cognitive internet of things. IEEE Commun. Mag. **56**(1), 29–35 (2018). <https://doi.org/10.1109/MCOM.2017.1700452>
74. Shinde, C., Lima, R., Das, K.: Multi-view geometry and deep learning based drone detection and localization. In: 2019 5th Indian Control Conference, ICC 2019—Proceedings, May 2019, pp. 289–294 (2019). <https://doi.org/10.1109/indiancc.2019.8715593>
75. Ritchie, M., Fioranelli, F., Griffiths, H., Torvik, B.: Micro-drone RCS analysis. In: 2015 IEEE Radar Conference—Proceedings, Oct. 2015, pp. 452–456 (2015). <https://doi.org/10.1109/radarconf.2015.7411926>
76. Seybert, A. et al.: Detect sense and avoid radar for UAV avionics telemetry item type text. In: Proceedings Detect Sense and Avoid Radar for Uav Avionics Telemetry (2020). <http://hdl.handle.net/10150/595802>. Accessed 12 Nov. 2020
77. Simulink—Simulation and Model-Based Design—MATLAB & Simulink (2020). <https://www.mathworks.com/products/simulink.html>. Accessed 12 Nov. 2020
78. Parrish, K.: An Overview of FMCW Systems in MATLAB, Cerc.Utexas.Edu, no. July, pp. 1–7 (2015). <http://www.cerc.utexas.edu/~kparrish/class/radar.pdf>
79. About GNU Radio GNU Radio. <https://www.gnuradio.org/about/>. Accessed 12 Nov. 2020
80. CST Studio Suite 3D EM simulation and analysis software (2020). <https://www.3ds.com/products-services/simulia/products/cst-studio-suite/>. Accessed 12 Nov. 2020
81. FE Modeling and Visualization\Altair HyperWorks (2020). <https://www.altair.com/hyperworks/>. Accessed 12 Nov. 2020
82. Wireless EM Propagation Software—Wireless InSite—Remcom (2020). <https://www.remcom.com/wireless-insite-em-propagation-software/>. Accessed 12 Nov. 2020
83. About USRP Bandwidths and Sampling Rates—Ettus Knowledge Base (2020). https://kb.ettus.com/About_USRP_Bandwidths_and_Sampling_Rates. Accessed 12 Nov. 2020
84. Nuss, B., Sit, L., Fennel, M., Mayer, J., Mahler, T., Zwick, T.: MIMO OFDM radar system for drone detection, Aug. 2017. <https://doi.org/10.23919/irs.2017.8008141>
85. Technologies\Ancortek Inc (2020). <https://ancortek.com/our-technologies>. Accessed 12 Nov. 2020
86. Jian, M., Lu, Z., Chen, V.C.: Drone detection and tracking based on phase-interferometric Doppler radar. In: 2018 IEEE Radar Conference, RadarConf 2018, June 2018, pp. 1146–1149. <https://doi.org/10.1109/radar.2018.8378723>
87. Syeda, R.Z., Adela, B.B., van Beurden, M.C., van Zeijl, P.T.M., Smolders, A.B.: Design of a mm-wave MIMO radar demonstrator with an array of FMCW radar chips with on-chip antennas 2019, pp. 33–36 (2020). <https://research.tue.nl/en/publications/design-of-a-mm-wave-mimo-radar-demonstrator-with-an-array-of-fmcw>. Accessed 12 Nov. 2020
88. Zhang, L., Wei, N., Du, X.: Waveform design for improved detection of extended targets in sea clutter. Sensors **19**(18), 3957 (2019). <https://doi.org/10.3390/s19183957>
89. Oechslin, R., Aulenbacher, U., Rech, K., Hinrichsen, S., Wieland, S., Wellig, P.: Cognitive radar experiments with codir. In: IET Conference Publications, vol. 2017, no. CP728 (2017). <https://doi.org/10.1049/cp.2017.0386>

90. Quilter, T., Baker, C.: The application of staring radar to the detection and identification of small Unmanned Aircraft Systems in Monaco, Aug. 2017. <https://doi.org/10.23919/irs.2017.8008145>
91. Ganti, S.R., Kim, Y.: Implementation of detection and tracking mechanism for small UAS. In: 2016 International Conference on Unmanned Aircraft Systems, ICUAS 2016, June 2016, pp. 1254–1260. <https://doi.org/10.1109/icuas.2016.7502513>
92. Poullin, D.: Counteracting illegal UAV flights: passive DVB radar potentiality. In: Proceedings International Radar Symposium, Aug.2018, vol. 2018-June. <https://doi.org/10.23919/irs.2018.8447902>

Prediction of Parkinson's Disease Using Machine Learning Models—A Classifier Analysis



A. T. Rohit Surya, P. Yaswanthram, Prashant R. Nair,
S. S. Rajendra Prasath, and Sundeep V. V. S. Akella

Abstract Among the chronic nervous system diseases, Parkinson's disease (PD) is known for its progressiveness in impairing the speech ability, gait as well as complex muscle and nerve actions. Hence an early diagnosis of PD will help in reducing the symptoms. Telemedicine offers a cost-effective and convenient approach, and several studies have used dysphonic features to remotely detect PD. In this study, we have used a data set from Kaggle, which included voice measurements from 31 people of whom 23 were diagnosed with PD. The data set included 22 different attributes pertaining to voice measurements, including the pitch period entropy with 195 voice recordings for each of the individuals. In the data pre-processing, the correlated attributes were removed and we used 10 non-correlating attributes (< 0.7) along with individual status (0 and 1 for healthy and PD, respectively). The data set after pre-processing was split into 70:30 ratio and also ascertained that the number normal versus PD are in equal ratios in both the training and testing data sets, respectively. The data set was evaluated with four different supervised classification machine learning (ML) models, namely random forest, XGBoost, SVM and decision tree. The XGBoost classifier model was found to be highly efficient in precise classification of PD with an accuracy of 0.93.

Keywords Machine learning · Classifiers · Parkinson's disease

1 Introduction

Parkinson's disease (PD) is a progressive nervous disorder system that usually starts with barely noticeable tremors and further leads to stiffness and slowing of movement. Due to the way in which the brain is connected to every part of our body, biomedical voice measurements can offer insight into neurological disorders. There is no cure for the illness and researchers have also noted that certain changes occur in the brains

A. T. Rohit Surya · P. Yaswanthram · P. R. Nair (✉) · S. S. Rajendra Prasath · S. V. V. S. Akella
Department of Computer Science and Engineering, Amrita School of Engineering, Amrita
Vishwa Vidyapeetham, Coimbatore, India
e-mail: prashant@amrita.edu

of people with PD, but it is not clear why these changes occur. Because the origin of Parkinson's is unclear, proven ways of preventing the disease still remain a mystery. Common symptoms are speech and writing impaired posture and balance, and loss of automatic movement. The progress of the disease might cause changes in the vocal cords. The muscle becomes thinner and less taut in the vocal cord. The vocal cords do not vibrate as they should, and a gap between the cords develops. In this work, we present a data set that contains information on voice measurements from 31 people, out of which 23 with Parkinson's disease. Data pre-processing techniques were applied to select the important features and applied on various machine learning (ML) classification models (random forest, XGBoost, SVM and decision tree). The ML models were trained and tested with the data to predict if the patient had PD or not. The ML models were evaluated on their accuracy to compare their performance and the XGBoost classification model was found to perform the best with an accuracy of 0.93.

2 Literature Review

Researchers pointed out the need for using non-motor features, such as RBD olfactory loss, RBD, and also important biomarkers. By using various machine learning models, they have also developed automatic diagnostic models like BayesNet, random forest and multilayer perception, and also boosted logistic regression. Boosted logistic regression offers the best performance. It additionally inferred that these models can be used for an early expectation of Parkinson's illness [1]. The specialists in this work zeroed in on examination of the voice estimation highlights of the patient data set to comprehend whether the patient is experiencing PD. They actualized logistic regression, decision tree, and K-closest neighbor (KNN) as base classifiers and have reported that in the vast majority of cases, the ensemble classifiers display much greater accuracy compared to the base classifiers [2]. Another research presents the characterization strategy and spotlight on discourse signs to the recognizable proof of Parkinson's infection. They used feed-forward strategy for the arrangement. They reasoned that the feed-forward strategy gives best order precision using discourse signals [3]. Another investigation examined Parkinson's illness and its investigation using AI calculations. The examination of information is finished using a directed AI approach. It additionally thinks and briefs about different managed learning calculations and their examination [4]. Some researchers proposed a technique for the expectation of Parkinson's infection seriousness using profound neural organizations. They used the "TensorFlow" python deep learning library to conduct their neural organization to foresee the seriousness of their neural organization. They additionally reasoned that the exactness esteems acquired by their strategy are better contrasted with precision gotten in past examination work [5]. Another research proposes three administered calculations for improving Parkinson's illness examination by recognition. Support vector machine (SVM), K-nearest neighbor (KNN) calculation and logistic regression (LR) were used for the conjecture of Parkinson's

illness. They reasoned that SVM got the best for breaking down the Parkinson data sets and KNN procured the most noticeably terrible execution [6–8]. Student Placement analyzer, a suggestion framework using machine learning can assist the situation cell within an association to recognize and focus on the targeted students and enhance their specialized skills as well as relational skills. Machine learning algorithms are likewise used for weld quality observing using acoustic signature classification algorithms and here the imperfections are ordered. The two calculations are random forest and J48 that were used and algorithm classification efficiencies are reported. We can likewise anticipate cross-site scripting assault using machine learning calculations. Using three machine learning calculations such as naïve Bayes, support vector machine and J48 decision tree, we can also classify gait features for stance and swing using machine learning. This gait study will be helpful for analyzing conditions during control and development-related infection Machine learning is widely used in a variety of diverse applications and contexts such as biometrics, vibration analysis and tool conditioning [9–11].

3 Proposed Architecture

In this investigation, we have implemented different classifiers on the Parkinson's data set and studied the accuracy of different classifiers. The data set used in this study did not contain any outliers and null values and hence was directly used for analysis. The correlation across the voice measurement attributes used in the study was determined using corr () function in python (pandas library). Among the 22 attributes, 12 attributes revealed better correlation (> 0.7) and were depicted as a heat map as shown in Fig. 1. These 12 characteristics were excluded and the remaining 10 attributes, namely MDVP:Fo(Hz), MDVP:Fhi(Hz), MDVP:Flo(Hz), MDVP:Jitter(percent), HNR, RPDE, DFA, spread1, spread2 and D2, were retained for use in ML model training and PDD prediction. The author of the data set had shown that the feature sets HNR, RPDE, DFA and PPE had better SVM classification performance, and we could find that PPE had a better correlation in the data pre-processing. Post removal, the non-correlating attributes were again shown in the heat map with the correlation values in Fig. 2.

Our proposed architecture is described in Fig. 3. We start by importing the data set and some required packages that are required. We then move on to the data pre-processing phase where we analyze the given data set and check and fill the missing data. Then we visualize the relationship between two variables using pair plots.

We use correlation to correlate all the columns pair-wise. When the correlation of a pair is near 1 it signifies that the correlation between them is stronger. From the heat map that is produced as shown in Fig. 1, we find that we have some columns that have high correlations, which means we can drop those columns without losing data. We then plot the correlation of the new data set, as shown in Fig. 2, which shows us that we have eliminated correlated columns. This completes our data pre-processing phase.

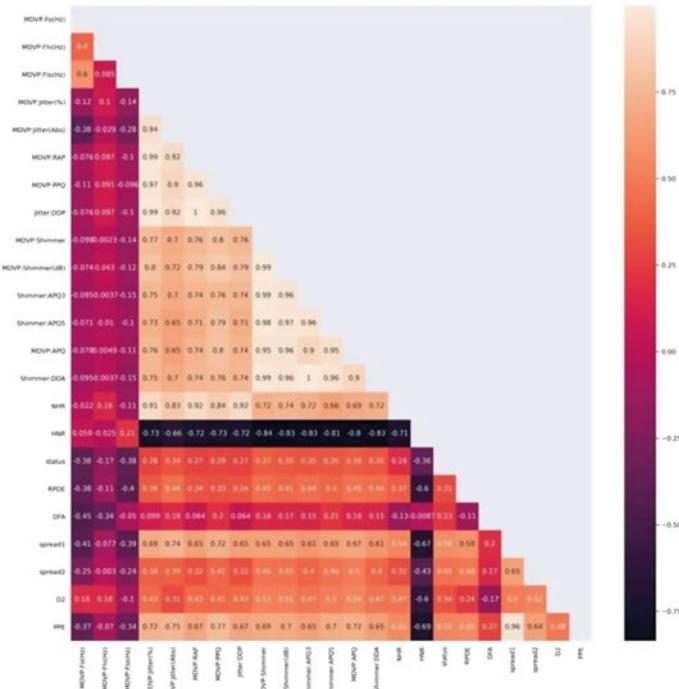


Fig. 1 Results of the correlation across the voice measurement attributes used in the study (depicted as a Heat map using seaborn library in python)

The next step is splitting the data into training and testing data. In order to have an accurate model, we have to use the same proportion of the positive and negative cases in both training and testing data. If our testing data is skewed toward negative results, then it would be hard for the model to predict positive cases. We mitigate this issue by splitting positive and negative cases in data sets, and in turn split them into train and test data sets. Then we combine the positive and negative data sets to get our training and testing data set which consists of 70% positive and 70% negative cases in the training data and 30% in the testing data.

We start to train different classifier models with the training data set that we created. The different classifiers used are random forest classifier, XGBoost classifier, SVM classifier model and decision tree classifier. We then test the accuracy of the models which we trained using the training data set. The accuracies of each model can be seen in Table 1.

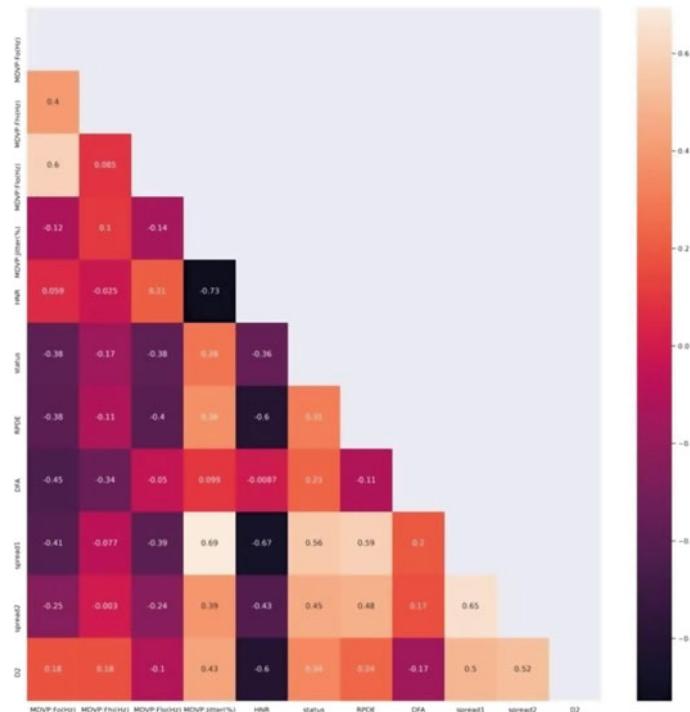


Fig. 2 The non-correlating voice measurement attributes selected for use in the ML models (depicted as a Heat map using seaborn library in python)

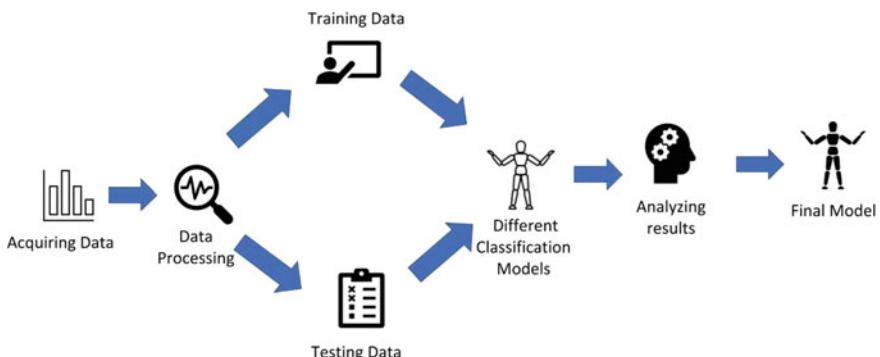


Fig. 3 System architecture of the proposed model

Table 1 Machine learning models and their accuracy in predicting PD

Sl. no.	Model	Accuracy
1	Random forest classifier	0.91666
2	XGBoost classifier	0.93333
3	SVM classifier model	0.85000
4	Decision tree classifier	0.85000

4 Results and Accuracy

(i) Random forest (RF)

This supervised algorithm can be applied for both classification and regression tasks and builds multiple classification decision trees (DT) employing random samples with a replacement to overcome the problem in DTs. We used classification tasks in this algorithm and set n_estimators to 101 to build the number of trees before taking the maximum voting or averages of the predictions. The RF-based prediction of PD had an accuracy of 0.91666.

(ii) eXtreme Gradient Boosting (XGBoost)

This algorithm is a method of ensemble learning, which uses gradient boosting decision trees to increase the computational speed and the model's performance. It is engineered to efficiently use computation time and memory resources to train the model. XGBoost dominates structured or tabular data sets on classification and regression tasks. The XGBoost-based prediction of PD had an accuracy of 0.93333.

(iii) Support Vector Machine (SVM)

For both classification and regression errands, this machine learning algorithm can be used. This calculation graphs each information element as a point in n-dimensional space with the approximation of a specific coordinate being the estimate of each element. The calculation distinguishes the two groups by locating the hyper-plane (Healthy/PD). The SVM-based expectation of PD had an exactness of 0.85000.

(iv) Decision Tree (DT)

For classification and regression assignments, this machine learning algorithm may be used. By learning basic decision rules derived from the data features, the algorithm predicts the calculation of a target variable. It characterizes the models by organizing them down the tree from the root to any leaf node; the leaf node classifies the ascribes used in the analysis. The SVM-based forecast of PD had a precision of 0.85000 (Fig. 4).

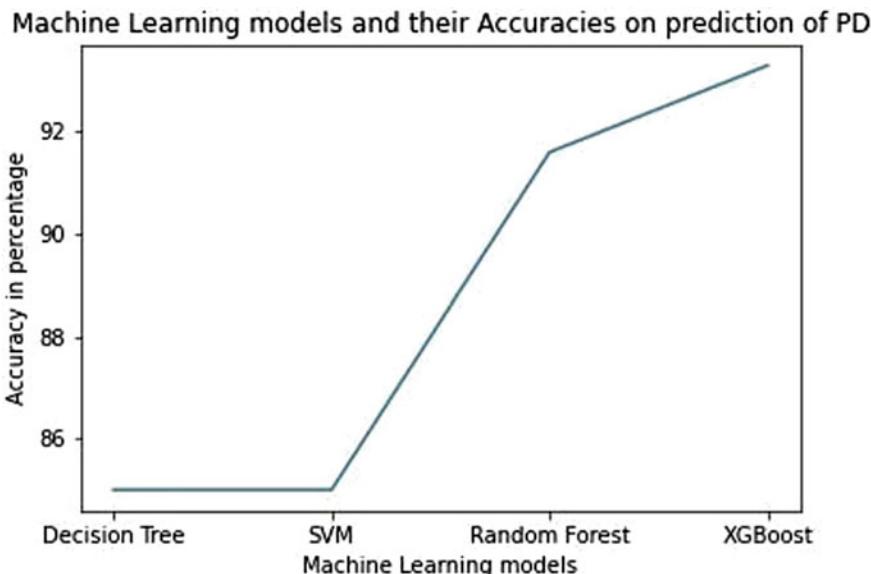


Fig. 4 Graph explaining the accuracy of all the classifiers studied

5 Conclusion

Parkinson's disease is a progressive nervous disorder which if detected in the early stages helps medical professionals deal with them better and make the patient's life better. In the study, we have analyzed biomedical voice and employed different machine learning models for its detection. Among all the classifiers, it was found that XGBoost (eXtreme Gradient Boosting) provided us with an accuracy of 93.333%. Based on this we recommend that the XGBoost technique should be used to develop a model for Parkinson's disease detection problems. Even though the model works efficiently, this is limited to the richness of the data set that was used to train this model. The selected data set that was used in this study had 195 instances; hence a data set with more instances would help to generalize better. This proposed model is a reliable model to detect Parkinson's disease due to its high accuracy.

References

1. Challa, K.N.R., Pagolu, V.S., Panda, G., Majhi, B.: An improved approach for prediction of Parkinson's disease using machine learning techniques. In: 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi, pp. 1446–1451 (2016)
2. Patra, A.K., Ray, R., Abdullah, A.A., Dash, S.R.: Prediction of Parkinson's disease using Ensemble Machine Learning classification from acoustic analysis. J. Phys.: Conf. Serv. **1372**, 012041

3. Akshay, S., Vincent, K.: Identification of Parkinson disease patients classification using feed forward technique based on speech signals. *Int. J. Eng. Adv. Technol.* **8**(5), 1769–1778 (2019)
4. Bala, N.D., Anusuya, S.: Machine learning algorithms for detection of Parkinson's disease using motor symptoms: speech and tremor. *Int. J. Recent Technol. Eng.* **8**(6), 47–50 (2020)
5. Shamrat, F.J.M., Asaduzzaman, M., Rahman, A.S., Tusher, R.T.H., Tasnim: A comparative analysis of Parkinson disease prediction using machine learning approaches. *Int. J. Sci. Technol. Res.* **8**(11), 2576–2580 (2019)
6. Nutakki, C., Narayanan, J., Anchuthengil, A.A., Nair, B., Diwakar, S.: Classifying gait features for stance and swing using machine learning. In: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, pp. 545–548 (2017)
7. Nutakki, C., Bodda, S., Diwakar, S.: Correlations of gait phase kinematics and cortical EEG: modelling human gait with data from sensors (2020). <https://doi.org/10.5772/intechopen.88465>
8. Nutakki, C., Mathew, R.J., Suresh, A., Vijay, A.R., Krishna, S., Babu, A.S., Diwakar, S.: Classification and kinetic analysis of healthy gait using multiple accelerometer sensors. *Procedia Comput. Sci.* **171**, 395–402 (2020)
9. Yuvaraju, E.C., Rudresh, L.R., Saimurugan, M.: Vibration signals based fault severity estimation of a shaft using machine learning techniques. *Mater. Today: Proc.* **24**, 241–250 (2020)
10. Muhammed, M.A., Aravindh, J.: CNN based off-the-person ECG biometrics. In: 2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET), pp. 217–221. IEEE (2019)
11. Krishnakumar, P., Rameshkumar, K., Ramachandran, K.I.: Acoustic emission-based tool condition classification in a precision high-speed machining of titanium alloy: a machine learning approach. *Int. J. Comput. Intell. Appl.* **17**(03), 1850017 (2018)

Social Distancing and Crowd Density Distribution System for Public Places and Public Transports Using Computer Vision and NLP



Sandeep K. Sharma, Rajiv K. Modanval, Prakhar Tayal, and N. Gayathri

Abstract The world population is growing day by day but the resources are limited. So, on one hand, we have to fulfill our requirements by using resources in an optimized manner, and on the other hand, we are currently in a pandemic for which social distancing is the key to prevent ourselves from the COVID-19 virus. We all travel by bus, train, metros, and other public transports, and we all have faced the problem of overcrowding in public transport once in our life especially during any festive season when there is a lot of rush among people to go from their workplace to their hometown. The problem of overcrowding is very common in developing countries like India where the population is relatively higher than the number of seats in public transports. For instance, if we look at the Delhi metro in which it is almost impossible to travel during the festive season. However, it's a regular problem, but the crowd becomes more when any festive season approaches. And due to such rush, people tend to just get inside the metro without even searching for a relatively less crowded or vacant seat/compartment, resulting in an unusual pattern of density seen across the metro, somewhere empty and somewhere highly crowded that even no place to stand. And now due to the COVID-19 pandemic, the situation worsens, on one hand, we are bound to travel in-crowd for work and on the other hand, we have to maintain social distancing for COVID prevention. So it becomes almost impossible to travel through such public transport. To solve such problems we are proposing a solution, that is, a system that takes care of the public places/transports by scanning and informing passengers where they can move to find a seat or less crowded. It also helps to ensure social distancing in the times of the COVID-19 pandemic.

Keywords Crowd management system · Social distancing system · Public transport · Computer vision · AI · NLP · COVID-19 · Pandemic

S. K. Sharma (✉) · R. K. Modanval · P. Tayal · N. Gayathri
School of Computer Science, Galgotias University, Greater Noida, India

N. Gayathri
e-mail: n.gayathri@galgotiasuniversity.edu.in

1 Introduction

The population is growing exponentially, so does the needs but we have a very limited amount of resources. The higher population is the biggest problem for any developing nation, as the growth rate of the population is relatively higher that puts immense pressure on the available resources. If this pace is not controlled then it becomes very hard for our future generations to live on earth. That's why we should opt for sustainable methods while using resources and use them in an optimized manner.

The system proposed is to support the mitigation of the spread of the novel coronavirus as well as reducing the crowd at public transports or public places by maintaining the crowd density across it. The system comprises a social distancing and crowd density system that uses person detection algorithms [1] so as to give accurate results. Apart from that, the system also facilitates various public transports to evenly distribute the crowd across it, thereby not putting any pressure on a particular compartment or coach. In case the compartment is over-crowded and there is no place to sit the door automatically closes so as to prevent further intrusion of the crowd. Everything is displayed on the public transport's display system and an announcement is made via the sound system of public transports.

Nowadays every public transport has cameras which are utilized to scan the vehicle and that scanning can find some interesting use cases, including detecting whether social distancing is followed or not, to get to know if the transport is less crowded, moderately crowded or highly crowded, etc. These use cases are put together to form a system known as "Social distancing and crowd density distribution system". Some of the key concepts that are utilized for the purpose of making the system work are the NLP (acronym for natural language processing), computer vision (CV), and microcontroller programming.

In this paper, we tried to put forward a system which is devised by combining some of the existing systems with a microcontroller chip that is better utilizing the systems for another great use case, i.e., maintaining the crowd distribution [2] in transport and public places keeping in mind the health of travelers/people and preventing the risk of spreading COVID-19 infection without the need of any additional set of hardware.

This paper is organized as follows:

Section 2 deals with the literature review in which comparative studies have been done based upon existing systems. Table 1 differentiates the different proposed techniques from different papers with their merits and demerits. Table 2 differentiates the accuracy of different models for this type of system. Section 3 deals with some of the general terms that are used across the paper. Section 4 deals with the preliminaries and description of the proposed system with a diagrammatic representation and flow diagrams. In Sect. 5, all the experimental aspects of the system are demonstrated diagrammatically. The proposed pseudocode and flow diagrams are described in Sect. 6 in which both the systems are described algorithmically. The feasibility

Table 1 Different kinds of proposed techniques with their merit/demerit

References	Publication year	Proposed technique	Demerits	Merits
[8]	2016, 2017	PIR-based human detection system	Unable to distinguish between significant and insignificant movement	Reduce wastage of electricity
[9, 10]	2017	RSSI-based human detection system	Processing time was too high that affects the accuracy	Reduction of electricity in campus, office
[11]	2019	Real-time passenger detection system	Unable to overcome contact protecting	Weight handling on the elevator to reduce accidents
[12]	2015	Safe driving system	Low accuracy	Reduce road accidents in traffic areas

Table 2 Accuracy of different available model

Model	Detected object	Time of execution (in sec)	Accuracy (%)
SSD Inception V2 COCO	2	298.22	97
SSD Mobilenet V1 COCO	2	219.58	94
Mask RCNN inception resenet V2 atrous	5	6008.02	99
Faster RCNN inception V2 COCO	3	420.71	99

analysis of the system done in Sect. 7 and different experimental results/outputs and conclusions are discussed in Sect. 8 and 9, respectively.

2 Literature Reviews/Comparative Study

Social distancing [6] is the biggest weapon to fight against this pandemic; some people are understanding it and taking care of it, but some people really don't care about it. They want to maintain social distance and other such kinds of precautions to fight against this pandemic, but they were not able to do it while traveling due to the crowd. Many researchers have realized this problem and come up with their solutions. Some solutions are very accurate and some have less accuracy but better compatibility. For example, one research was done by three researchers, namely Afiq Harith Ahamad, Norliza Zaini, and Mohd Fuad Abdul Latip from Malaysia, and the research study was published in 2020 at 10th IEEE International Conference on Control System, Computing and Engineering (ICCSCE) [13]. They claimed that

their model is up to 99% accurate in special conditions. The research is mainly based on movement detection in an area. If there is no movement in that area then the light of that area gets turned off automatically. They took this concept as the basis for their research.

2.1 Automatic System-Based Human Detection

An examination [14] introduced programmed control of intensity supply in the study hall, where movement identification was utilized as the principal sign for distinguishing human presence. On receiving picture preparation, foundation deduction was utilized to identify the movement made by the human presence. The strategy proposed is to separate between two pictures and casings at the pixel level. Foundation deduction measure utilizes one picture to be the reference picture and go about as a foundation picture. The restriction of this strategy is when there is insignificant movement, for example, solid breeze and creatures will likewise be considered as “human presence”.

Khedkar and Malwatkar [15] introduced an examination that utilizes RPi as the primary segment for controlling a computerization framework that is incorporated with GSM for correspondence. This investigation is performed to the extent of a home robotization framework that utilizes a concentrated regulator for controlling the electrical machines in the house. RPi has been a standard subject furnishing accessibility in managing mechanization. It tends to be utilized as a regulator in any electronic and electrical computerization with the wide adaptability in interfacing with numerous outer electronic gadgets. With the adjustable component of RPi and the upheld Python programming language, it is appropriate for robotization as an assortment of open-source libraries that are accessible.

2.2 Area-Based Human Detection

The research [8] focused on human habits, like leaving classrooms, personal rooms without switching off the lights, air conditioner, fans, etc., which results in exhaustive use of electricity. To control this wastage of electricity, the whole system should be automatic. To implement it, researchers used Arduino as a controller in a project. In that project, the area was divided into grids which are responsible for controlling fans and light of that particular area based on human motion. Detecting humans in that particular area is used as passive infrared technology, also known as PIR. The disadvantage in utilizing the PIR sensor is heat touchy. It distinguishes different wellsprings of warmth produced including sun heat and some other radio-recurrence radiation. Other than that, the discharged infrared radiation can be handily obstructed

by other items and it relies upon the speed of the dropping article which cuts down the precision in recognizing the human presence [9].

To overcome the above problem and to increase accuracy, image processing is used to detect human presence in a particular frame [10]. But in this approach the target is to detect the absence of humans in a particular area/frame. The targeted fields are classrooms, offices, houses, etc. This research proposed an RSSI-based human presence recognition framework for energy-sparing computerization. A received signal strength indicator (RSSI) is a sign for strength estimation of the sent sign by the passageway. In this investigation, a human presence location strategy is created by utilizing the example of RSSI understanding worth. The framework has a deferral in the time reaction, where the module takes 40 s for 20 readings. The technique proposed is appropriate in little and moderate rooms as the Wi-Fi signal proliferation is better inside the modest space. Such impediment prompts utilizing picture preparation as the human presence identification to acquire precision in distinguishing humans.

In Fan et al. [11] research, a technique to avoid accidents and overloading in elevators while people use them is proposed. In this process, a number of humans counted in real time to calculate the weight of humans using image processing. In view of the directed investigation, the creators found the method was unacceptable. This is on the grounds that the form of the human body may have contact protection that makes the framework unfit to precisely include the individual in the elevator.

Hariyono and Jo [12] proposed to distinguish and restrict the person on foot from moving vehicles concerning the street region and relative separation from the vehicle. They utilized the area characterization to separate between the immobile foundation and moving closer view. The pictures are grouped into two districts, which are the assessment street zone and the zone where the walker should stroll with the proposed strolling human model. The dataset utilized in this task is removed from Caltech and ETH [13, 14], web, and creator's own pictures. The area of the picture taken is separated into two districts which are the street zone got from the street path limits. Area arrangement will decide and recognize the person on foot from the framework. From the examination, the strategy creators utilized in their exploration the identification of human presence dependent on the district order which can be actualized by separating the auditorium and research facility zone.

2.3 Some Pre-trained Models

As indicated by the performed researches, most existing article recognition applications are utilizing pre-prepared CNN-based models. Thus, this examination is planned to utilize a comparative methodology, so the further investigation has been done to decide the most reasonable pre-trained model for this venture. Table 2 shows the accuracy of different models [13].

3 Some General Terms

3.1 What is AI?

Artificial intelligence (AI) [3] is a large branch of computer science that deals with the creation of smart machines capable of performing tasks that usually require human intelligence. AI is an interdisciplinary science with various methods.

From decision-making to the community to bottom-line advantages, AI has the ability to greatly enhance the way your company operates. But don't make the mistake of only implementing AI in a few instances of customer use, feeding it with details once a week, and locking it away from the rest of the company.

3.2 Types of AI

There are four types of artificial intelligence:

- Reactive machines
- Limited memory
- Theory of mind
- Self-awareness.

3.3 What is an AI Technique?

AI problems have little in common except for the fact that it is hard. One of the few results to come out in the first three decades of AI research is that intelligence requires knowledge.

Knowledge possesses some less desirable properties, which includes:

- It is constantly changing
- It differs from data by being organized in a way that corresponds to the ways it will be used.

AI technique is a method that exploits knowledge that should be represented in such a way that:

- The knowledge captures generalization.
- It can be understood by people.
- In many AI domains, most of the knowledge a program has must ultimately be provided in terms they understand.
- It can be easily modified to correct errors.

The kind of human capabilities that should be provided in AI entailed system are:

- Learning
- Understanding
- Recognition
- Decision-making
- Adaptability
- Knowledge base.

3.4 Natural Language Processing

The field of computer science that focuses on interactions between humans using spoken language and computers is natural language processing (NLP) [4]. Huge amounts of natural language data are processed by computerized speech recognition and analytics and conclusions are drawn.

Natural language processing (NLP), in short, helps computers to interpret, understand, and infer meaning from human languages.

3.5 Applications of NLP

- Classify text into categories
- Index and search large texts
- Automatic translation
- Speech understanding: Understand phone conversations
- Information extraction: Extract useful information from resumes
- Automatic summarization: Condense 1 book into a single page
- Question answering
- Knowledge acquisition
- Text generations/dialogues
- Language Translator
- Chatbots
- Voice Assistants
- Grammar Checkers.

3.6 Computer Vision

Computer vision [5] is a field of artificial intelligence that trains computers to interpret and understand the visual world using digital images from cameras.

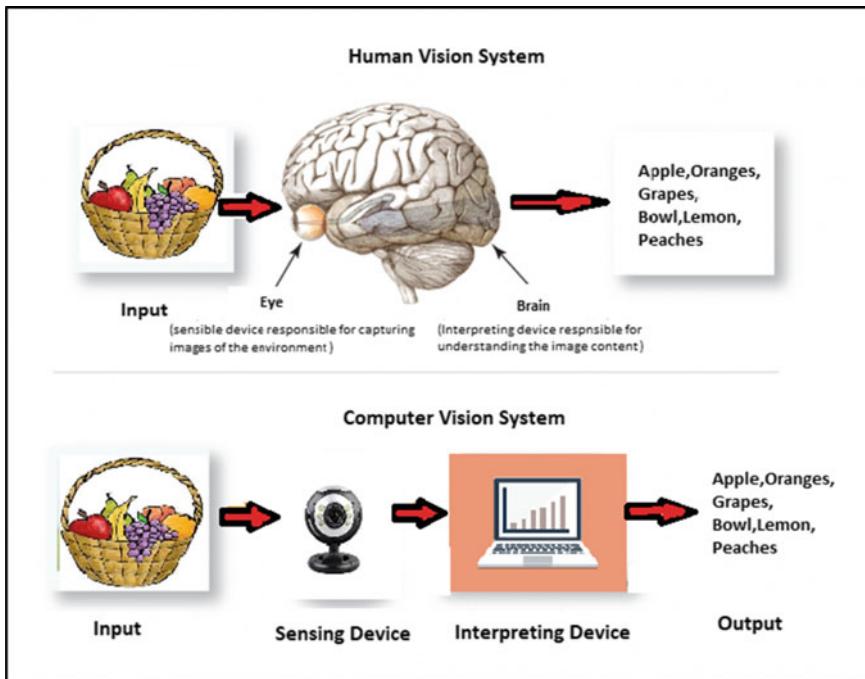


Fig. 1 Computer vision versus human vision

Application of computer vision:

- Object detection
- Classification
- Security
- Education
- Object tracking
- Robot and driverless car navigation and control (Fig. 1).

4 Proposed System Overview and Preliminaries

Our proposed method utilizes computer vision and NLP, where computer vision is used to detect the density of the crowd and whether the person is managing social distancing or not. NLP is used for announcements of observation detected by the machine learning model via a speaker to maintain social distancing and tell about the empty spaces in other compartments.

4.1 Design Requirements of Social Distancing and Crowd Density Observation System

A. Basic Flow Diagram

The basic flow diagram of our proposed method is depicted in Fig. 2.

In Fig. 2 the ML model is connected to each node, i.e., input device for taking inputs and based upon the knowledgebase predict accordingly and return through the output device. This depicts the basic flow diagram of the ML model.

B. Detailed flow diagram

Figure 3 demonstrates the working of our proposed social distancing and crowd management system. In this system, we use the cameras as an input device that are generally attached in public transports, and that camera is attached with a microcontroller chip which contains the trained model for recognizing people and density calculator based upon which it recognizes the safe distance between two or more persons. Based upon the observation it utilizes NLP to announce the public to maintain the safe distance and tell them where the crowd is less so as to distribute evenly across the transport. The microcontroller chip that is attached to the input and output devices is connected with the cloud via a gateway. In our model, everything is happening in the cloud from the training model to validation and storage. The training that happens across the cloud utilizes a data set for training purposes and after the training is done with at most accuracy the model gets stored in the database and also a copy is sent to the microcontroller chip. The basic components of our system are a social distancing system and a crowd management system that contain various tools and techniques attached to the cloud and the data is transmitted to and fro so as to give accurate results.

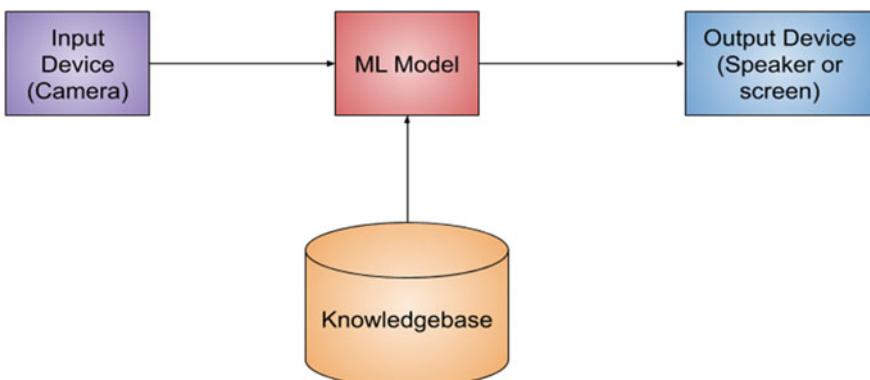


Fig. 2 Basic flow diagram of the proposed idea

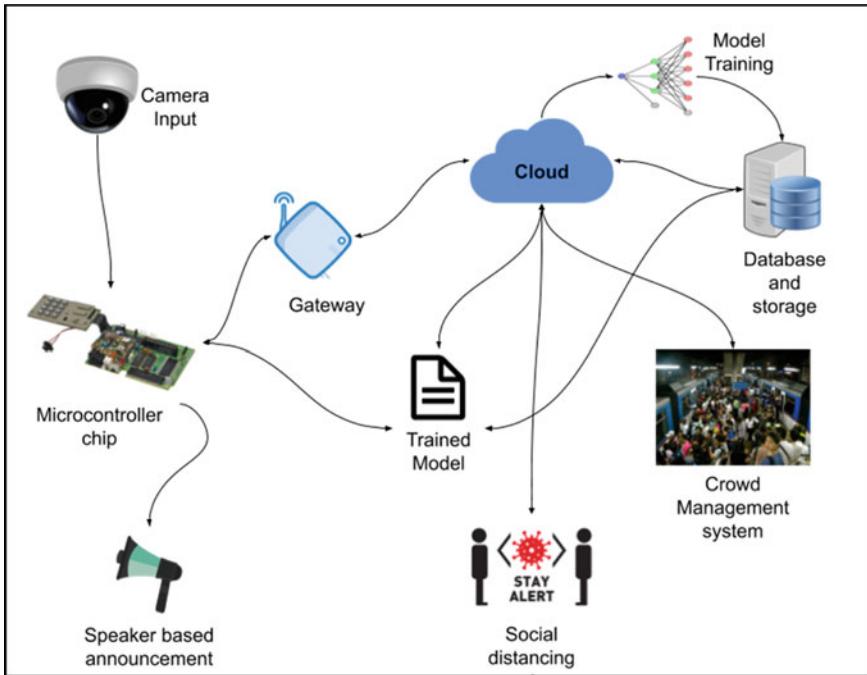


Fig. 3 Social distancing and crowd management system

4.2 Design Goals

The proposed system is based on various design goals, including:

- *Modular and Embedded Design:* The design consists of a simple microcontroller chip embedded inside a tiny box that can be easily installed inside the public transport with direct connection with cameras, sound, and display system of the transport.
- *Upgradability:* The microcontroller installed in any system can be easily upgraded to the newer versions as the chip can automatically connect with the public Wi-Fi system to fetch the latest updates. The key benefit is that it only requires few bytes of data for upgradation and only requires connectivity once a week.
- *Maintainability:* The system design is so much easy to maintain and does not require much maintenance as it doesn't require additional power, but it consumes the power from within the vehicle.
- *Scalability:* As the population is continuously in growth which requires additional systems to track on, the system comprises additional slots that can be utilized for making it more scalable in case of need.

5 Experimental Design

5.1 Implementation and Description of Project Modules

The implementation of our project goes like this: We will use the CCTV camera of public transport [7] as an input device that will capture continuous images and pass it to the microcontroller where the machine learning model recognizes and observes the social distancing and crowd density and trigger announcements as per our observation. If the compartment is found to be so much crowded and another compartment of the same vehicle is found to be less crowded, then our system would suggest the crowd to distribute evenly across that less dense point. Our system will also observe whether a person or group of persons is in the safe distances or not. If they found to be violating the social distancing norms then the announcement will be triggered suggesting them to move apart. If the vehicle found to be so much crowded, then our system would suggest other passengers kindly wait for another vehicle and don't board onto the vehicle as it is for their own benefit only.

For the demonstration purpose, we will use a laptop with any configuration connected to the internet (Fig. 4).

The two main components of our system are:

1. Social distancing system
2. Crowd density management system

Each component has subcategories and both are connected to the microcontroller chip via the cloud where training for this type of model happens, and the trained model gets stored into the microcontroller chip.

Fig. 4 Demonstration setup



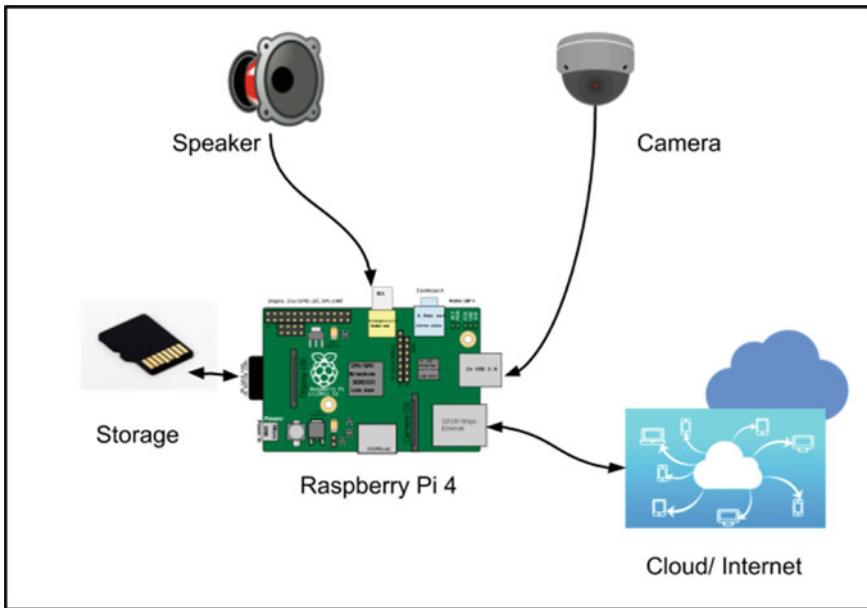


Fig. 5 Connection diagram of different components

Various models attached to the components are:

1. Vision module: the module utilizes computer vision [16] for the purpose to identify the density of the crowd and the distance between the two or more individuals.
2. Speech module: this module utilizes NLP, which stands for natural language processing, to make the announcement via speaker based upon the observation and ML model response.
3. Storage module: this module is attached with the microcontroller chip connected with the cloud, which in turn is connected to the databases from where the local storage keeps on updating based upon the trained model so as to perform optimally.
4. Network module: the model's main function is to keep connected to the device via the internet for data transmission. However, the internet is not important all the time; it is just for the sake of showing updates in optimal results.

The connection diagram of different modules is depicted in Fig. 5.

6 Flow Diagram and Pseudocode

See Fig. 6.

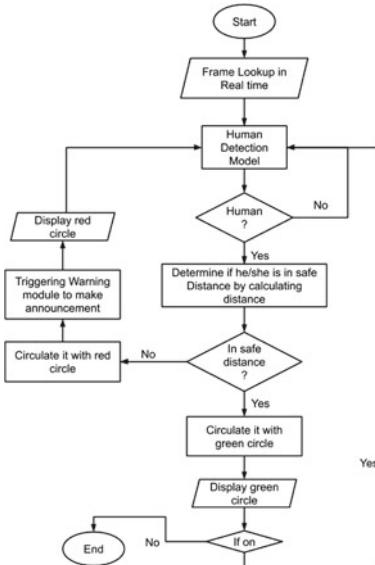


Fig. 6 Flow diagram for social distancing system

6.1. Pseudocode for Social Distancing system

Inputs: the coordinate axis of the objects in the frame.
safe distance: $D=1\text{m} * 3779.5\text{px}$ [$1\text{m}=3779.5\text{ px}$] as frames are in pixels.

Step1: while [ON] do recognize the human in the frame.

Step2: If any human in the frame then

Step3: Locate the human and its coordinate.

Step4: If multiple humans with coordinates $p1(x1,y1)$ and $p2(x2,y2)$ then

Step5: Calculate distance: $d=\sqrt{(x2-x1)^2+(y2-y1)^2}$

Step6: If $d>D(\text{safe distance})$

Step7: Display green circle

Step8: end If

Step9: Display red Circle

Step10: Trigger warning using audio

Step11: end If

Step12: Announce to maintain a safe distance

Step13: end If

Step14: end While

Step15: Stop

See Fig. 7.

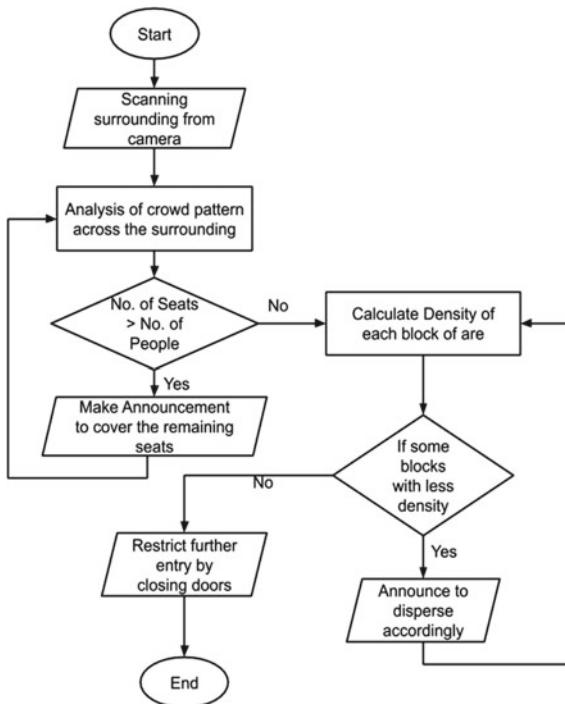


Fig. 7 Crowd density distribution system

6.2. Pseudocode for Crowd Density Distribution System

Inputs: real-time scanning of the crowd in the frame with the number of **headcounts:** $N(h)$ and **seat count:** $N(s)$.

Step1: for humans in the frame

Step2: If $N(s) > N(h)$ then

Step3: Locate the seat and make an announcement to cover it.

Step4: else

Step5: Calculate the density of the crowd in each block with area A.

$$D_p(\text{density of crowd}) = N(h)/A$$

Step6: If some blocks density $D_b < D_p$ then

Step7: Trigger announcement module and locate the blocks to disperse accordingly.

Step8: else

Step9: Restrict further entry by closing the doors
Step10: end If
Step11: Announce and display the density pattern across the transport or area.
Step12: end If
Step13: end for
step 14: Stop

7 Feasibility Analysis

In this pandemic we have to fight against COVID-19 mainly with two weapons: social distancing and washing hands at least for 20 s. Maintaining social distancing is the biggest challenge in a crowd. We are trying to solve this problem. Our project has many pros over some cons. These are as follows:

- It helps to identify the crowded and uncrowded areas.
- It helps to shift the crowd to empty space.
- It has an NLP-based voice feature that makes it comfortable with all types of languages.
- Our system will also work after this pandemic. It will help passengers to find seats based on availability.
- Our system can detect fire and any type of unnatural events that will help to minimize the loss.

8 Experimental Results and Discussion

The experimental setup consists of raspberry pi with a high-resolution camera and audio and display as the output unit. The camera captures the picture in real time and the image is transferred to the microcontroller for processing and analysis so as to trigger the output signal accordingly via output devices, including sound and display unit. The results observed by the experimentation setup are depicted in Fig. 8.

Cameras available at public places and public transports are of high resolution and can easily capture the area so as to detect if a person is following social distancing norms or not. If the person is at a safe distance from others they are encircled green by the system, whereas the person who is not in the safer limits are marked in red circles and are warned by the audio systems installed at that area so as to prevent the spread of the COVID-19 virus and further break the chain of transmission so as to contain it (Fig. 9).

The working of the crowd density management system is the same as the social distancing system. The only difference is that the system calculates the density of

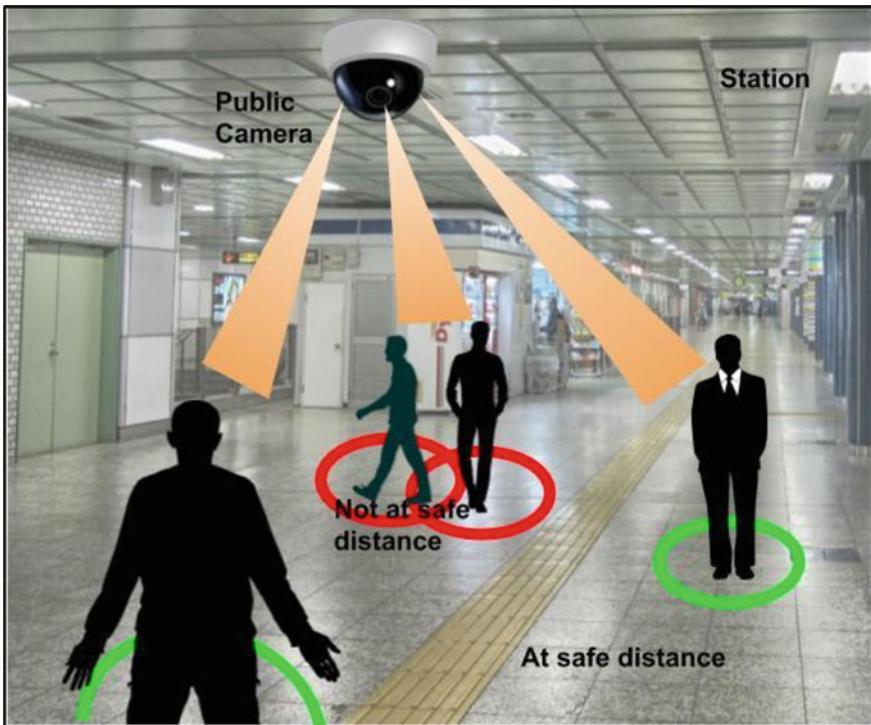


Fig. 8 Working of social distancing system

population and then acts upon it by classifying the crowd as least crowded, moderately crowded, or highly crowded. While the less or moderate crowd is somewhat allowed until a certain extent by throwing some simple announcement, the situation of overcrowding should not be tolerated as they possess the greatest risk of spreading of the virus. Also, it can be critical for many people in terms of their health. So to prevent overcrowding the system generates warnings and announces passengers to disperse and prevent further crowd by closing the automated doors (Fig. 10).

The above-scattered chart displays the comparison between the number of real existing cases versus the number of cases that could be if the system similar to our proposed system would have been utilized. Since it was very much a sudden pandemic and the virus was so new that no one would have thought of it, a lesson to learn from this is that everything is uncertain around us. So we must prepare for any type of uncertainty.

Our proposed system that is designed keeping in mind the upgrade and scalable functionality can be utilized today to contain further spreading and also prevent us from future uncertainties.



Fig. 9 Working of crowd density management system

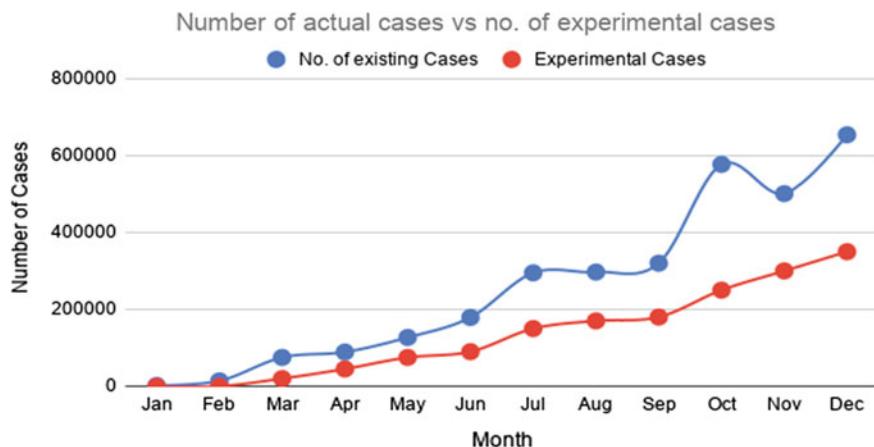


Fig. 10 Number of existing cases versus no. of experimental cases

9 Conclusion

In this paper, we have presented the social distancing problem as the estimation and characterization of interpersonal distances from images. Solving such problems allows quick screening of the population for detecting potential behaviors that can cause a health risk, especially related to recent pandemic outbreaks. The proposed system delivers information to people for making social distancing who are not obeying it and tells the importance of masks and social distancing. Thus, this proposed system would work in an efficient manner after the lockdown period ends and helps in easy social distance inspection in an automatic manner. The algorithm can be embedded in public cameras and then details can be fetched to the camera unit same as the drone unit which receives details from the drone location details and stores them in a database. Thus, the proposed system favors society by saving time and helps in lowering the spread of coronavirus.

Acknowledgments This research was partially supported by N. Gayathri (Assistant professor, Galgotias University). The authors thank their team members (Sandeep Kumar Sharma, Rajiv Kumar Modanval, Prakhar Tayal) for their dedication toward work. We all are working together to complete this research.

References

1. Gupta, S., Kapil, R., Kanahasabai, G., Joshi, S.S., Joshi, A.S.: SD-measure: a social distancing detector. In: 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN), Bhimtal, India, pp. 306–311 (2020). <https://doi.org/10.1109/cicn49253.2020.9242628>
2. Lijun, C., Kaiqi, H.: Video-based crowd density estimation and prediction system for wide-area surveillance. *China Commun.* **10**(5), 79–88 (2013). <https://doi.org/10.1109/CC.2013.6520940>
3. Williams, C.: A brief introduction to artificial intelligence. In: Proceedings OCEANS '83, San Francisco, CA, USA, pp. 94–99 (1983). <https://doi.org/10.1109/oceans.1983.1152096>
4. Surabhi, M.C.: Natural language processing future. In: 2013 International Conference on Optical Imaging Sensor and Security (ICOSS), Coimbatore, pp. 1–3 (2013). <https://doi.org/10.1109/icoiss.2013.6678407>
5. Rosenfeld, A.: Computer vision: basic principles. *Proc. IEEE* **76**(8), 863–868 (1988). <https://doi.org/10.1109/5.5961>
6. An, L., Hawley, S., Lee Van Horn, M., Bacon, E., Yang, P., Resnicow, K.: Development of a coronavirus social distance attitudes scale. *Patient Education and Counseling* (2020). ISSN 0738-3991. <https://doi.org/10.1016/j.pec.2020.11.027>
7. Khoudour, L., Deparis, J.P., Duvieubourg, L.: Linear image sequence analysis for passengers counting in public transport. In: 1996 International Conference on Public Transport Electronic Systems (Conf. Publ. No. 425), London, UK, pp. 100–104 (1996). https://doi.org/10.1049/osi_cacp:19960462
8. Suresh, S., Anusha, H.N.S., Rajath, T., Soundarya, P., Vudatha, S.V.P.: Automatic lighting and control system for classroom. In: Proceedings of 2016 International Conference on ICT in Business Industry and Government ICTBIG 2016 (2017)

9. Rao, V.S., Kowshik, S.S.: PIR sensor characterization and a novel localization technique using PIRs. In: Proceedings of the 14th International Conference on Information Processing in Sensor Networks (2015)
10. Habaebi, M., Rosli, R., Islam, M.R.: RSSI-based human presence detection system for energy saving automation. Indones. J. Electr. Eng. Informatics **5**(4), 339–350 (2017)
11. Fan, H., Zhu, H., Yuan, D.: People counting in elevator car based on computer vision. IOP Conf. Ser.: Earth Environ. Sci. **252**, (2019)
12. Hariyono, J., Jo, K.H.: Detection of pedestrian crossing road. In: Proceedings International Conference on Image Process. ICIP, vol. 2015, December, no. January, pp. 4585–4588 (2015)
13. Ahamad, A.H., Zaini, N., Latip, M.F.A.: Person detection for social distancing and safety violation alert based on segmented ROI. In: 2020 10th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), Penang, Malaysia, pp. 113–118 (2020). <https://doi.org/10.1109/iccsce50387.2020.9204934>
14. Ganiger, D., Patil, K.A., Patil, P., Anandhalli, M.: Automatic control of power supply in classroom using image processing. In: Proceedings—2017 International Conference on Recent Advances in Electronics and Communication Technology, ICRAECT 2017, pp. 230–234 (2017)
15. Khedkar, S., Malwatkar, G.M.: Using Raspberry Pi and GSM survey on home automation. Int. Conf. Electr. Electron. Optim. Tech. ICEEOT **2016**, 758–761 (2016)
16. Dandil, E., Çevik, K.K.: Computer vision based distance measurement system using stereo camera view. In: 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, pp. 1–4 (2019). <https://doi.org/10.1109/ISMSIT.2019.8932817>

TV Viewing Behaviour: Analysis Using Machine Learning Algorithms



C. Karthika, A. G. Hari Narayanan, and P. P. Vijayalakshmi

Abstract Television is becoming an important part of our everyday life as a mass medium for communication. With its dramatic and demonstrative strength, to change it is a true source of knowledge, education and entertainment. It is convincing to take into account the promotion of social values and norms in a civilized society. The structure and shape of the shape play a major role. Here, we propose a new method called multi-class classification with the help of big data analytics. In this study, we have provided a framework for comparing the quality of various classification methods using statistical simulation when individuals belong to one of the two groups that are mutually exclusive. Here we compare naïve Bayes classification, multi-layer perceptron classification and decision tree classification as a test case. From the results, we found that the classification accuracy of multi-layer perceptron classification is higher than the other two in analysing the television viewing behaviour. The data for the study is collected through a direct questionnaire survey. Here, accuracy is assessed over correctly and incorrectly categorized instances. The analysis found that MLP is the best algorithm to predict TV viewing behaviour.

Keywords First keyword · Second keyword · Third keyword. Television · Viewing behaviour · Machine learning · Classification algorithms · Big data analytics · Naïve Bayes classification · Multi-layer perceptron classification and decision trees classification

C. Karthika (✉)

Department of Visual Media and Communication, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Kochi, India

A. G. Hari Narayanan

Department of Computer Science and IT, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Kochi, India

P. P. Vijayalakshmi

Department of English and Languages, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Kochi, India

1 Introduction

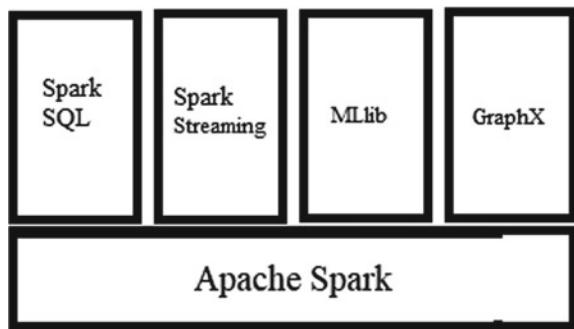
Television has been the largest mass media since the second half of the twentieth century. Globally, the exposure to TV entertainment programs is growing [1, 2]. The audience will draw a range of programmes, which require our eyes, ears and minds. It is capable of having a colder, seductive and compelling effect on our lives than other media [3]. It is important to note that audiences are concerned with it. It has the ability to affect viewers' lives and change their living standards. It contributes to visions and dreams.

The effect of television on culture is penetrating. The culture is transformed by new programs. It has improved social values in some way. The only cause of socialization and social learning should not be called television. Other sources, including parents, siblings, teachers and peer groups, have shaped the social learning process. The rapid technological development and rapid living have brought various changes to the content and distribution patterns of television programmes. Till last year, in India, the BARC India provided approximate and controlled TV audience measurement data for the television audience. They tell India is a country with 197 trillion TV viewers (Çığşar and Ünal [4]). Now, there are controversies regarding the data provided by BARC, and some investigations are going on it. In order to evaluate the target markets of different services, their desires and expectations, the press industry often seeks the measurement of the public so as to optimize the material and care accordingly. This helps to improve the programmes. This also helps advertisers to initiate or grow new ads depending on the target audience.

BARC's approach begins with a survey of organizations to collect particular demographic data of households and people. The following is a listing analysis in addition to the establishment study in order to include a more randomly selected sample of TV owners. After that, the control variables and the configuration of the panels are fixed. Panel management is responsible for the supervision of panel operators in compliance with the specified standards. Domestic measuring equipment is often used to measure the viewership and capture TV viewing events. The data processing, audience estimates and reports are then presented; the rules and validation are allocated; and the customers weighted, plugged and distributed in an understandable manner. The results are then given. In the current research, we intend to create a new method to calculate TV viewership and reach machine learning algorithms. This framework would provide data on the scale of the package to potential programme producers and advertisers.

As a result of the rapid and unavoidable growth of technology, the volume of data by various means is growing globally. Here in the case of TV, the data relating to the watching of a specific show easily and quickly take the media's entertainment role into account. Unbelievable audience levels can be obtained from the big data via television. The collection of broad data using appropriate strategies and techniques is close to the excavation of maximum admissible of newly discovered mineral.

Data mining is a tool for statistical data processing based on application. It aims to derive information from vast volumes of data not previously determinable [5]. Big

Fig. 1 Apache Spark model

data is a global focus for researchers. Interestingly enough, the entire world looks forward to its growth, because it is an important tool for business. For any company to achieve market leadership, efficient procurement and analysis of accessible large data is important.

Our research complements additional approaches in order to evaluate personal characteristics and target content based on these attributes. Decisions were proposed on theoretical and collaborative filtering methods for creating customized TV programme guides [6]. Here we also use machine learning techniques for the prediction (Fig. 1).

2 Related Works

Jain and Jakate's [7] study demonstrated the limitations of the TRP method at that time. They suggested a new approach to addressing these shortcomings. The proposed software is cost-efficient and flexible and can simultaneously allow the public to measure and analyse several shows. The public libraries and machine learning, massive data and social media have been used for research. In a study conducted by Splanger et al. [8] the selling of viewer profiles to advertisers typically poses legal questions regarding privacy. They can classify the target audience through these data. Therefore, it is very important to confidentially control the collection and use of the viewer information. To this end, it is important to develop a clear policy.

Research is being carried out on cumulative attention growth. Kaltenbrunner et al. [9] proposed in their analysis that news stories adopt a constant pattern of progress based on publication times. In Szabo and Huberman [10] a log-linear model surpassed constant growth models in relation to medium-squared errors (MSE).

In the study by Zhu et al. [11], the behaviour data of the television viewers were analysed. The researchers presented a method to predict the popularity of TV programmes. They used a dynamic time warping (DTW) distance-based *K*-Medoids algorithm to group programmes with similar popularity into four evolutionary trends, which has the ability to capture the inherent heterogeneity of program popularity.

The study by Kadam et al. [12] assesses the performances of TV shows and calculated the rating of each actor in the show based on the text reviews. Another recent study proposed an advanced prediction model that defined the performance analysis of the popular show of that time and found the factors that influence the popularity of the show. The study inferred that awards, nominations, characters are the predictive factors of popularity and appreciation [13].

The main focus of the preceding studies is on developing a general model to forecast the success of some media but to disregard the huge gap that arises as the popularity of content progresses. Thus, these methods are ineffective for predicting program popularity for broadcast TV, especially when predicting early peaks and subsequent popularity explosions. No other work explored the predictive capacity of features derived from an interactive programme guide to the best of our knowledge. A survival analysis model has been used to detect posts with over 100 comments on MySpace by Lee et al. [14]. This model has a precision of about 80%.

Kun et al. [15] proposed a big-data planning scheme focused on wireless big data computing in their report. The study of Anand et al. [16] suggested a method for forecasting TV popularity using two K-means and K-means algorithms with an accuracy of 97%.

In a study done in Kerala on viewing habits of news satires, it has been found that people are interested in watching such programmes to gain information through entertainment [17]. Another study inferred that the TV news on flood in Kerala was watched by women for several reasons, like gaining information regarding the flood, rescue operations etc., and they attained gratifications like relief and security through it [18].

From the current research survey, we find that the majority of individual services are being examined. But here we suggest a new approach known as the multi-class classification with the help of APACHE Spark with Jupyter.

3 Proposed System

TV is a perpetual entertainment tool. It delivers a variety of programmes from soap opera to news and satires. More than 500 raw data have been gathered by direct survey questionnaires. The study has the following parameters: “Age, Gender, Education, Income, Occupation, Area, Watch Time, Day/Night, Others Opinion, Watch While Having Food, Watch Commercial, Create Role Models, Top-singer, Saregama, Comedy Stars, Ningalkum Akam Kodeeswaran, Laughing Villa” are the labels collected from the respondents. These data are then coded in binary forms. Among these data, we take only the last five for this study, that is, the reality shows of different Malayalam channels into consideration. The programmes focused are Topsinger of Flowers Channel, Sa Re Ga Ma of Zee Keralam, Comedy Stars of Asianet, Ningalkum Akam Kodeeswaran of Mazhavil Manorama and Laughing Villa of Surya TV (Fig. 2).

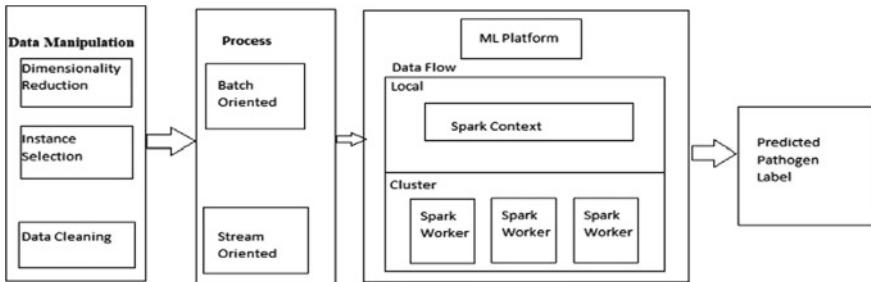


Fig. 2 Model for TV viewing behaviour analysis

Root Element

Age, gender, education and income programme lists.

‘Big data’ is a rapidly expanding term that defines the vast number of data. Analysis of data involves extracting useful data by identifying all possible relationships between various data. It makes huge data even bigger. The quantity of data called big data is too much for a personal data analyser.

Machine education has contributed to better big data management as a data care service. Consider a situation in which vast numbers must be compiled that is time-consuming and repetitive. By linking data, data cropping and data patterns, you get deeper into the understanding process. The results allow the company to quickly determine. Thus, companies may use machine learning for big data analytical research to unlock the potential for big data adoption.

4 Machine Learning Classification Algorithm

Classification is one of the most commonly used computer algorithms that almost everybody should know about. For our research study, after loading and preparing data, we used three classification algorithms.

Naïve Bayes classification

Multi-layer perceptron classification

Decision tree classification

Our data collection obtained from the static data explores the supervised classification algorithms.

4.1 Naïve Bayes Classification

We are ready to use the first classification algorithm when the data is prepared. naïve Bayes uses the idea of posterior likelihood to check for models as a classification technique. When the features are autonomous, naïve Bayes produces the best results. This model does not require more qualified analytical data. That is why it would be more rapid to forecast.

4.2 Multi-layer Perceptron Classification

A multi-layered perceptron is the second classifier we are investigating. The ANN.NN Multi-layer Perceptron Classification (MLP) maps the nodes in the output list in which both nodes are in separate modes. We must carry out weight-dependent calculations to render this node interaction. We must also recognize nodes that are hidden to increase map and prediction levels by helping to achieve precision.

4.3 Decision Tree Classification

Another common classifier in the ML family is the decision tree classifier. DT starts with a root node and then makes a number of branches (solutions), just like a normal tree. Independent prediction is done by the trees from the existing tree result. As a result, we can say that it will reduce the error value of mean, or in other words, it reduces the inconsistency in the classification.

4.4 Programming Model Steps

Here we are trying:

1. Load and parse the data input the dataset and do the cleaning process to make ready the dataset for execution. Here we used dataset that collected different features from the viewers to identify people behaviours. In that TV programmes are made as a group of list depending on the time what they view. Some are made with the popularity sense. Having the minimum information from a particular location grouping may be big task, which makes the system process execution not within a time bound to make the proper prediction.
2. Build and evaluate the model on training data.

For Naïve Bayes classification, two multi-layer perceptron classification, decision tree classification.

We used Spark model for this experiment. In this process four virtual machines are used inside Ubuntu. In that one system has a master node with Spark implementation files for making model. For achieving this, we are also made to execute in multiple nodes across the system. YARN did the memory management, and HDFS contains the data for this. For running Spark's API, the main task is to create a collected different features (programmes) grouping into frames that will assign for classification. These datasets may be the same or different categories as in the above dataset collection detail using Spark's libraries (MLlib, ML) for implementing our three classification models. Using this model we are having a few number of features. Multi-class type support is used here. For this prediction we are using three metrics: accuracy, precision and prediction.

Also, we will measure ROC curve for the identification of true and false positive rates. It ranges from 0 to 1. Large area predicts high positive and low false rates. PR curve for identification also used true or false, where the area used in the graph represents high, true and precision value for positive rates of viewership.

3. Save and load model.

5 Results

Here, we intend to find the best algorithms to identify TV viewing behaviour. There are studies in communication and journalism that predict the difference in TV viewing behaviour according to demographic variables. In this study, we take the demographic variables age, gender, education and income etc. as discussed earlier in our database design. We want to detect which algorithm predicts the difference in the viewing behaviour of these variables.

At first, overall database representation inside Spark-based ML system is done (Table 1).

From the analysis, MLP produces the best outcome for our study. It has been found that the naïve Bayes (80.2%) did the worst with a precision of percent in predicting viewing behaviour.

Table 1 Overall prediction of the dataset we collected from the below results

Method/Algorithms	Metrics		
	Accuracy (%)	Precision (%)	Prediction (ms)
Naïve Bayes	82.75	80.2	68.48
MLP	98.75	96.14	20.46
Decision tree	96.25	95.45	56.54

Table 2 Age-wise prediction of the dataset

Method/Algorithms	Age		
	Accuracy (%)	Precision (%)	Prediction (ms)
Naïve Bayes	81.2	79.3	65.4
MLP	96.2	94.6	18.5
Decision tree	95.8	93.7	54.5

Table 3 Gender-wise prediction of the dataset

Method/Algorithms	Gender		
	Accuracy (%)	Precision (%)	Prediction (ms)
Naïve Bayes	83.7	80.3	64.4
MLP	96.2	95.6	19.5
Decision tree	94.8	92.7	53.5

5.1 Prediction Group-Wise Result

Naïve Bayes of 79.3% did the worst with a precision of percent in predicting viewing behaviour difference age-wise; 80.3% did the worst with a precision of percent in predicting viewing behaviour difference gender-wise; 88.2% did the worst with a precision of percent in predicting viewing behaviour difference education-wise, and 85.2% income-wise.

But MLP of 94.6% was found best in the preparation and prediction times for the variable age; 95.6% was found best in the preparation and prediction times for the variable gender; 98.14% was found best in the preparation and prediction times for the variable education; 96.4% was found best in the preparation and prediction times for the variable income.

MLPs (18.5 ms) have been very successful with subtle edges and both worked similarly in terms of precision and training period of age. MLP of 21.4 ms was found best in the preparation and prediction times for the variable income; 19.5 ms have been very successful with subtle edges and both worked similarly in terms of precision and training period of gender; 19.46 ms was found best in the preparation and prediction times for the variable education.

Evaluation details of each main attribute and the result are given in Tables 2 and 3.

MLP did well in the areas under ROC curve and achieved an almost perfect score for age, gender, education and income. Figure 3 depicts the overall prediction as per our matrix. Figure 4 illustrates the True and False rate ROC curve for TV viewership dataset (Fig. 5 and Tables 4 and 5).

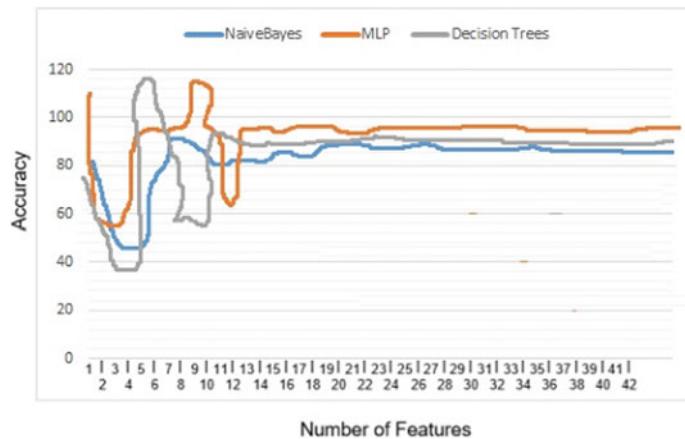


Fig. 3 Selection Graph I diagram for overall prediction accuracy versus features

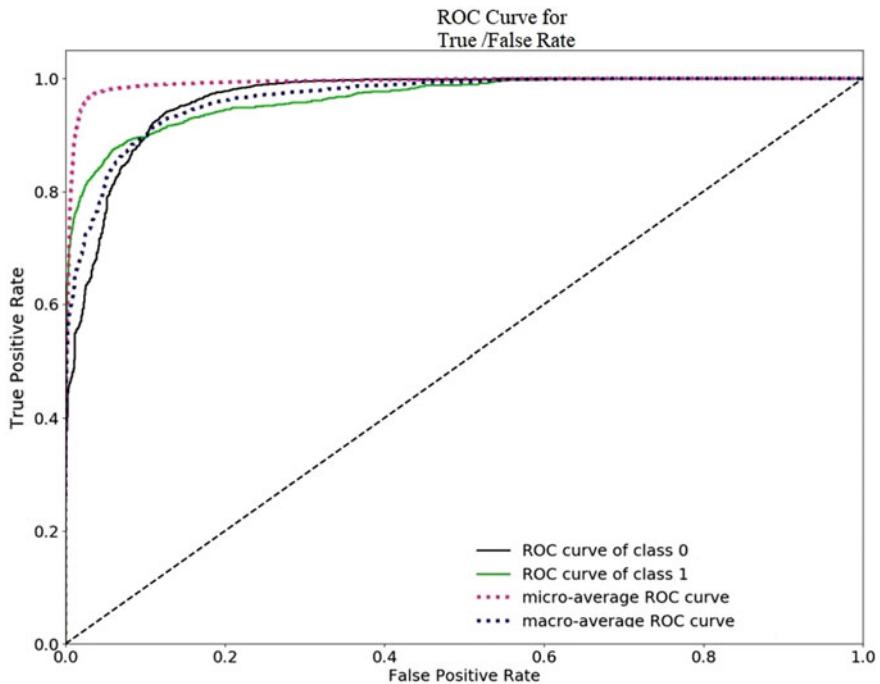


Fig. 4 True and false rate ROC curve for TV viewership dataset

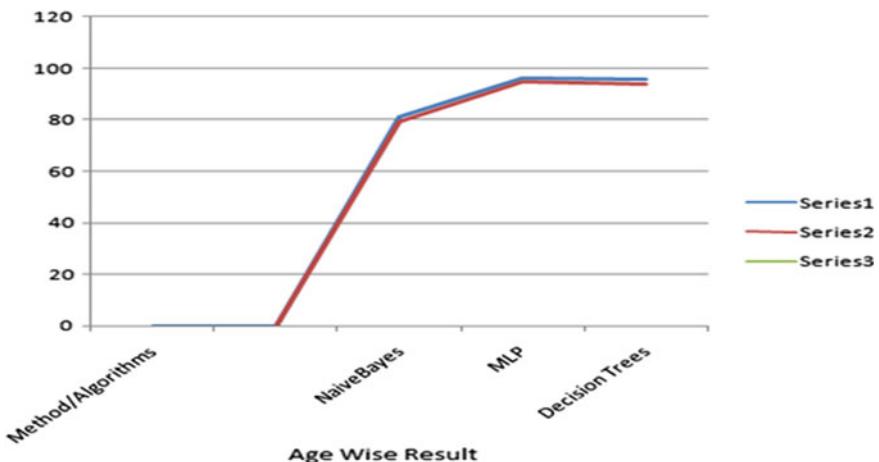


Fig. 5 Age-wise prediction graph

Table 4 Education-wise prediction of the dataset

Method/Algorithms	Education		
	Accuracy (%)	Precision (%)	Prediction (ms)
Naïve Bayes	80.75	88.2	70.8
MLP	97.75	98.14	19.46
Decision tree	96.25	95.45	52.54

Table 5 Income-wise prediction of the dataset

Method/Algorithms	Income		
	Accuracy (%)	Precision (%)	Prediction (ms)
Naïve Bayes	84.75	85.2	69.8
MLP	97.2	96.4	21.4
Decision tree	96.25	94.45	56.54

The measurements of the MLP dataset were much higher than the measurements from other datasets. Almost always lower measurements were used to add a number of characteristics to datasets. As it stands, it is likely that Spark will benefit from further review of the results. This can only be supported by using more methods of collection and datasets.

6 Conclusion and Future Work

Television is a universal tool for education, entertainment, information and so on. The television viewing behaviour may vary depending on different demographic variables. There may be a difference in programme preference with males and females, youth and old etc.

Here we intended to find the best algorithm to predict TV viewing behaviour. In our analysis, it has been found that MLP is the best in predicting TV viewing behaviour for all features. Here we tested three deep learning algorithms, viz., naïve Bayes classification, multi-layer perceptron classification and decision tree classification.

Spark does not have much deep learning in-house at present. For execution, however, separate third-party APIs are available. It might be appropriate for us to adjust the implementation. The Python API in Spark is available as various deep learning resources for language.

Since Apache generated Spark, other Apache services exist that can improve our way of implementing and handling cluster of Spark. Apache Ambari, for example, would allow us to start and maintain a web user interface with our cluster. That comes with a selection system of own metrics. This is what we are talking about. The request comes with a little more overhead, but it can be compensated in the cluster of appropriate computers.

In the analysis, MLP is found to be the best for predicting the difference in television viewing behaviour gender-wise, age-wise, education-wise and income-wise. This framework would provide data on the scale of the package to potential programme producers and advertisers. They can craft the media text based on these results. This will give better programme content, better identification of target audience and better reach to the programmes.

References

1. Brown, W.J.: The use of entertainment television programs for promoting prosocial messages. *Howard J. Commun.* **3**(3–4), 253–266 (1992). <https://doi.org/10.1080/10646179209359754>
2. Narasimhamurthy, N.: Television as a dominant source of infotainment among youths in Bangalore city. *IOSR J. Res. Method Educ.* **4**(5), 21–28 (2014)
3. BARC (n.d.). <https://www.barcindia.co.in/index.aspx>. Accessed 15 Jan 2021
4. Çiğşar, B., Ünal, D.: Comparison of data mining classification algorithms determining the default risk. *Sci. Program.* 1–8 (2019). <https://doi.org/10.1155/2019/8706505>
5. Inmon, W.H.: Building the Data Warehouse. Wiley, Indianapolis (2005)
6. Smyth, B., Cotter, P.: A personalized television listings service. *Commun. ACM* **43**(8), 107–111 (2000). <https://doi.org/10.1145/345124.345161>
7. Jain, P., Jakate, P., Dhotre: A novel approach to analysis of TV shows using social media, machine learning and big data. *IJTEL*, 4 (6) (2015)
8. Spangler, W., Gal-Or, M., May, J.: Using data mining to profile TV viewers. *Commun. ACM*. **46**, 66–72 (2003). <https://doi.org/10.1145/953460.953461>
9. Kaltenbrunner, A., Gómez, V., López, V.: Description and prediction of slashdot activity. In: Proceedings—2007 Latin American Web Conference, LA-WEB 2007 (2007), pp. 57–66. <https://doi.org/10.1109/LA-WEB.2007.59>

10. Szabó, G., Huberman, B.: Predicting the popularity of online content. *Commun. ACM.* **53** (2008). <https://doi.org/10.2139/ssrn.1295610>
11. Zhu, C., Cheng, G., Wang, K.: Big data analytics for program popularity prediction in broadcast TV industries. *IEEE Access* **5**, 24593–24601 (2017). <https://doi.org/10.1109/access.2017.2767104>
12. Kadam, T., Saraf, G., Dewadkar, V., Chate, P.: TV show popularity prediction using sentiment analysis in social network. *Int. Res. J. Eng. Technol.* **4**, 1087–1089 (2017)
13. Krishnamoorthy, N.: TV shows popularity and performance prediction using CNN algorithm. *J. Adv. Res. Dyn. Control Syst.* **12**, 1541–1550 (2020). <https://doi.org/10.5373/jardes/v12sp7/20202257>
14. Lee, J.M., Moon, S., Salamatian, K.: Modeling and predicting the popularity of online contents with cox proportional hazard regression model. *Neurocomputing* **76**, 134–145 (2012). <https://doi.org/10.1016/j.neucom.2011.04.040>
15. Kun, W., Wang, Y., Hu, X., Sun, Y., Deng, D.-J., Vinel, A., Zhang, Y.: Wireless big data computing in smart grid. *IEEE Wirel. Commun.* **24**, 58–64 (2017). <https://doi.org/10.1109/mwc.2017.1600256wc>
16. Anand, D., Satyavani, A.V. Raveena, B., Poojitha, M.: Analysis and prediction of television show popularity rating using incremental K-means algorithm. *Int. J. Mech. Eng. Technol.* **9**, 482–489 (2018)
17. Karthika, C., Vijayalakshmi, P.P.: More than news: viewing habits of television news satires. *Int. J. Adv. Sci. Technol.* **29**(03), 6963–6972 (2020)
18. Karthika, C., Vijayalakshmi, P.P.: Television news and women: impact of television news of flood in Kerala on women (6), 739–744 (2019). <https://doi.org/10.35940/ijeat.F1143.0886S19>

Inverse Kinematics of Robot Manipulator Integrated with Image Processing Algorithms



**Rajesh Kannan Megalingam, Santosh Tantravahi,
Hemanth Sai Surya Kumar Tammana, Nagasai Thokala,
Hari Sudarshan Rahul Puram, Naveen Samudrala,
and Chennareddy Pavanth Kumar Reddy**

Abstract Computer vision has many applications in various fields, such as remote sensing, face detection, and fingerprint detection. In this paper, various algorithms for motion detection, hazmat detection, and QR code detection are presented. These algorithms are implemented in OpenCV which are ROS-integrated. A camera is set up on a robot with a 5 degrees of freedom (DOF) manipulator. This robot is sent to remote locations and can gather information about the environment. This information can be the type of hazmat detected, the coordinates of the hazmat signs, etc. An inverse kinematic model of the arm is presented. A simulation of the arm configurations based on the coordinates obtained from the camera has been done using MATLAB.

Keywords Hazmat sign detection · Motion detection · QR detection · SAR · Robot operating system (ROS) · Open CV · Inverse kinematics · MATLAB

1 Introduction

A search and rescue robot is used to get information from a remote area where humans cannot reach and the information collected is sent to the operator station which is integrated with ROS. The information from the camera is checked with the algorithms and the corresponding algorithm is executed. In this paper, a proposal relating to a combined method for hazmat detection (is used to detect harmful substances such as flammable liquids, poisonous gases, etc.), QR detection, and motion detection (can be used to know whether a person is alive in case of a disaster such as earthquake) combined with ROS has been made. Cv_bridge is used to convert information from

R. K. Megalingam (✉) · S. Tantravahi · H. S. S. K. Tammana · N. Thokala · H. S. R. Puram
Department of Electronics and Communication Engineering, Amrita Vishwa Vidyapeetham,
Amritapuri, Kerala, India

N. Samudrala · C. P. K. Reddy
Department of Electrical Engineering, University of Dayton, Dayton, OH, USA
e-mail: lnup01@udayton.edu

the camera driver to the ROS messages. When a motion is detected in the frame, the corresponding area will be highlighted. Motion detection is a real-time application used in survival cameras and motion cameras. When the QR code is detected, the information which is encrypted in the QR will be displayed on the QR code. In the case of a hazmat sign, the corresponding sign name will be stored in a text file when a hazmat sign is detected. A camera is used to obtain the coordinates when there is a detection of hazmat/QR. The obtained coordinates are used for simulating the inverse kinematic model of a 5 degrees of freedom (DOF) robot arm in MATLAB. Our paper is divided into five sections. The first section introduces what the paper is about. The second section includes the purpose of the passage, which describes the goals which are to be achieved. The third section discusses the background studies and related works. System architecture and implementation and a set of specific computer tools are mentioned, and analysis on three algorithms is explained in Section 4. Finally, the last section contains the conclusions, future works, and acknowledgment.

2 Motivation

Autonomously working robotic arms have significant applications in many fields, such as the manufacturing industry, etc. It has considerable applications in the field of disaster search and rescue operations as well. An autonomously working robotic arm can be used to reach places where humans cannot reach physically. Gathering information from these locations is a complex task that can be easily achieved with the help of these robotic arms. This information can be useful during the time of disaster rescue operations. A theoretical modeling of a robotic arm and its simulation was performed and presented in this paper. The modeling presented here is a simple one without any dynamic constraints from the environment.

3 Related Works

The author in [1] used face detection, skin detection, color detection, shape detection, and target detection algorithms and implemented them in MATLAB to detect various types of objects for many surveillance applications. The author in [2] proposed a complete closed-form solution to the inverse kinematics problem for a 4-DOF manipulator. This paper [3] presents an inverse kinematics analysis of a humanoid robot arm under geometric constraints. The author in [4] transformed the QR code into many runs of data in alternate pixels of black and white. The experiments like image binarization, image seeking, and localization adjustment are accomplished in sequence. Here the QR code used a decoding system. The work in [5] proposed a localization involving a convolutional neural network that could detect barcodes. Image processing algorithms were implemented. In paper [6], a novel approach is used to detect the QR code in arbitrarily acquired images that are proposed, and the

results show an effective way of detecting the QR code. A monitoring system in [7] used a motion detection technology and is implemented using C# and MATLAB programming language and the author used four different motion detectors. A DMD module is used in [8] and it is added to a video coder to determine whether the current block is in motion or with zero motion and a dynamic calculator of local variance of difference frame as a threshold is used to detect the motion activity. The author in [9] uses two different thresholding techniques and the blacks motion estimation technique which is used for comparison based on the measure of overall derived tracking angle. The teleoperation of a robot on ROS connected through SSH is given in [10]. A robot model simulation in different scenarios with input from a steering wheel is presented in [11]. The authors of [12] discussed the simulation of a 6 DOF arm ROS gazebo. The authors of [13] designed a robotic coconut tree climber called Amaran and presented results of various experiments and tests conducted on the robot to test its viability and other performance metrics.

4 System Architecture

The system architecture comprises three major blocks: camera, detection block, and Rqt GUI. The detection block consists of motion detection, hazmat sign detection, and QR code detector and decoder. The images are viewed in the operator console with the help of Rqt GUI. Rqt is an inbuilt library in ROS which is used to monitor various sensor readings and camera feed in the form of ROS topics. The camera gives the coordinates in the form of x, y, z after detection. The system architecture layout is shown in Fig. 1. The img_topic mentioned in the system architecture is a ROS entity that carries the camera data and provides this data to various detection algorithms for further process.

Hazmat detection deals with the identification of various hazmat signs that are used to identify substances with certain regulations to be met. It helps to identify the type of hazardous material in a package. The QR detection mainly deals with

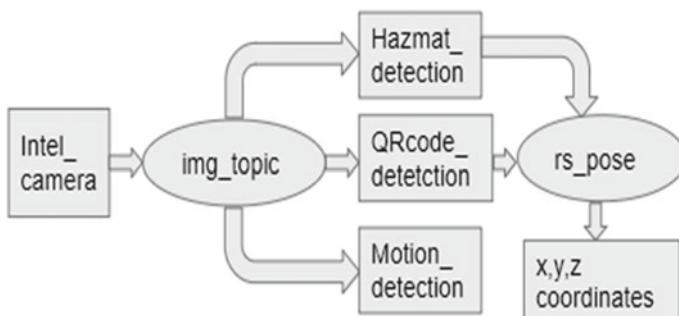


Fig. 1 System architecture

the translation of QR (quick response) code into information that can be easily understood by humans. Similarly, motion detection is used to identify the change in position of an object with respect to the surrounding environment.

5 Inverse Kinematic Model of Arm

Figure 2 is the robotic arm for which the inverse kinematic model has been developed. The robot's work space is assumed to be a cylinder. The coordinates obtained are in the form of (x, y, z) . These coordinates are obtained after transforming the coordinates that are obtained in the camera frame (obtained from the camera) to the base frame of the robotic arm.

$$\Theta = \text{atan}2(y, x) \quad (1)$$

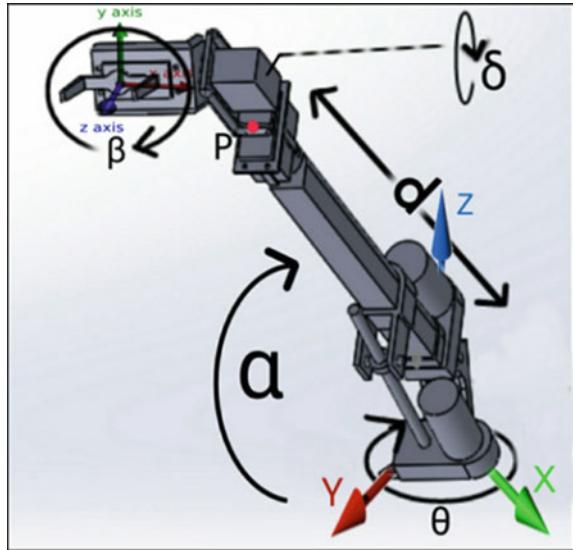
$$\alpha = \text{atan}2(z, y) \quad (2)$$

$$d = \sqrt{(x^2 + y^2 + z^2)} \quad (3)$$

The orientation matrix obtained from Θ , α and d is given

$$R_I = [R_{I11} R_{I12} R_{I13} R_{I21} R_{I22} R_{I23} R_{I31} R_{I32} R_{I33}] \quad (4)$$

Fig. 2 Robotic arm



R_I is the orientation matrix of a coordinate frame present at point P.

The end effector's frame from the intermediate frame is the rotation about the x -axis by an angle and rotation about z -axis by an angle β .

$$R_{EF} = \text{Rot}(x, \delta) * \text{Rot}(z, \beta) \quad (5)$$

$\text{Rot}(z, \beta)$ is a rotation matrix which represents the rotation along z -axis by an angle β . $\text{Rot}(x, \delta)$ is a rotation matrix which represents the rotation along x -axis by an angle δ .

Let R be the desired orientation of the end effector.

$$R = R_I * R_{EF} \quad (6)$$

$$R_{EF} = R_I^{-1} * R \quad (7)$$

For a rotation matrix,

$$R^{-1} = R^T \quad (8)$$

This implies

$$R_{EF} = R_I^T * R = A \quad (9)$$

$$A = [A11\ A12\ A13\ A21\ A22\ A23\ A31\ A32\ A33] \quad (10)$$

Comparing A and R_{EF} we get

$$\beta = \text{acos}(A11) \quad (11)$$

$$\text{If } \sin(\beta) > 0 \quad (12)$$

$$\delta = \text{atan2}(A13, -A12) \quad (13)$$

$$\text{If } \sin(\beta) < 0 \quad (14)$$

$$\delta = \text{atan2}(-A13, A12) \quad (15)$$

6 Implementation

The camera feed is visualized by running the `usb_cam` driver. This is a ROS library for visualization. OpenCV is used to perform operations on images and ROS as a platform. Different algorithms access the camera feed through ROS topics. The video feed from the camera is read from ROS topics and copied to two frames, and the difference between the two frames is computed. Then the result is converted to a grayscale image and the threshold is applied. Then the difference frame is dilated. Contours on dilated frames are used to sense the motion in the frame. SIFT and “`find_2d_object`” is used for real-time image detection. The computer is trained to detect the necessary images. `Find_2d_object` fetches camera feed from ROS topics. `Find_2d_object` plays a key role in searching the images in disaster-struck areas. If detected, the corresponding hazardous names will be displayed on the screen and saved in a text file.

`Zbar` is an open-source library used for real-time scanning of video streams to decode the information hidden in the QR codes. `Zbar` keeps looking for a QR, and whenever it finds one it will decode and the regarding text information is displayed on the screen, then save it in a text file for future use. The `Z` library consists of three components: finder patterns, alignment patterns, and timing patterns. Finder pattern is used to detect three squares present in a QR code. The detection of the three squares confirms the presence of *qr* codes and its orientation in the image/video feed. Alignment patterns are used for QR detection when there is a distortion in the QR. Timing patterns are dotted lines, which help in the identification of the location of the data grid. The black/white-colored blocks form bits.

6.1 *Rqt_gui*

`rqt_image_view` is a `rqt` version of `image_view`. `image_view` used to view images on ROS topics. `rqt_gui` allows us to open multiple `rqt_image_view` windows and dock into a single window. Through `rqt_gui` the final output can be viewed.

6.2 *Intel i5 NUC*

The processor used is Intel Core i5 NUC. This is one of the best processors that support computer vision techniques. This is seventh generation of Intel Core i5-7260U with a processing speed of 2.2 GHz along with frequency of 3.4 GHz. It has four USB ports of which two ports support version 3.0. This processor has an Ethernet port and an HDMI port. It uses memory technology of DDR4-2133 and the power input is 12–19 V.

6.3 Arduino UNO

The microcontroller used in the board is ATmega328P. It can operate at a voltage of 5 V and it has a flash memory of 32 KB. It has 6 analog pins and 14 digital pins of which six can be used as PWM also. It works on a frequency of 16 MHz. It also has TX and RX pins that can transmit and receive data with the help of Bluetooth.

6.4 Logitech Camera and Intel RealSense

A Logitech camera is an optical instrument that can capture images and it can also record moving images. The camera has 2.4 GHz Intel Core TM2 Duo and it has 2 GB RAM with 200 MB hard drive space and with USB 2.0 port. It also has 1 Mbps upload speed with a high screen resolution of 1280×720 .

The Intel RealSense Depth Camera D435 extracts 3D information from an environment. This feed combines 3D depth to robotics navigation, object recognition, and other applications. It is a USB-powered depth camera and features a pair of depth sensors, an RGB sensor, and an infrared projector.

6.5 Robot

The robot is a wheeled robot with a setup of CPU, Arduino, motor driver, DC motors, camera, and router. SSH connection is used to access the robot which is in a remote location. The user can view the environment on the screen so that required tasks can be executed. Joystick is used to give the input commands to the robot so that it can move to the necessary locations. The algorithms are launched using Linux and ROS. The algorithms of image processing used are motion detection, QR code detection, and hazmat detection. The algorithms are executed and the results appear on the console screen. The detected output is saved in a text file.

7 Experiments and Results

The algorithms were successfully launched from the user console. Joystick is used to control the robot. All results can be viewed in the user console screen and detection can be observed.

Figures 3 and 4 show that inhalation hazard and radioactive hazmat signs are detected with a boundary of red line, and the name of the hazmat sign is displayed. Figure 5 shows that two hazmat signs are detected at a time and the name of the signs is displayed. Figure 6 shows that the motion of the fan is detected with a red line

Fig. 3 Detection of inhalation hazard



Fig. 4 Detection of radioactive hazard



where there is a movement in the frame. Figure 7 shows that a QR code is detected and the information in QR is displayed. Figure 8 shows the launch of algorithms at a time. Figure 9 shows the MATLAB simulation for a given set of inputs as shown in Table 1.

The end effector can have multiple orientations for the given set of inputs.



Fig. 5 Two hazmat detections at a time

Fig. 6 Motion detection of fan



8 Conclusions

Detection of hazmat signs, QR code, and motion was tested under using different examples and has been successfully accomplished. All the algorithms were launched successfully and are working efficiently. All the detections were exactly traced from the environment. Simulation of the inverse kinematic model of the arm has been done using MATLAB. Different and more advanced image processing algorithms which can aid in rescue processes can be implanted in the system. Thermal cameras can also be used which sense the thermal image and detect the presence of human beings. Real-time autonomous control of the robotic arm can be implemented.



Fig. 7 QR detection

```
hal
/home/hutlab/Desktop/test/src/ohm_rrl_perception/perception_launch/launch/perception.launch
[2]+ Stopped                  roslaunch ohm_rrl_perception_launch perception.lau
nch
hutlab@07CPU0118L:~/Desktop/test$ roslaunch ohm_rrl_perception_launch perception
.launch
... logging to /home/hutlab/.ros/log/341aef98-dc89-11e9-85de-00e04c63518c/roslau
nch-07CPU0118L-31349.log
Checking log directory for disk usage. This may take awhile.
Press Ctrl-C to interrupt
Done checking log file disk usage. Usage is <1GB.

started roslaunch server http://07CPU0118L:42645/
SUMMARY
=====
PARAMETERS
* /hazmat_detection/gui: False
* /hazmat_detection/objects_path: /home/hutlab/Des...
* /hazmat_detection/settings_path: /home/hutlab/Des...
* /hazmat_detection/subscribe_depth: False
* /hazmat_viz/input_topic: img_perception
* /hazmat_viz/viz_topic: img_hazmats
* /motiondetection/debug: False
* /motiondetection/image_topic: img_perception
```

Fig. 8 Launch of image processing algorithms

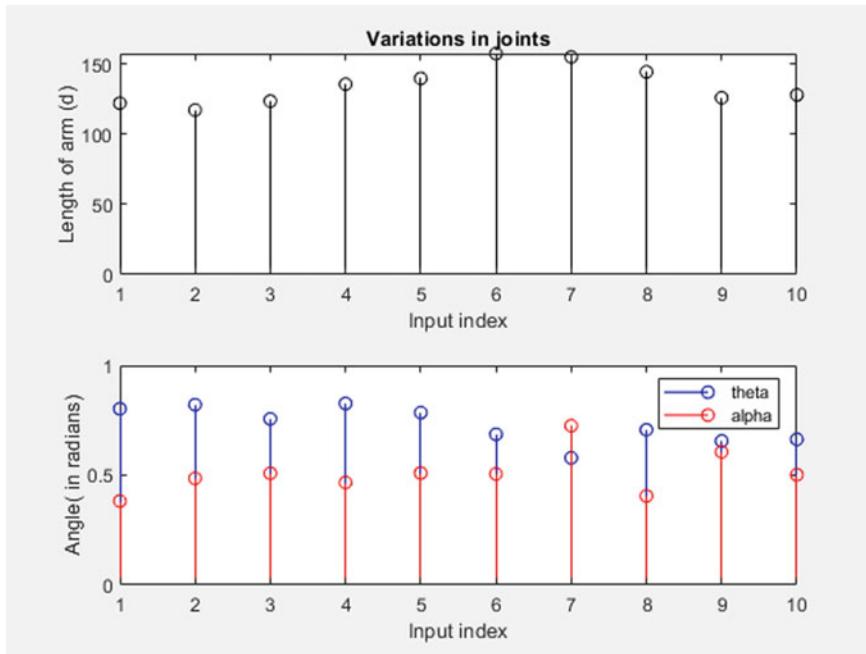


Fig. 9 MATLAB simulation of the inverse kinematic model

Table 1 Input table

Input	(X, Y, Z) in cm
1	(85, 88, 40)
2	(80, 86, 50)
3	(90, 85, 52)
4	(92, 100, 55)
5	(99, 99, 60)
6	(122, 100, 60)
7	(130, 85, 80)
8	(110, 94, 45)
9	(100, 77, 58)
10	(101, 79, 48)

Acknowledgments The authors are indebted to HuT labs and the Department of Electronics and Communication Engineering at Amrita Vishwa Vidyapeetham, Amritapuri for their continuous guidance and support.

References

1. Raghuandan, A., Mohana, Raghav, P., Aradhy, H.V.R.: Object detection algorithms for video surveillance applications. In: 2018 International Conference on Communication and Signal Processing (ICCSP), pp. 0563–0568 (2018). <https://doi.org/10.1109/ICCSP.2018.8524461>
2. Si, Y., Jia, Q., Chen, G., Sun, H.: A complete solution to the inverse kinematics problem for 4-DOF manipulator robot. In: 2013 IEEE 8th Conference on Industrial Electronics and Applications (ICIEA), pp. 1880–1884 (2013). <https://doi.org/10.1109/ICIEA.2013.6566674>
3. Song, D.H., Jung, S.: Geometrical analysis of inverse kinematics solutions and fuzzy control of humanoid robot arm under kinematics constraints. In: 2007 International Conference on Mechatronics and Automation, pp. 1178–1183 (2007). <https://doi.org/10.1109/ICMA.2007.4303715>
4. A-Lin, H., Yuan, F., Ying, G.: QR code image detection using run-length coding. In: Proceedings of 2011 International Conference on Computer Science and Network Technology, pp. 2130–2134 (2011). <https://doi.org/10.1109/ICCSNT.2011.6182398>
5. Chou, T., Ho, C., Kuo, Y.: QR code detection using convolutional neural networks. In: 2015 International Conference on Advanced Robotics and Intelligent Systems (ARIS), pp. 1–5 (2015). <https://doi.org/10.1109/ARIS.2015.7158354>
6. Zhang, X., Luo, H., Peng, J., Fan, J., Chen, L.: Fast QR code detection. In: 2017 International Conference on the Frontiers and Advances in Data Science (FADS), pp. 151–154 (2017). <https://doi.org/10.1109/FADS.2017.8253216>
7. Yong, C.Y., Sudirman, R., Chew, K.M.: Motion detection and analysis with four different detectors. In: 2011 Third International Conference on Computational Intelligence, Modelling & Simulation, pp. 46–50 (2011). <https://doi.org/10.1109/CIMSim.2011.18>
8. Metkar Shilpa, P., Talbar Sanjay, N.: Dynamic motion detection technique for fast and efficient video coding. TENCON 2008—2008 IEEE Region 10 Conference, pp. 1–5 (2008). <https://doi.org/10.1109/TENCON.2008.4766715>
9. Zhang, S., Stentiford, F.: Motion detection using a model of visual attention. In: 2007 IEEE International Conference on Image Processing, pp. III - 513–III - 516. <https://doi.org/10.1109/ICIP.2007.4379359>
10. Megalingam, R.K., Tantravahi, S., Tammana, H.S.S.K., Thokala, N., Puram, H.S.R., Samudrala, N.: Robot operating system integrated robot control through secure shell (SSH). In: 2019 3rd International Conference on Recent Developments in Control, Automation & Power Engineering (RDCAPE). <https://doi.org/10.1109/rdcape47089.2019.8979113>
11. Megalingam, R.K., Nagalla, D., Pasumarthi, R.K., Gontu, V., Allada, P.K.: ROS based, simulation and control of a wheeled robot using Gamer's steering wheel. In: 2018 4th International Conference on Computing Communication and Automation (ICCCA). New Delhi, India. <https://doi.org/10.1109/ccaa.2018.8777569>
12. Megalingam, R.K., Katta, N., Geesala, R., Yadav, P.K., Chanda, R.: Keyboard-based control and simulation of 6-DOF robotic arm using ROS. In: 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India. <https://doi.org/10.1109/ccaa.2018.8777568>
13. Megalingam, R.K., et al.: Amaran: an unmanned robotic coconut tree climber and harvester. IEEE/ASME Trans. Mechatron. **26**(1), 288–299 (2021). <https://doi.org/10.1109/TMECH.2020.3014293>

A Comprehensive Study on Workloads in Cloud Computing



K. Mallikharjuna Rao and Aravapalli Rama Satish

Abstract The outstanding burden is a measure of work performed or it tends to be proficient to perform inside a particular timeframe. In cloud computing, the term outstanding burden alludes to the sort and qualities of the application that are needed to be facilitated on the cloud. As we realize that every application has its prerequisites, trademark, and behavior. So that what applications behave on the cloud is to a great extent subject to the workload. Essentially workloads are either genuine or manufactured remaining burden. This paper clarifies the various kinds of genuine remaining tasks at hand and their uses which are existing openly and what are the measurable and benchmarking tools used to examine and age of manufactured outstanding burden from the genuine outstanding tasks at hand.

Keywords Cloud computing · Reliability · IBM SPSS statistics · R-programming · Real workload · Synthetic workload · Application benchmarks

1 Introduction

Distributed (Cloud) computing is an innovation that is getting more famous step by step. This is a direct result of its versatile nature: clients can utilize the assets on interest and pay just for the assets they need. Those assets are basically in Virtual Machine (VM) structure. Various organizations can utilize the cloud for various assets like used to run group occupations, stockpiling, or reinforcement reason and to have web applications.

The three various types of distributed computing framework:

K. Mallikharjuna Rao (✉) · A. Rama Satish

Assistant Professor Sr. Grade 1, School of Computer Science and Engineering, VIT-AP

University, Amaravati, India

e-mail: mallikharjuna.rao@vitap.ac.in

A. Rama Satish

e-mail: rama.satish@vitap.ac.in

- **Software-as-a-Service (SaaS):** It is utilized to have seller applications. Example: Google apps (Google docs, Google destinations, Google calendar), Microsoft office365, Cisco WebEx. Regularly these product frameworks will confront an issue the product is solid or not because of the progressions and assumption for wanted high dependable results [1, 2].
- **Infrastructure-as-a-Service (IaaS):** It is utilized to give the assets which are identified with IT and organization like overseeing, getting to, and checking distant server farms. Model: Amazon Web Services (AWS), Microsoft Azure, Google Compute Engine.
- **Platform-as-a-Service:** It gives programming conditions to the clients to build and build up the applications on the cloud framework. Model: Google APP Engine, Heroku, Force.com, Microsoft Windows Azure.

In distributed computing, various applications are existing where every application can have various attributes and conduct dependent on its current circumstance. Notwithstanding the choice to try the stage, controlled outstanding burdens are utilized to run the analysis. The word remaining task at hand characterizes the number of clients demands with appearance time-stamps. Outstanding tasks at hand are two sorts it very well may be either manufactured or genuine remaining burden. The engineered remaining burdens are created with explicit projects. In any case, the genuine outstanding tasks at hand are put away in traces (.csv records) which are accessible on the web. These outstanding tasks at hand are utilized to test the presentation of uses. Both the genuine and manufactured outstanding burdens characterize momentarily in the next segments.

2 Background: Performance Evaluations on Cloud

The word remaining burden characterizes the number of clients demands with appearance time-stamps. Outstanding burdens are utilized to test the exhibition of utilizations in the cloud. Outstanding tasks at hand are either genuine or manufactured remaining burdens. Genuine outstanding tasks at hand are accessible freely on the web which is put away in traces. These traces are in the forms as .CSV documents. Engineered remaining burdens are created with a particular program.

2.1 Experimental Platforms

Experimentations can occur either from genuine cloud suppliers (providers) or private cloud suppliers. While doing tests in private cloud suppliers the primary disservice is to set the entire foundation without fail. In any case, in genuine cloud suppliers, the foundation is now set yet, we need to arrange the framework to send the applications benchmark. As indicated by cloud suppliers, they will be charged for every execution.

To keep away from the experimentation cost there is the best arrangement called custom testbeds. There will be a charge as per the framework setup. The least demanding advance is to introduce the VMware programming that can oversee both the OS level, Server level, and Application level. Virtualization worker level is required for custom testbeds, for example, VMM (virtual machine monitor) and Hypervisor which runs each or additionally working frameworks in turn. A portion of the famous hypervisors is ZEN, Kernel Virtual Machines (VMM), VMware ESXi. There are various sorts of open-source custom testbeds that are accessible to convey various applications: Eucalyptus, open stack. Also, some sort of business custom testbeds like Vcloud Director created by VMware. Open stack suppliers open-source uphold for some, ventures like HP, Intel, AMD, however, it is as yet under development.

In the genuine framework, a test system is expected to test the working of cloud stages including assignment and deallocation of assets. Making another test system identified with cloud stages is a period taking cycle. In this way, to keep away from that we need to choose the test system which as of now exists in the cloud and change it to our necessities. It might require hours to do the trial on the genuine foundation. Be that as it may, the occasion-based climate analysis is completed in minutes. Genuine frameworks are exceptionally configurable, so it permits the client to assemble the exhibition condition of an application and the outcomes rely upon the degree of usage of an application. A portion of the exploration situated cloud test systems is Cloud Sim, Ground Sim, and Green Cloud.

2.2 *Real Workloads*

In cloud-based frameworks applications, outstanding burdens are sorted into two different ways conditional and bunch remaining tasks at hand. In value-based outstanding tasks at hand, the web applications are used to serve the online HTTP solicitations of customers.

This is primarily used to play out the substance of HTML pages, online video transfers, or pictures. Every one of these substances of value-based outstanding burdens can be put away statically and progressively gave by workers. In cluster outstanding burdens, it used to deal with the long-running positions, asset-based positions. As per our knowledge, there are no genuine workloads that are accessible freely and publicly. The creators who need to perform tests utilizing genuine traces they should make their genuine traces dependent on the need for their investigation. In any case, a portion of the genuine remaining tasks at hand are freely accessible, for example, ClarkNet trace [3], World Cup98 [4], and Google traces [5]. ClarkNet trace keeps up the data of online HTTP demands that are gotten by the ClarkNet worker in 1995 from about fourteen days' timeframe. ClarkNet is loaded with web access gave by the Metro Baltimore Washington DC territory. In Fig. 1 ClarkNet trace shows some cyclic example it clarifies the remaining (workload) tasks at hand are more in day

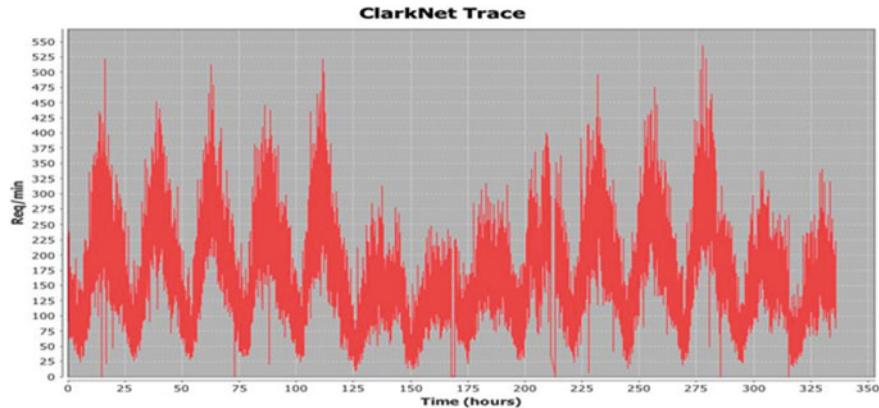


Fig. 1 The number of requests processed per minute in ClarkNet Trace

time contrasted with evening time and the outstanding burdens are high on ordinary days contrasted with workdays.

For instance, World Cup trace 98 contains online HTTP demands made by the World Cup in the year 1998 between the period April 30, 1998, to July 26, 1998. In genuine remaining burdens, different traces are openly accessible called Google trace version1 and Google trace version2 which are accessible in Google Data Center. Google trace version1 the main informational index was taken over from alludes 7 hours timeframe. The Google trace is an assortment of informational collections that are introduced in the Google server farm as a CSV record. In version1 each undertaking alludes to solitary work and the work may allude to different assignments. Each errand can execute for 5 min timeframe. The underneath Fig. 2 shows the Google trace version1.

3 Synthetic Workloads Generators

Engineered outstanding burdens are a counterfeit program that is created by breaking down the attributes and conduct of genuine remaining tasks at hand. Engineered Workloads use trace for an outstanding task at hand age and benchmarks for assessment, so the manufactured remaining burdens are more reasonable and adaptable. There are various sorts of engineered remaining tasks at hand instruments and benchmarks are accessible on the web. Table 1 shows the different kinds of engineered remaining burden Generators.

Google trace version2 keeps up some more data about positions, limitations, and machine qualities contrasted with trace 1. The version2 information is taken from 11 k machines over 30 days' timeframe. The beneath Fig. 3 shows the Google trace version2.

	A	B	C	D	E	F
1	Time	ParentID	TaskID	JobType	NrmlTaskCores	NrmlTaskMem
2	90000	757745334	1488529826	0	0	0.0311296
3	90000	975992247	1488529821	0	0	0
4	90000	1468458091	1488529832	1	0.021875	0.00235309
5	90000	1460281235	1488529840	0	0	0
6	90000	1164728954	1488529835	0	0.003125	0.0016384
7	90000	1288997448	1488529848	0	0.003125	0.0049152
8	90000	1488529845	1488529847	1	0.003125	0.000719232
9	90000	1263655469	1488529844	2	0	0
10	90000	1164728954	1488529851	0	0.003125	0.0016384
11	90000	981883546	1488529849	3	0.003125	0.0838861
12	90000	1263655469	1488529853	2	0	0
13	90000	757745334	1488529860	0	0	0
14	90000	1487094655	1488529866	0	0.003125	0.00382931
15	90000	1458411965	1488529868	0	0	0.000402513

Fig. 2 Google trace Version1 loads in excel file**Table 1** Represents different types of synthetic workload Generators in synthetic workload generation

S. No	Synthetic workload Generator	Description
1	Rain Workload Generator	Rain toolkit [10] is a statistical-based workload generation. To show various remaining tasks at hand it used observational and defined circulations
2	Faban	Faban [11] is a Markov-chain-based Workload generator and is generally utilized for worker execution and burden testing likewise alluded to as benchmarking
3	Apache JMeter	Apache JMeter [12] is utilized for both burden Testing and execution assessment. It plays out the assessment test in both static and dynamic
4	httperf	httperf [13] used to quantify the exhibition of web servers. It assists with creating a few HTTP remaining tasks at hand and measures the exhibition in various situations

Untitled2 [DataSet2] - IBM SPSS Statistics Data Editor																				
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help																				
Visible: 26 of 26 Variables																				
		starttime	endtime	jobid	taskid	ex	machined	cpuusage	memoryusa	assignedmem	memory	maxmemor	diskusage	localdiskop	cpuusage	disk	cpi	mai	sa aggr	ti
1	8000000000	1 E+10	3 E+6	0	4155527081	.0014	.0679	.0757	.0679	.0000	.0002	.0112	.00	.24290	.0071	0	1	5.70e3x6gR	3	9 13 09 00
2	8000000000	1 E+10	3 E+6	1	329150663	.0012	.0679	.0756	.0679	.0000	.0002	.0104	.00	.23340	.0073	0	1	0.70e3x6gR	3	9 13 09 00
3	1000000000	1 E+10	3 E+6	0	4155527081	.0013	.0679	.0757	.0679	.0000	.0002	.0107	.00	.24630	.0077	0	1	1.70e3x6gR	3	9 13 09 00
4	1000000000	1 E+10	3 E+6	1	329150663	.0012	.0679	.0756	.0679	.0000	.0002	.0256	.00	.21450	.0063	0	1	5.1X1Q440	3	9 13 02 00
5	4000000000	1 E+10	3 E+6	0	4155527081	.0014	.0679	.0757	.0679	.0000	.0002	.0127	.00	.23610	.0070	0	1	5.1X1Q440	3	9 13 02 00
6	4000000000	1 E+10	3 E+6	1	329150663	.0012	.0679	.0756	.0679	.0000	.0002	.0181	.00	.21520	.0066	0	1	5.1X1Q440	3	9 13 02 00
7	7000000000	1 E+10	3 E+6	0	4155527081	.0013	.0679	.0757	.0679	.0000	.0002	.0182	.00	.24910	.0076	0	1	2.1WsgqT	0	0 03 00 00
8	7000000000	1 E+10	3 E+6	1	329150663	.0012	.0679	.0756	.0679	.0000	.0002	.0173	.00	.20600	.0061	0	1	0.1WsgqT	0	0 03 00 00
9	0000000000	1 E+10	3 E+6	0	4155527081	.0013	.0679	.0757	.0679	.0000	.0002	.0132	.00	.27290	.0083	0	1	1.1WsgqT	0	0 03 00 00

Fig. 3 Google trace Version2 loads in IBM SPSS Statistics file

3.1 Application Benchmark's

Application Benchmarks assess worker adaptability, dependability, and execution. Frequently unwavering quality issue happens with programming frameworks. A portion of the arrangement methods for dependability issues are proposed in [6–9] Some of the normally utilized benchmarks for the age of engineered outstanding tasks at hand are RUBiS, Cloud Stone, TPC-W. Be that as it may, RUBiS and TPC-W the two benchmarks are obsolete, yet these benchmarks are as yet utilizing in the examination region.

3.2 Synthetic Workload Generation with the Help of Real Workloads

In engineered remaining task at hand age initial step is to describe the genuine outstanding burdens utilizing some sort of dissecting apparatuses called R programming or IBM SPSS insights 22. After portrayal, the subsequent advance is to produce the manufactured outstanding tasks at hand dependent on the qualities and conduct of the genuine remaining burden.

Central processor utilization, memory use, RAM solicitation of work.

To begin with, we dissected that remaining burden utilizing IBM SPSS statistics 22. In this work, start time and end time is known as an errand. The Fig. 4 shows the beginning time of work at 10800000000 and the consummation time is at 13200000000. It speaks to one errand. The remaining computations are given in Fig. 4.

After investigating the information in IBM SPSS measurements. By utilizing Linear relapse in SPSS formulae is produced in each group for CPU usage and memory usage utilizing the accompanying articulations.

$$CPU\text{ usage} = c + m.jobid + n.memory\text{ usage}. \quad (1)$$

where c is constant, m is RAM request, n is CPU request

$$\text{Mem usage} = c + m.jobid + n.cpu\text{ usage} \quad (2)$$

Considering this, the engineered outstanding task at hand is produced for CPU and Memory use. Figure 5 speaks to the manufactured outstanding burden age for CPU usage mean. Here cpuusage1 speaks to the manufactured remaining burden and CPU usage mean speaks to the genuine outstanding task at hand.

The computations of CPU usage mean qualities of various cases are given in Fig. 6.

The computations of synthetic workload CPU usage mean qualities of various cases are given in Fig. 7.

	starttime	endtime	jobid	Rjobid
1	10800000000	11100000000	3418309	1.000
2	11100000000	11400000000	3418309	1.000
3	11400000000	11700000000	3418309	1.000
4	11700000000	12000000000	3418309	1.000
5	12000000000	12300000000	3418309	1.000
6	12300000000	12600000000	3418309	1.000
7	12600000000	12900000000	3418309	1.000
8	12900000000	13200000000	3418309	1.000
9	10800000000	11100000000	3418309	1.000
10	11100000000	11400000000	3418309	1.000
11	11400000000	11700000000	3418309	1.000
12	11700000000	12000000000	3418309	1.000
13	12000000000	12300000000	3418309	1.000
14	12300000000	12600000000	3418309	1.000
15	12600000000	12900000000	3418309	1.000
16	12900000000	13200000000	3418309	1.000
17	10800000000	11100000000	3418314	2.000
18	11100000000	11400000000	3418314	2.000
19	11400000000	11700000000	3418314	2.000
20	11700000000	12000000000	3418314	2.000

Fig. 4 Analyzing Google trace version2 in IBM SPSS Statistics

caseno	starttime	endtime	cpuusage1	cpuusagemean	err
5178.00	11100000000	11400000000	.11629222	.10170000	-.01
5179.00	11400000000	11700000000	.11629222	.12870000	.01
5180.00	11700000000	12000000000	.11629222	.12550000	.01
5181.00	12000000000	12300000000	.11629222	.11730000	.00
5182.00	12300000000	12600000000	.11653012	.13040000	.01
5183.00	12600000000	12900000000	.11656062	.10510000	-.01
5184.00	12900000000	13200000000	.11649962	.10280000	-.01

Fig. 5 Generation of duplicate CPU usage mean form Real CPU usage mean

Figure 8 shows the error rate difference between the real and synthetic workload of the CPU usage mean.

As a similar cycle, we accomplished for the manufactured remaining task at hand age of memory usage mean.

Likewise, this paper provides the various open-source application benchmarks for synthetic workload generation, and the tools useful for Real workloads characterization are given in Tables 2 and 3.

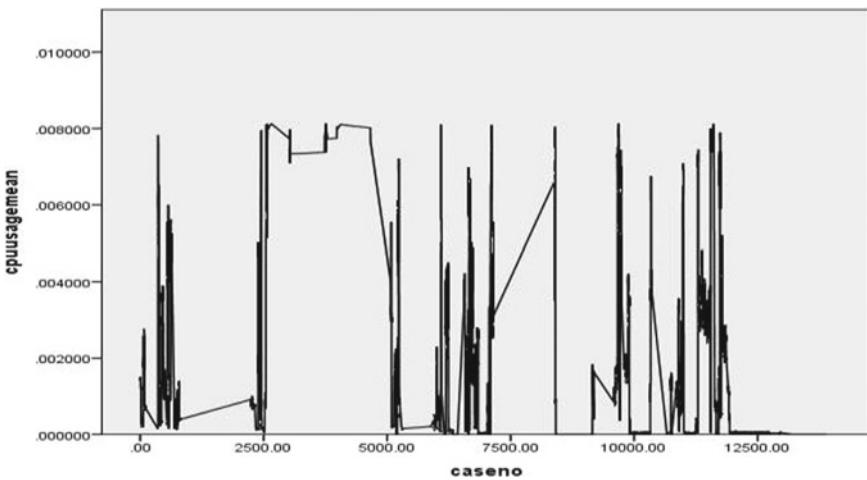


Fig. 6 The graph represents the real CPU usage mean

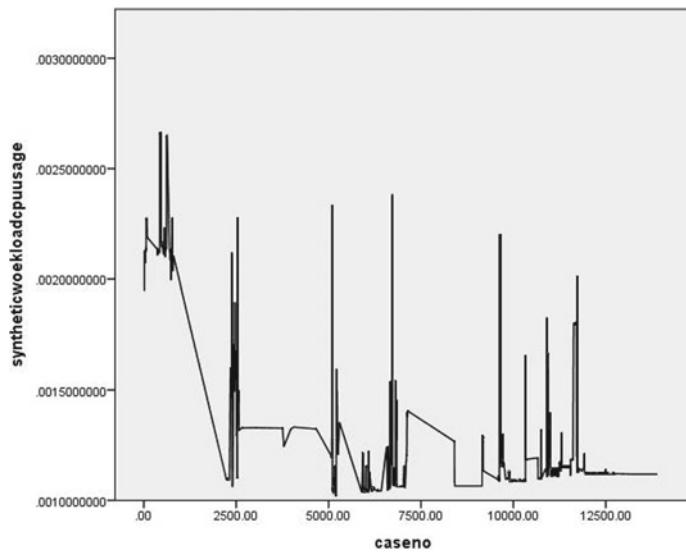


Fig. 7 The graph represents the Synthetic workload CPU usage mean

4 Conclusion

This paper clarifies that in distributed computing not many numbers genuine remaining tasks at hand accessible on open mists, so dependent on dissecting the attributes and conduct of genuine outstanding burden by utilizing distinctive examining instruments like R programming or IBM SPSS statistics tools which are used to

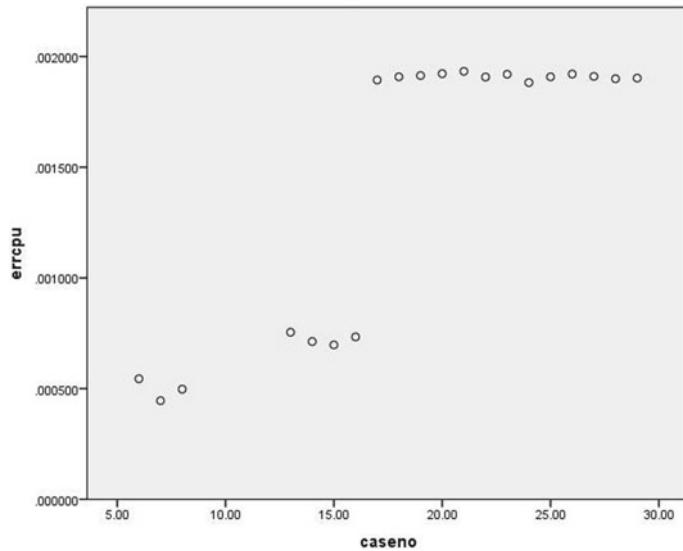


Fig. 8 The graph represents the error rate between real CPU usage mean and Synthetic CPU usage mean

Table 2 Represents different types of Application Benchmark's in synthetic workload generation

S. No	Application Benchmark's	Description
1	RUBiS	RUBiS [14] is a model for a bartering Website planned after ebay.com. It is used to check the server performance and adaptability
2	Cloud Stone	CloudStone [15] is a Multilanguage, Multi-stage apparatus for cloud computing created by the RAD lab bunch at the college of Berkeley. It includes the age of adaptable and sensible outstanding burden generator (Faban) to create a remaining task at hand against the genuine outstanding task at hand web 2.0
3	TPC-W	TPC-W [10] is used to characterize the information base and exchange measure. It is a web-based (e-commerce) business application explicitly for online book shopping. It gives three distinctive sorts of profiles, for example, ordering, shopping, and browsing. Also, it is a non-profit association

generate synthetic workloads or engineered remaining burdens. The remaining tasks at hand are utilized to tackle the exploration issues in various territories like asset provisioning, load adjusting, and testing the exhibitions of the cloud. Also provided the various benchmark applications for synthetic workload generation, and the tools useful for Real workloads characterization.

Table 3 Tools used for Characterization of Real Workloads

S. No	Application Benchmark's	Description
1	R programming	R programming [16] is open-source programming it is utilized for statistical computing and graphical illustrations. R writing computer programs is generally utilized for information investigation like data analytics
2	IBM SPSS statistics	IBM SPSS [17, 18] is a product-line software that is utilized for statistical computing in sociologies. It is additionally utilized in a few areas like training, finance, health, and organizations

References

1. Mallikharjuna Rao, K., Anuradha, K.: A Hybrid Method for parameter estimation of software reliability growth model using modified genetic swarm optimization with the aid of logistic exponential testing effort function. In: IEEE Proceedings of International Conference on Research Advances in Integrated Navigation Systems (RAINS—2016), pp. 82–89 (2016)
2. Mallikharjuna Rao, K., Kodali, A.: An efficient method for parameter estimation of software reliability growth model using artificial bee colony optimization. SPRINGER-Lecture Notes in Computer Science (LNCS) Volume 8947, pp. 765–776 (2015)
3. Clarknet Trace. <ftp://ita.ee.lbl.gov/html/contrib/ClarkNet-HTTP.html> Accessed 4 May 2014
4. Worldcup Trace 98. https://en.wikipedia.org/wiki/1998_FIFA_World_Cup
5. Google trace. <https://github.com/google/cluster-data>
6. Mallikharjuna Rao, K., Anuradha, K.: An efficient method for software reliability growth model selection using modified particle swarm optimization technique. In: International Review on Computers and Software, Praise Worthy Prize Publishers, Italy, Vol. 10, No. 12, pp. 1169–1178, December 2015
7. Mallikharjuna Rao, K., Anuradha, K.: A new method to optimize the reliability of software reliability growth models using modified genetic swarm optimization. Int. J. Comput. Appl., Foundation of Computer Science (FCS), NY, USA, Volume 145 No. 5, pp. 1–8, July 2016
8. Mallikharjuna Rao, K., Anuradha, K.: An efficient method for enhancing reliability and selection of software reliability growth model through optimization techniques. J. Software, IAP Publishers, Finland, Vol. 12, No. 1, January 2017
9. Mallikharjuna Rao, K., Anuradha, K.: Performance evaluation of software reliability growth models using optimization techniques. Int. J. Comput. Sci. Inform. Secur. (IJCSIS), Pittsburgh, USA, Vol. 14. No. 11, pp. 355–362, November 2016
10. RainWorkloadToolkit. <https://github.com/yungsters/rain-workload-toolkit/wiki> (2012) Accessed 13 September 2012
11. Faban. faban.org
12. Apache JMeter. <http://jmeter.apache.org/>, 2012. Accessed 18 September 2012
13. Htpperf. <https://github.com/htpperf/htpperf>
14. RUBBoS: Bulletin Board Benchmark. <http://jmob.ow2.org/rubbos.html/> (2012). Accessed 18 September 2012
15. CloudStone Project by Rad Lab Group. <http://radlab.cs.berkeley.edu/wiki/Projects/Cloudstone/> (2012) Accessed 13 September 2012
16. TPC-W. <http://www.tpc.org/tpcw/default.asp> (2012)
17. R programming. <https://www.tutorialspoint.com/r/>
18. IBM SPSS Statistics. <https://www.ibm.com/in-en/marketplace/statistical-analysis-and-reporting>

Inset Feed Micro-Strip Patch Antenna for Communication Application Using CST



Priyanka Shah and Niraj Tevar

Abstract For many modern communication systems nowadays we require cost effective with lower dimensional antennas. Micro-strip fix reception apparatuses speak to one group of minimized radio wires that offers the advantages of a conformal nature and the ability of prepared mix with a correspondence framework's printed hardware. The different concept of feed is introducing over here because of its unique importance. So, it can be either feed by coaxial probe method or we can also use the inset micro-strip line. It is more useful as a narrow band application and that is the main reason according to its applications so it can be more applicable as a printed resonant antenna. In this paper, for the Bluetooth application, a patch antenna, using inset feed has been implemented using CST microwave Studio with its operating frequency 2.4 GHz.

Keywords Micro-strip patch · Dielectric constant · Fresnel reflection loss · Inset fed

1 Introduction

According to the physical flexibility with the Micro-strip patch antennas, i.e., cost effective and light in weight, they are more preferable for many different applications. Its geometrical structure like dimensions, shapes, and properties of material are affecting the input impedance of these antennas, Also the position and with which material the feed is applied also needs to be considered to achieve the good response at its resonance. So, we need to adjust the physical parameters of the antenna to get the best response. For the physical structure of the rectangular Micro-strip patch, the patch length generally lies within the range of $0.33 * \lambda_0$ to the half of its wavelength.

P. Shah (✉) · N. Tevar
Parul University, Vadodara, Gujarat, India
e-mail: Priyanka.shah8278@paruluniversity.ac.in

N. Tevar
e-mail: Niraj.tevar@paruluniversity.ac.in

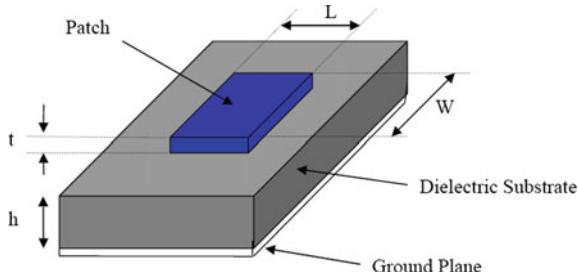


Fig. 1 Physical design of patch antenna

its thickness, i.e., patch, must be much less compared with the free space operating wavelength (λ_0). The height of the substrate, which is a dielectric material, usually between $0.003 * \lambda_0$ and $0.05 * \lambda_0$. In the Micro-strip Patch antenna, the radiation is occurred mainly due to the effect of fringing field. The ground layer and patch layer are separated by the substrate material. So due to the sandwich structure this fringing field is generated and it radiates. To achieve a better efficiency with larger bandwidth and more forward radiation in major lobe, a thick substrate requires with a low dielectric constant. But while using these kinds of low dielectric constant, the physical dimensions of the antenna increase, so need to compromise with the efficiency or the performance of antenna, the higher dielectric constant substrates are used. Patch antennas have a very high-quality factor (Q) with the higher Q, the antennas are low efficient ant with a narrow bandwidth (Fig. 1).

2 Feed Techniques

There are majorly four types of feeding techniques are used to feed the patch, i.e., Coaxial Probe method, Micro-Strip Line method, Aperture Coupling method, and Proximity Coupling method. The main focus to design and developed the patch antennas are reduction in dimension with a wider BW, also in terms of increased gain. Some of the physical parameters of the antenna, like feeding technique, the dielectric constant, and the height of the substrate are depending upon the geometry of antenna. Over here the inset fed antenna is introduced, because it is easy to design and implement. The property of antenna can be easily controlled by the length and gap of inset. While using this feed method, impedance is controlled with a planner type of feed [1, 2].

3 Design Parameters

1. Resonant Freq. (fr): As per the application, this antenna is applicable for the Bluetooth applications, 2.4 GHz frequency is selected. According to the Universal Mobile Telecommunications Service, the operating range is 2.402 and 2.480 GHz, So the antenna must be radiate between these frequency ranges.
2. Dielectric constant, also known as Relative Permittivity of the substrate (ϵ_r): Here, main objective to design the inset feed patch antenna is to reduce the physical dimension of the antenna, so higher dielectric constant is preferable. Here the relative permittivity 2 is used for the substrate material.
3. Substrate Height (hs): According to application of patch antenna as it is used in mobile phone, we must consider the parameter weight for the same. So, for light weight, we need to select as low as possible for substrate height.

Width of Patch (W_p). The mathematical representation for the Width of Patch (Fig. 2).

$$W_p = \frac{c}{2f_r \sqrt{\frac{(\epsilon_r+1)}{2}}} \quad (1)$$

Dielectric constant (Effective) (ϵ_{ref}). The effective dielectric constant is given by

$$\epsilon_{ref} = \frac{\epsilon_r + 1}{2} + \frac{\epsilon_r - 1}{2} \left[1 + 12 \frac{h_s}{W_p} \right]^{-\frac{1}{2}} \quad (2)$$

Length (Effective) (L_{ef}). The effective length is given by

$$L_{ef} = \frac{c}{2f_r \sqrt{\epsilon_{ref}}} \quad (3)$$

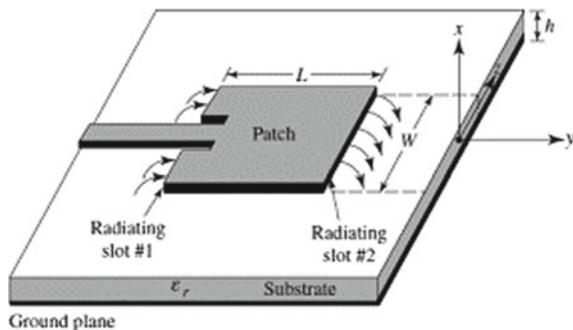


Fig. 2 Proposed Geometrical structure of an antenna

Table 1 Physical parameters

Name	Description	Value (mm)
fr	Resonant frequency (operating)	2.4 GHz
LP	Length of patch	42.55
WP	Width of patch	51.00
Wf	Width of feed line	8.669
Lf	Length of feed line	47.47
HS	Height of substrate	2.650
ϵ_r	Dielectric constant (relative permittivity)	2
Si	Distance to feed inset from patch	14.49

Extensive Length (ΔL_E). The length extension is given by

$$\Delta L_E = 0.412 h_s \frac{(\epsilon_{ref} + 0.3) \left(\frac{W_p}{h_s} + 0.264 \right)}{(\epsilon_{ref} - 0.258) \left(\frac{W_p}{h_s} + 0.8 \right)} \quad (4)$$

Patch Length (LP). The actual length is obtained by

$$L_P = L_{ef} - 2\Delta L_E \quad (5)$$

The implementation and design of the inset feed Micro-strip Patch antenna according to the parameters mentioned in Table 1, which is operating at 2.4 GHz is shown as below in Fig. 3.

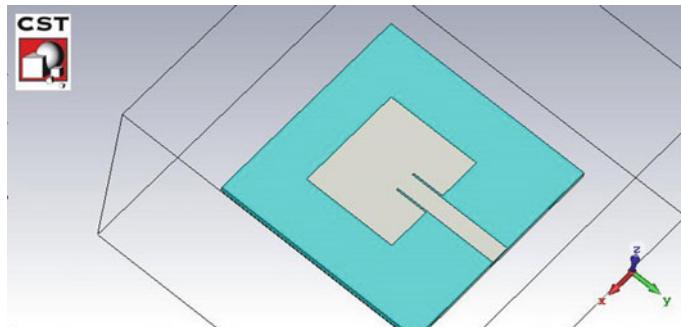


Fig. 3 CAD in CST for inset feed patch

4 Result

The Micro-Strip patch antenna with an inset feed is implemented and simulated at operating frequency 2.4 GHz for the communication applications, i.e., Bluetooth. The simulation is done in CST Microwave Studio at center frequency 2.4 GHz. The results for the implemented design as below.

Return Loss. Fresnel reflection loss or return loss is shown in Fig. 4. The value of the return loss S_{11} is -12.41 dB at the working frequency, which is 2.4 GHz. The simulated result is shown in Fig. 4.

VSWR: The VSWR value for the Micro-strip patch antenna for the resonant frequency 2.4 GHz is 1.6, which is shown in Fig. 5.

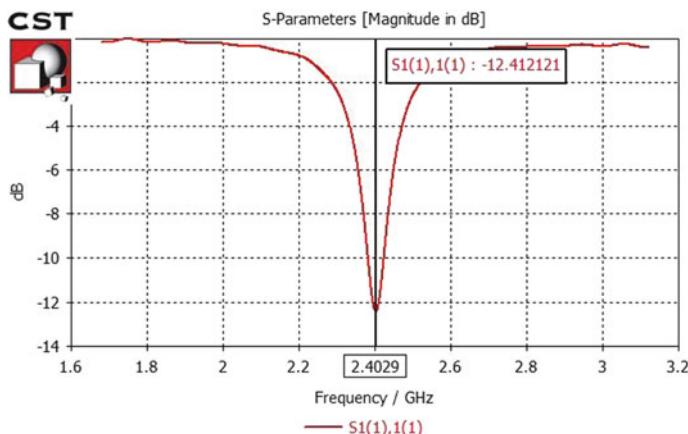


Fig. 4 Reflection Co-efficient for Micro-strip patch

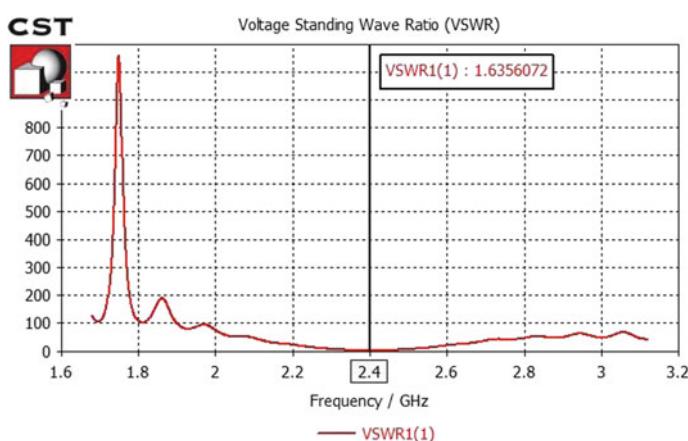


Fig. 5 VSWR for Micro-strip patch antenna with inset feed

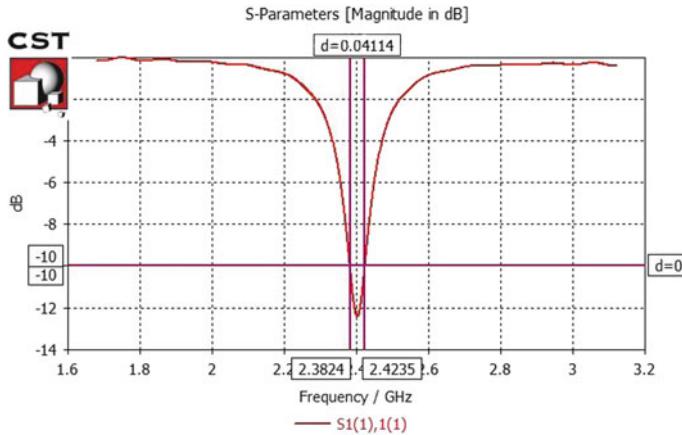


Fig. 6 Bandwidth of the Antenna at 2.4 GHz

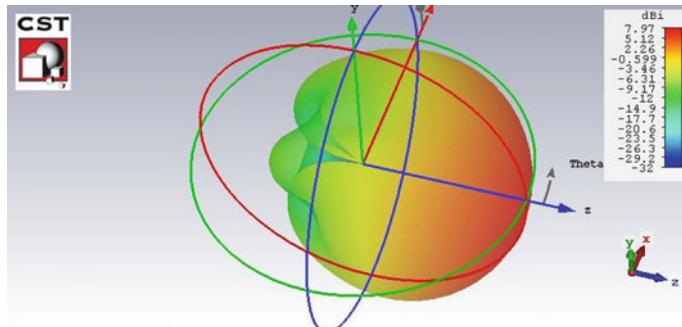


Fig. 7 3D View for directivity of the antenna

Bandwidth. The bandwidth of the simulated antenna is approximately 0.04 GHz for the communication application as shown in Fig. 6.

Gain and Directivity Pattern: The Directivity 3D pattern is shown in Fig. 7 and the 2D polar plot for the Gain pattern is shown in Fig. 8, which is 7.98 dBi

5 Conclusion

After simulating the inset feed Micro-strip patch antenna, the S₁₁ (return loss) is lower than -10 dB for center frequency 2.4 GHz and the bandwidth is nearly 0.04 GHz, which is actually a wideband. So, this antenna can be used for the communication purpose like Bluetooth.

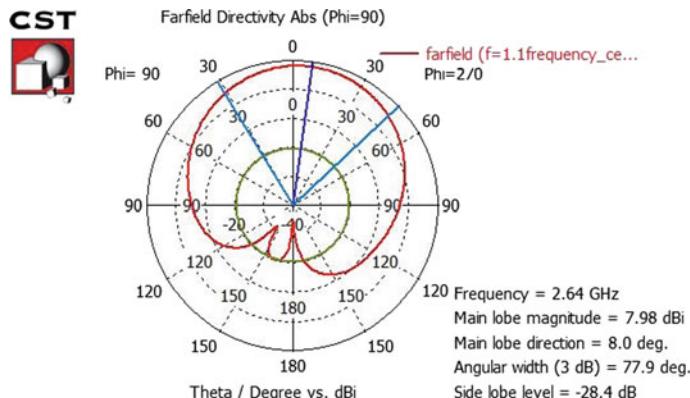


Fig. 8 2D Polar plot

Table 2 Result

Parameter	Value
Frequency	2.4 GHz
Gain	7.97 dBi
VSWR	1.6
Beam Width	77.9°

The gain value at the center frequency 2.45 GHz is 7.97 dBi and generated beam width is 77.9° (Table 2).

References

1. Basilio, L.I., Khayat, M.A., Williams, J.T., Long, S.A.: The dependence of the input impedance on feed position of probe and Microstrip line-fed patch antennas: IEEE Trans. Antennas Propagation **AP-49**, 45–47 (2001)
2. Samaras, T., Kouloglou, A., Sahalos, J.N.: A note on the impedance variation with feed position of a rectangular microstrip antenna. IEEE Antennas Propag. Mag. **46**, 90–92 (2004)
3. Garg, R., Bahl, I.J., Bhartia, P., Ittipiboon, A.: Microstrip antenna Design Hand Book: Artech House. Dedham, MA (2000)
4. Tevar, N., Chavda, N.: Designing of Microstrip Patch Antenna for 3G-WCDMA Application: National conference PIET, Vadodara Distinct ISBN 978–93–82880–33–2 in April-2014
5. Robinson, Rahmat-Samii, Y.: Particle swarm optimization in electromagnetic. IEEE Trans. Antennas Propagation **52**(2), 397–407 (2004)
6. Kaschel, H., Ahumada, C.: Design of rectangular Microstrip patch antenna for 2.4 GHz applied a WBAN. IEEE International Conference on Automation, 14 January 2019
7. Tevar, N., Mehta, P., Bhatt, K.: Numerical analysis and modeling of corrugated horn antenna with high gain and low SLL: ICMARS (2015)
8. Shankar, S., Chaurasiya, H.: Inset feed microstrip patch antenna: IEEE International Conference on Computer, January 2016

9. Surjati, I., Yuli, K.N., Astasari, A.: Microstrip patch antenna fed by inset microstrip line for Radio Frequency Identification (RFID): Asia-Pacific International Symposium on Electromagnetic Compatibility, June 2010

Load Balancing Approach and Diminishing Impact of Malicious Node in Ad Hoc Networks



Shrikant V. Sonekar, Rohan Kokate, Manoj Titre, Aniket Bhoyar, Merajul Haque, and Sachin Patil

Abstract Two aspects are very important in any network, one is the uniform distribution of load among all the nodes and another one is addressing the security issues. The battery drain problem and power consumption by the nodes are more or less related with the security issues as nodes are continuously moving and the nodes which are very close to the transmission area generally gets overloaded with traffic. Ad hoc wireless network is defenseless and more exposed to almost all categories of the attacks with supplementary to susceptible, whether the attack is novice or old one. Hence, there is a need of some mechanism which could identify the attacks at the early stage. Few of the attacks uses the union of statistical and algebraic techniques for targeting the key element. They also target the scientific properties of the cryptographic algorithms. The trusted server issues the certificate to all the nodes and could help in identifying the malicious node to some extent. Load balancing scheme is very useful to distribute the load traffic uniformly in the network. The paper basically focuses on two important aspects of networks, i.e., equal distribution of load among the nodes in the network and diminishes the stringent impact of attack to greater extent in the premature stage.

Keywords Further route request · Further route reply · PIRRS · PORRS · Power consumption · Load balancing · Trusted server · Ad hoc wireless network

1 Introduction

Ad hoc wireless network also called as network without infrastructure is defined as the network which drives without the provision of any motionless infrastructure and makes use of multi-hop radio transmitting technique. Moreover, achieving the time synchronization is challenging task and it consumes more bandwidth. The central goal of routing is to discover paths with least possible overhead and also rapid reconfiguration of shattered routes. But, Ad hoc wireless network has strengthen its

S. V. Sonekar (✉) · R. Kokate · M. Titre · A. Bhoyar · M. Haque · S. Patil
JD College of Engineering and Management, Nagpur, Maharashtra, India

importance in almost all the wireless applications due to their speedy and economically a lesser amount of demand for setting out networks. Due to the nonexistence of centralized planner, it is defenseless and more exposed to almost all categories of the attacks as it is more susceptible and routing becomes more complex, whether the attacker is beginner or deep-rooted is immaterial. Sometimes the system fails due to misleading in recognizing the attacks [1].

The false rejection rate and acceptance rate are majorly happening in the ad hoc wireless networks, the network becomes helpless when most of the existing routing protocols don't try to find out new route if there is no change in the topology though there is a congestion problem.

- False Rejection Rate (FRR): System supposed to recognize a person but fails to recognize the system. It is measured in terms of the ratio of false refusal to the total quantum of attempts in percentage.
- False Acceptance Rate (FAR): System supposed to recognize an intended person but fails to recognize the intended one by the system. It is measured in terms of the ratio of false recognition to the total quantum of attempts in percentage.

This leads to the concentration of more traffic on some specific nodes. This is not desirable as the wireless devices have shortage of computing capability and battery power. Hence, there is a need of resolving the traffic concentration problem otherwise the nodes give up their role of packet forwarding and unnecessarily system declares it as malicious nodes, though they are not malicious. The node which is at center location leads to faster battery drain as it is continuously accepting and transmitting data packets, which ultimately forms the ring, thus reducing the normal life of the closer nodes and due to this the overall efficiency of the nodes can be significantly come down, thus leads to the coverage problem in the ring area which is referred as ring formation problem [2–4].

This problem motivates toward the mechanism for the systematic load balancing scheme which is also referred as efficiency enhancing scheme. There are few cases of attacks wherein the restriction of the intermediate nodes from originating route reply packets could decrease the impact of attacks severity. In such cases, the destination node only would get a chance to initiate the process of route reply packets.

This process still cannot give complete assurance for the security from the malicious nodes in the networks. As the size and density of the network region increases, the postponement and interruption involved in the route discovery process also increases as both are equally proportionate with each other. In order to get assurance for the path existence from intermediary to the extreme node, the route request packets have to be sent from the starting node to the neighbor node of the intermediary node as soon as the packet meant for route reply is received from one of the intermediary nodes [5].

2 State-of-The-Art Scenario

Consider the state represented diagrammatically wherein the route request packets are sent by the source node and obtain route reply in the course from intermediary malicious node M_n , whatever the packets received from the node M_n contains information about the subsequent hop neighbor node E_n . Figure 1 shows that the starting node S_n sends further route request packets to the neighbor node E_n whereas node E_n responds to the starting node S_n with a further route reply packet. Moreover, the malicious node M_n is not listed in the neighbor's node routing list, i.e., node E_n .

The best part of this scenario is that the further route reply packet sent by neighbor node E_n does not contain a route to the malicious node M_n . This technique completely eliminates the stern impact of the attacks as new route to the extreme node through node E_n is selected if at all there is a route to the extreme node D_n and previously preferred route through the malicious node M_n is discarded.

The issue with this technique is that the control overhead increases extensively and technique fails if the malicious nodes are in group, i.e., the technique works more significantly against single attacker. The fake symmetric route is established between the nodes P_1 and P_2 . The disproportionate routes are more likely to be witnessed in heterogeneous networks, i.e., MANET. Consider an example shown in Fig. 2 which depicts two unequal routes or links wherein the node- P_1 connects with P_2 and another link shows that P_2 connects with Z .

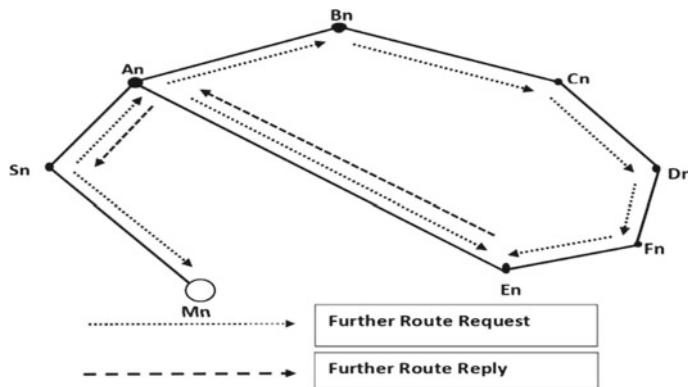
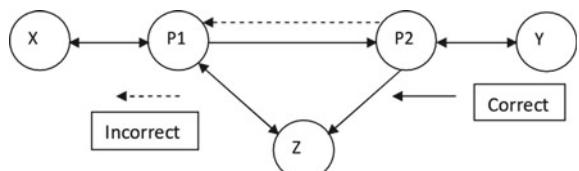


Fig. 1 Broadcast of further route request and further route reply

Fig. 2 Counterfeit symmetric route between nodes P_1 and P_2



Due to this scenario, the malicious node M gets benefitted and could launch an attack. In order to get benefit, the malicious node-Z attempts to construct a deceitful symmetric link between two nodes, i.e., P1 and P2.

The formation of this counterfeit symmetric route takes following process, it is depicted in terms of different messages:

message – 1 : Node – P1 → {Φ : Hello, {Nil}}

When node-P1 wants to find out the neighbor node, it broadcasts blank Hello message that reaches to both the nodes P2 and malicious node-Z.

message – 2 : Node – P2 → {Φ : Hello, {Node – P1, Asymmetric malicious node – Z}}

After that P2 announces that P1 is an Asymmetric neighbor in its Hello message, this is shown in message-2

message – 3 : Malicious node – Z → {Node – P1 : Hello, {Node – P1, Asymmetric malicious node – Z}}

After getting the message-2 by the malicious node-Z, it forwards to node-P1 with the intention that node-P1 should not reply, shown in message-3.

message – 4 : Node – P1 → {Φ : Hello, {Node – P2, Symmetric malicious node – Z}}

The node-P2 updates its route status with node-P1 to symmetric. When node-P1 receives the message-3, it comes to know that its distinctiveness has been advertised in the neighborhood slant and therefore node-P1 settles on the decision that node-P2 is symmetric neighbor.

Therefore, based on this conclusion, it decided and floated this newer route information in hello message as

message – 5 : Node – P2 → {Φ : Hello, {Node – P1, Symmetric malicious node – Z}}

Due to this, the duped nodes, i.e., P1 and P2 puts the conclusion that they are associated by means of symmetric link but actually they are not, thus all the generated packets does not reach to the complete network and therefore the network is divided in to different parts.

The notation Φ signifies the message broadcasting. The following algorithm depicts the importance of load balancing for deciding the controlling node. Though the controlling node manages the proportionate of load among all the nodes, still the nodes keep on changing its location in the region of high load and low load band, due to this the problems arises in the formation of ring [6–8].

```

Input: Allotment of some quantum of initial energy
to the nodes
Output: The node which has long lasting energy is
declared as controlling node
//Setting the finalization values for controlling node
    Set node id= create node;
    Set preliminary Energy= $preenergy;
        Set long lasting Energy= $llenergy;
    Set default as radio range value;
// Set conditions
    If ((required node exists in radio range) &&
        (subsequent hop!= Empty))
    { Capture the load of all the nodes in the
        network (network node_all)};
    Configure the node { packet_type (routereq, routerep,
        node_distance, preenergy, llenergy, Time, pktsend,
        pktrcv, pktdrop)}
        //Set node_distance=node_range
        for (n=0; n<mm; n++)
    usedenergy[n] = preenergy[n]-llenergy[n]
        fullenergy[n] = usedenergy[n]
    //Set conditions
        if(fullenergy[n] < usedenergy[n])
            { fullenergy = usedenergy[n] node id = i;};
    //Set conditions
        if(node_distance <=300 && fullenergy[node_id]>
            neighbor_node_energy [node_id])
    //check condition
    {fullenergy node is declared as controlling node;}

```

The high mobility scenario of the nodes need not be considered as severe but the problem of dropping the calls in the network cannot be ignored. The objective of load balancing scheme is to distribute the load traffic uniformly amongst all the nodes in the network so that none of the node becomes overloaded [9].

The position of the node and its region density matters a lot in the network and due to this, there should be proper load balancing scheme in the network. Generally the shortest path is chosen for transmission of the packets, whenever such scheme of selection of path is chosen for transmission of the packet, the node which is closer to the center gets more traffic than the node which is away from center. The generated traffic is always highest at the center and it goes on decreasing with the increase in the distance from the network center, as it is inversely proportionate with each other as shown in Fig. 3a, b. The load balancing plays vital role to solve this problem in the networks, as the shortest path distribution creates load imbalance problem in some part of the network, thus the longest path concept has been used at large which redistribute the load, thus the load balancing could be performed in proper manner and this way the battery life of the node gets improved.

However, the overall battery consumption problems could not ignore as the longer path is used which ultimately increases the time for packet delivery and there is huge

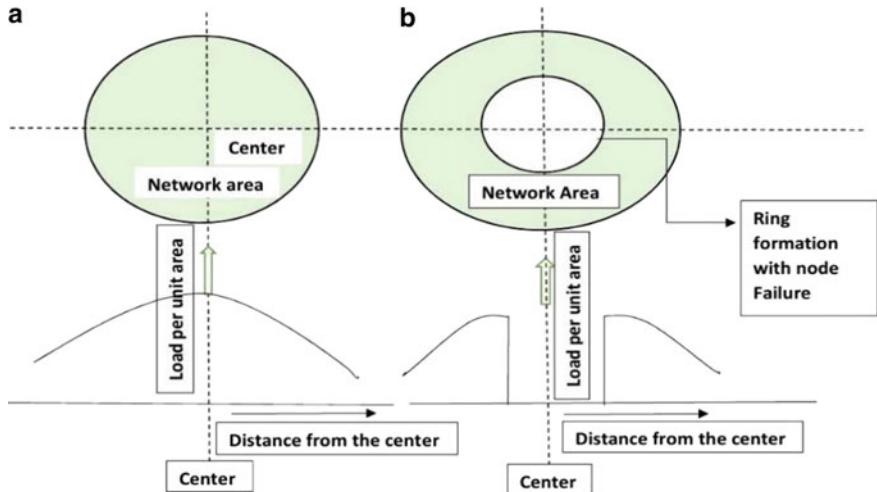


Fig. 3 **a** Variation of load density with distance, **b** Ring formation due to node failure

quantum of reduction in overall efficiency of nodes which also adversely affects the efficiency of the network.

The impact of this issue is that the mechanism known as automatic rate fallback also reduces to various lower rates as the distance between the sender and receiver increases. The Preferred Inner Ring Routing Scheme (PIRRS) and Preferred Outer Ring Routing Scheme (PORRS) tackle the problem of edge weight parameter in the computation of shortest path. In former, the packet preferably routed through the inner layer of the two rings whereas when the nodes belong to different rings, the packet goes in the radial direction and when the nodes belong to same rings, it is transmitted in the same rings. The former is exactly opposite to PIRRS, the packet to be transmitted remains in the outer of the source and destination rings for the maximum time. Both the routing schemes are static in nature and could not manage the load balancing in proper manner when the generated traffic is not distributed uniformly [10–12].

3 Dynamic Power Adjustment Technique

It is based on the link affinity, the node stability is the major issue and therefore there is no assurance of node stability in Ad hoc wireless networks as it is susceptible to frequent link let-downs due to constant node moving feature, thus leads to the reduced efficiency of the network. Such reduced efficiency network is not acceptable for transmission of data packet and therefore there is a huge requirement of mechanism or a protocol which selects a path that assures good response in terms of link stability. The mechanism which defines and decides the route stability is known as an affinity.

Consider that node “x” illustrate a set of signals from the node “y”, the affinity “A_{xy}” is calculated as follows:

$$\text{high, : if } \delta S_{sy}(\text{avg}) > 0$$

$$A_{xy} = \frac{\{S_{\text{thre}} - S_{xy}(\text{current})\}}{\delta S_{xy}(\text{avg})}, : \text{if no conditions field} \quad (1)$$

The node “x” and “y” signal strength denoted as “S_{xy}(current)” is below the threshold level (S_{thre}) then chances of discontinuing the established link between two nodes “x” and “y” increases. The rate at which the signal strength is changed for some samples is the average for both the nodes “x” and “y”.

In the cluster form networks, each node’s responsibility is to broadcasts the hello packets, the purpose is show that the node is alive. In order to transmit hello packets periodically, it requires continuous power. Some time is required for arrival of the hello packets and the time interval (T_i) for Hello packets is represented

$$\begin{aligned} S_{HP} - \{S_H - \{S_{\text{thre}}|la\} * T_i\} &: \text{if moving far and } T_i < la \\ S_{s,s+r} - \{S_H : \text{if moving closer and } T_i < la \\ S_{\text{thre}} &\text{ if no condition satisfied} \end{aligned} \quad (2)$$

where SHP is the signal strength of the Hello packet received, T_i is the time period between two successive hello packets and “la” is the link affinity between a node and its neighbor.

The receiver on the other hand calculates the signal strength of the sender node on receiving the hello packets from the sender. The signal strength of the sender is denoted as (S_s, s+r). After calculating the signal strength the node adjusts its transmission power (P_r) accordingly. The new adjusted transmission power (P_s, s+r) is given by

$$P_{c,s+r} = \{P_r * (S_{\text{thre}}|S_s, s+r)\} \quad (3)$$

The dynamic power adjustment helps in saving the power and using this scheme each node could transmits the data packet with the minimum power [11].

4 Multiple Exclusive Shortest Path Routing

It searches for a variety of the paths in the band region which is closest to the shortest path and tries to reduce the average hop length of the selected paths along with the traffic fairness. It finds computationally difficult the first k-shortest paths as the paths may not be sufficiently node disjoint and defines the paths as follows. The path which

is shortest in the network could be identified by Dijkstra's shortest path algorithm. All edge weight reaches to the high value which is not an infinity, a high value which prevents selection of nodes as far as possible and at the same time allows the nodes if no other path exists [13]. The shortest path algorithm, i.e., Dijkstra's algorithm needs to run again for finding the path which is shortest. When there is a choice for selection of paths, the path which is least loaded is selected.

```

Multiple Exclusive Shortest Path Routing Based Algorithm
Define (high_value)
    high_value(hv)= a high value(hv)<> ∞ (INFINITY)
Then go for Path_Search
    [Source (SRC), Destination (DST)]
    { Define (edge_weights_value)
        Edge_Weights_Value=1
        //Search using Dijkstra's Algorithm with
        Edge_Weights_Value=1 }
        MESPRBA[Source (SRC), Destination (DST)]
        { For i:=1 to n do {
            Path_Search(ps) [Source(SRC), Destination(DST)] }
            {{ It is applicable to all the nodes in the
            coming back route }}
        Modify and update all the edge weights to hv
        Find maximum load on each of the nodes for "n"
        Numberof paths
        The path which has least value of Maximum (load on
        each nodes of "n" paths)will be returned back }

```

5 Small Minimum Energy Communication Network

It is used to construct a sub-network from a given network, thus minimizes the usage of the energy at large. If graph G represents the entire network then the subgraph is represented by “G-1” and reason behind constructing is to minimize the energy utilization which plays very important role for the functioning of the network. Though the subgraph “G-1” has less number of edges as compared to G but the nodes which are offered for the graph G are all being retained in subgraph.

The subgraph “G-1” is constructed in such a manner that the connectivity among the nodes is not dislocated. Moreover, the energy required for transmitting the data packets among the nodes is lower in the subgraph than in graph G. Consider two neighbor nodes “s” and “t”, the minimum power requires for transmitting data packets between two neighbor nodes “s” and “t” are shown as

$$p(s,t) = m * d(s,t)^n \quad (4)$$

Where the term “m” indicates a constant, “n” is the path failure exponent which indicates the loss of power as the node moves far from the originator, i.e., transmitter. $d(s,t)$ is the space between the neighbor nodes “s” and “t”. Consider the term “cp” which indicates the power needed for receiving the data within the time. The transmission of the data packet is more economic when it is being transmitted by smaller hops because as the nodes move far and far, the transmission power increases exponentially with the moving data packets. Consider the path of two neighbor nodes and is denoted as s (i.e., s_0) and t (i.e., s_k) and is represented by $j = (s_0, s_1, \dots, s_k)$ such that each (s_i, s_{i+1}) is an edge in the “G-1” which is subgraph of G. The total quantum of power consumed for the transmission of data packets is

$$PC(j) = \sum_{i=0}^{k-1} (p(s_i, s_{i+1}) + cp) \quad (5)$$

If the condition $PC(j) \leq PC(j-1)$ satisfies all the paths “j-1” between “s” and “t” in G then it is minimum energy path. “G-1” is said to satisfy the property of minimum energy, if there exists a path j in “G-1” which is minimum energy path in G, for all the node pairs (s,t). It uses only the minimum energy paths from subgraph for data transmission therefore the overall energy consumed for data packet transmission is minimized. Power consumption by all the nodes in the uniform way in the cluster and distributing the power among the nodes in the uniform manner thus leads to minimize the overall transmission power has ambiguity. Whenever the data packet is being transmitted from node A to B, the power required is inversely proportional to the nth power of the space(s) between the two nodes, i.e., $1/d^s$ where “s” varies from 2 to 4 based on the space distance between the nodes and also the terrain. In order to assure the successful transmission of data packets from node S_x to S_y , the signal-to-noise ratio value of node “y” is to be greater than a specific threshold value $thre_j$. The SNR at the receiving node s_j can be mathematically represented using the condition as

$$SNR_y = \frac{PTxGPx,y}{\sum_{k*x} PTkGPk, y + ny} thre_j(BER) \quad (6)$$

where the term PTx is used to denote transmission power of host S_x . The path gain is denoted using $GPx,y = 1/d^s x, y$ for the hosts S_x and S_y . S_y is the thermal noise and BER which is based on the value of $thre_j$ is the bit error rate [14–16].

6 Least Power Consumption Routing

It prefers the route from source to destination that has minimum total transmission power among all the available routes. LPCR algorithm can be considered by altering the Dijkstra's shortest path algorithm. But by doing this, the algorithm may pick up a path with more number of hops, thus the nodes require a lesser amount of power. This leads to increase in the end-to-end delay of the data packets in the cluster. Along with this, the more number of hops ultimately reduces the route stability as the node mobility is one of the intrinsic appearances of the ad hoc wireless networks. Due to such constraints, the Bellman ford algorithm is taken into consideration as one of the remedies. This algorithm takes transmission power as a cost metric into the account. The power cost is given by

$$\{(C_{x,y}) = Q_{\text{transmit}}(nx, ny) + \text{Cost}(ny) + Q_{\text{transceiver}}(nx)\} \text{ where } nx, ny \text{ are nodes} \quad (7)$$

The term $Q_{\text{transmit}}(nx, ny)$ is the transmitter power of node x to reach node y and $Q_{\text{transceiver}}(nx)$ selects the route with small quantum of number of hops and is the transceiver power of node j. The cost function at node "nx" is given by

$$\text{cost}(nx) = \min_{y \in \text{NN}(x)} C_{x,y} \quad (8)$$

The term $\text{NN}(x) = \{y, ny\}$ and is a neighbor node for the node "nx". Though the algorithm assures the reduction of the overall power consumption of the cluster but does not assure whether there is a reduction of overall power consumption at individual nodes or not. It is a secure routing protocol which works at the network layer and is based on cryptographic certificates, able to defeat all the identified attacks in same layer. Though authenticated routing for Ad hoc networks takes precaution for authentication of messages, its integrity, and non-repudiation, but at the same it expects a small quantum of prior security coordination among all the nodes.

The source node broadcasts route request packets during the route discovery process whereas the destination node responds by unicasting back reply on the selected path on receiving the route request packets from the source node. This protocol assures for secure route establishment by using an introductory cryptographic certification process followed by an end-to-end route authentication process [17–19].

7 Issue of Certificates for Authentication

The certificate is issued to the nodes in the ad hoc wireless network as there exists an authenticated trusted server, the public key of this trusted server is known to all legal nodes in the network. The protocol has no specific key distribution algorithm whereas

it assumes that keys are generated a priori by the trusted server and distributed to all nodes in the network. Consider the network “N” which has node “B”, the node gets certificate only after completing the process of joining to the network. The trusted server TS issued the following certificate to the node B:

$$((T_s \rightarrow B: Cert_B = [LAB, KB+, TOCC, TOEC]PK_{TS})) \quad (9)$$

where the terms “B” indicates the node, LAB is the logical address for node “B”, KB+ is the public key for node “B”, TOCC is the time of creation of the certificate, TOEC is the time of expiry of the certificate, and PKTS is the private key of the trusted server. This process helps in authentication of the nodes in the network as when the trusted server verifies the intended nodes in the network, it uses the certificates issued mechanism, the node which could not revert back with correct information is treated only as suspicious node, not the malicious node, the certificate information cannot reveal the node is a malicious node or not [20].

The foremost goal mouth of this destination to destination route validation and confirmation process is to make certain that whatever the data packets are being sent from the source node should reach to the truthful intended destination, i.e., the end user node. The starting node “Sn” broadcasts a Route Request or Route Discovery data packet meant for the end user node “En”. The packet which is requesting the route contains the packet identifier also known as path detection procedure (PDP), the logical address of the end node (LAEN), the certificate of the starting node Sn (CertificateSn), the up-to-date time (t), and nonce value (NSN). A nonce is a random or pseudo-random number which is issued to ensure that the previous communications are not reclaimed in attacks, i.e., it assures the authentication process in replay attacks. The procedure is given as

$$S_n \rightarrow \text{broadcasts} := [PDP, LA_{EN}, Cert_{Sn}, N_{SN}, t]PK_{Sn-} \quad (10)$$

Here, PKSn- is the private key of the starting node Sn. Whenever the starting node sends a route discovery message, the arbitrary value of nonce will automatically increment. It is used in combination with the time stamp which acts as counter in order to ensure the recycling process of nonce easier. Whenever a node receives a route discovery process packet from the starting node with a greater value of the starting nonce than that in the formerly received route discovery process packets from the same starting node, then it creates a record of the neighbor from which it received the packets. Then encodes the packet further with its own credential in terms of the certificate and then adopts the broadcasts process. The following equation denotes the process.

$$I_n \rightarrow \text{broadcasts} := [[PDP, LA_{EN}, Cert_{Sn}, N_{SN}, t]PK_{Sn-}]K_{Sn-}, Cert_{Sn} \quad (11)$$

After sending the data packets from the starting node “Sn” to the End node “En”, an intermediate node “In” receives route discovery process from starting node “Sn”, then an intermediate node “In” removes the certificate from the data packet and

inserts the certificate of the intermediate node “In” and then forward the data packet further. The responsibility of the end node “En” is to verify certificate and the value of tuple (NSN,t) on receiving route discovery process packet by the destined node. The end node “En” uses the process of unicasts (REPLYPACKET) to the starting node “Sn” along with inverse path as

$$E_n \rightarrow N_x := [REP, LA_{Sn}, Cert_{EN}, N_{SN}, t]K_{En-} \quad (12)$$

The term “Nx” is the neighbor node to the end node “En” and is the node which has initially started the process of forwarding the route discovery process packet to the end node. LASn denotes logical address of the starting node, KEn denotes public key for the end node.

$$\{ \langle \text{Generated Certfails} | \text{Time Stamp} | \text{NonceValue} \rangle \neq \\ \langle \text{Required Value} \rangle \text{then} \langle \text{Generate ERROR Message} \rangle \} \quad (13)$$

The REPLYPACKET uses the same process followed by a PDP packet. If the generated certificate fails or the time stamp or the nonce value does not matched as per the requirements mentioned above then an error message is generated. The generated error message is same as that of other packets, only with the addition that the identifier field of the packet is replaced with ERROR message [21, 22].

8 Conclusion

An ad hoc wireless network is defenseless and more exposed to almost all categories of the attacks as it is more susceptible, where the attack is novice or old one. Therefore, there is a need of some mechanism or technique which could recognize the attacks at the premature stage. Few of the attackers are using the union of statistical and algebraic techniques for targeting the key element. They also target the scientific properties of the cryptographic algorithms. Some of the proficient attackers use the permutation and combination technique for detecting the property of the key as well as divide and conquer mechanism for reducing the complexity of key guessing parameters. The trusted server issues the certificate to all the nodes and could help in identifying the malicious node to some extent. Load balancing scheme is very useful to distribute the load traffic uniformly in the network if properly implemented. The position of the nodes and their density matters a lot for uniform load distribution and hence equilibrium of load throughout the network. The overall efficiency of the network is highly dependent on the appropriate distribution of load in the network otherwise there could be battery drain problem and dislocation of the nodes in the destined path. Dividing the network into sub-networks helps in improving the power consumption and is one of the parameters for improving the overall life of the network. Still there is a scope for researchers to improve the overall efficiency of the network and minimum the power consumption of the network.

References

1. Saidi, A., Benahmed, K., Seddiki, N.: Secure cluster head election algorithm and misbehavior detection approach based on trust management technique for clustered wireless sensor networks (2020). <https://doi.org/10.1016/j.adhoc.2020.102215>
2. Aslan, Aslan, Ö., Samet, R.: A comprehensive review on malware detection approaches. *IEEE Access* **8**, 6249–6271 (2020)
3. Fu, H. et al.: A Data clustering algorithm for detecting selective forwarding attack in cluster-based wireless sensor networks. *Sensors* **20**(1), 23 (2020)
4. Baza, M. et al.: Detecting sybil attacks using proofs of work and location in vanets. *IEEE Trans. Dependable Secure Comput.* (2020)
5. Gomathy, V. et al.: Malicious node detection using heterogeneous cluster based secure routing protocol (HCBS) in wireless adhoc sensor networks. *J. Ambient Intell. Humanized Comput.*, 1–7 (2020)
6. Yang, H. et al.: A novel algorithm for improving malicious node detection effect in wireless sensor networks. *Mobile Netw. Appl.*, 1–10 (2020)
7. Sonekar, S.V., Pal, M., Tote, M., Sawashere, S., Zunke, S.: Computation termination and malicious node detection using finite state machine in mobile Adhoc networks. In: 2020 7th IEEE International Conference on Computing for Sustainable Global Development, New Delhi, India, pp. 156–161 (2020). <https://doi.org/10.23919/indiacom49435.2020.9083710>
8. Liu, L., Yang, J., Meng, W.: Detecting malicious nodes via gradient descent and support vector machine in Internet of things. *Comput. Electr. Eng.* **77**, 339–353. <https://doi.org/10.1016/j.compeleceng.2019.06.013> (2019)
9. Yadav, S. et al.: An Effective approach to detect and prevent collaborative Grayhole attack by malicious node in MANET. *International Conference on Intelligent Systems Design and Applications*. Springer, Cham (2019)
10. She, W. et al.: Blockchain trust model for malicious node detection in wireless sensor networks. *IEEE Access* **7**, 38947–38956 (2019)
11. Yang, H., Zhang, X., Cheng, F.: a novel wireless sensor networks malicious node detection method. *International Conference on Security and Privacy in New Computing Environments*. Springer, Cham (2019)
12. Jamal, T., Butt, S.A.: Malicious node analysis in MANETS. *Int. J. Inform. Technol.* **11**(4), 859–867 (2019)
13. Kumar, J.A., Tokek, V., Shrivastava, S.: Security enhancement in MANETs using fuzzy-based trust computation against black hole attacks. *Information and Communication Technology*. Springer, Singapore, pp. 39–47 (2018)
14. Sonekar, S.V., Kshirsagar, M.M., Malik, L.: Cluster head selection and malicious node detection in wireless Ad Hoc networks. *Advances in Intelligent Systems and Computing*, vol. 638. Springer, Singapore. https://doi.org/10.1007/978-981-10-6005-2_55 (2018)
15. Zawaideh, F., Salamah, M., Al-Bahadili, H.: A fair trust-based malicious node detection and isolation scheme for WSNs. *2nd International Conference on the Applications of Information Technology in Developing Renewable Energy Processes & Systems (IT-DREPS)*. IEEE (2017)
16. Malathi, M., Jayashri, S.: Design and performance of dynamic trust management for secure routing protocol. *IEEE International Conference on Advances in Computer Applications (ICACA)*. IEEE (2016)
17. Singh, U. et al.: Detection and avoidance of unified attacks on MANET using trusted secure AODV routing protocol. *Symposium on Colossal Data Analysis and Networking (CDAN)*. IEEE (2016)
18. Sonekar, S.V., Kshirsagar, M.M.: Mitigating packet dropping problem and malicious node detection mechanism in Ad Hoc wireless networks. *Advances in Intelligent Systems and Computing*, vol 404. Springer, https://doi.org/10.1007/978-81-322-2695-6_27 (2016)
19. Qureshi, N.: Malicious node detection through trust aware routing in wireless sensor networks. *J. Theor. Appl. Inform. Technol.* **74**(1) (2015)

20. Hui, C., Yan, D.-D., Pan, J.-I.: Malicious node detection algorithm based on reputation with voting mechanism in wireless sensor networks.. *China Univer. Metrol.* **24**(4), 353–359 (2013)
21. Su, M.-Y.: Prevention of selective black hole attacks on mobile ad hoc networksthrough intrusion detection systems. *Comput. Commun.* **34**(1), 107–117 (2011)
22. Jiang, W., Li, Z., Zeng, C., Jin, H.: Load balancing routing algorithm for Ad Hoc networks. 2009 Fifth International Conference on Mobile Ad-hoc and Sensor Networks, Fujian, pp. 334–339 (2009). <https://doi.org/10.1109/msn.2009.81>

ELEMENT: Text Extraction for the Dark Web



Ashwini Dalvi , Irfan Siddavatam, Apoorva Jain, Smit Moradiya, Faruk Kazi, and S. G. Bhirud

Abstract The increasing amount of data on the Internet has been a constant challenge when text cleaning and relevant-text extraction are of interest. One of the areas of focus on the internet is the Dark Web; the data here is much more volatile and dynamic. With more researchers looking for data and extracting information, the algorithms have always been in a state of constant improvement. The solutions currently offered, all work based on text feature extraction algorithms like TF-IDF, Bag of Words, Word2Vec. Discussion on these extraction methods is well documented but a critical evaluation among these algorithms is amiss from standard literature. This paper discusses a balanced approach for tagging extracted data; ELEMENT (Effective Lemmatization, Efficient Management of Extracting Noteworthy Tags) which is a modified form of TF-IDF. Having a balanced approach like ELEMENT will benefit from being able to perform well under any given circumstance. The paper discusses and compares the proposed approach with existing strategies of text feature extraction. This comparison spans across accuracy of feature extraction, efficiency concerning Time and Space, concluding with a simplistic view of the strengths and weaknesses of each algorithm.

Keywords Text feature extraction · TF-IDF · Dark web

1 Introduction

Researchers have always been interested in looking into what text signifies, what it means and represents; recognition of context, based upon the text and then deducing what any excerpt of text is about. This steady progression of domain classification, domain understanding, and then context recognition, all are based on the roots of

I. Siddavatam · A. Jain · S. Moradiya

K. J. Somaiya College of Engineering, Vidyavihar, Mumbai 400077, India

A. Dalvi · F. Kazi · S. G. Bhirud

Veermata Jijabai Technological Institute, Mumbai 400077, India

e-mail: ashwinidalvi@somaiya.edu

text feature extraction algorithms. These algorithms recognize and decide what text is relevant and what is not. Some of the standard text feature extraction algorithms are TF-IDF, Bag of Words, and Word2Vec. However, when one look for the evaluations for these algorithms; their performance measure is mostly unknown, making it tedious even to select an algorithm that provides what is required with the most efficiency. Moreover, with the ever-increasing amount of data in the world, it has become crucial that these algorithms perform as expected and efficiently. A closer look into these algorithms shows room for another more balanced methodology, ELEMENT.

A major challenge is the complexity of the identification of data interfaces due to the data characteristics of Dark web data. The efficient, accurate and adaptive dynamic tagging of content on a webpage is among the various requirements of a resource identifier that need to be fulfilled [1]. Research about the nature and structure of the web and its content has focused primarily on manually labeling a set of crawled sites against a series of categories, sometimes using those labels as testing set for subsequent automated crawls (web crawling is widely used to get huge amounts of data) [2]. The manual labeling of content is a work-intensive as well as time-consuming task. Methodologies, like feature extraction, are used to reduce the burden of labeling the content, making it essential for these methodologies to be efficient.

The ELEMENT methodology is a variant of TF-IDF (Term Frequency—Inverse Document Frequency) algorithm. It is a numerical statistic designed to determine the value of a word to a document in a series or a corpus. The TF (Term Frequency) value is directly equivalent to the number of times a term appears in the text. It is further offset by the IDF (Inverse Document Frequency) value, which denotes the number of documents in the corpus containing the phrase. [3] The IDF factor requires multiple documents, it is not ideal to use the TF-IDF algorithm to extract text features from a single web page, making this a drawback for the algorithm. The methodology of ELEMENT mitigates this drawback as it is designed for text feature extraction, even from a single webpage.

Another aspect of this paper is to compare all existing Text Feature Extraction methodologies along with ELEMENT to perform an analysis of which one would suit any dataset. To conduct this research, a dataset with diversified domains which resemble each domain as accurately as possible is required. Given the amount of data present on the web, it has been a primary use case for Text Feature Extraction algorithms, performing the tests on a multi-domain web dataset will provide accurate results. The web dataset, DBpedia, which has Wikipedia data from various domains is chosen to ensure proper and diverse testing which will provide quality results.

In the further sections, we discuss similar research works followed by the methodology followed for the development of ELEMENT. The works discussed under the related works section include almost all the important works done in the field of tagging and classification from various authors working in the field for a long time. The methodology discusses the approach we have employed for the development and explains the components of ELEMENT. The paper further moves on to discuss the

evaluation and results of various text extraction methodologies by comparing them for various selected factors.

2 Literature Review

Text extraction is the objective of a text feature extraction methodology and that requires a huge amount of data to be tested. In order to do so, a data source of equally vast amount and type of data would be the necessary. Fortunately, the Internet presents itself an answer there, given the amount of data present on the web, it is the primary location where such algorithms are tested. And in order to collect this huge amount of data, the assistance of a web crawler is taken. A Web Crawler is a mechanism that automatically collects and stores data from all interconnected webpages of a given website. And it comes as no surprise that most papers have talked about text feature extraction methodology in terms of the Internet itself. In this section of the paper, we have reviewed some of the common text feature extraction methodologies employed in/with crawling methodologies.

As discussed in the introduction, approaches from various research works have been studied. It is realized that some commonly used and highly explored algorithms include top-k, bag of words, TF-IDF, and word2vec algorithm. Most present research work on the classification module focuses on these approaches and their adaptations.

One of the approaches involved the use of model-based on: single term BOW model and comma-separated terms BOW model. It was found that this model outperformed the traditional techniques based on calculated precision and recall value. It takes both structured and unstructured sources from the deep web as well as both advanced and straightforward query interfaces into consideration [4]. Another approach uses the TF-IDF as keyword or appropriate word extraction method, link classification method, and classification technique for its crawler. In some approaches, it is proposed to use the TF-IDF algorithm with various modifications based on the specific requirements of the crawler [4–6]. An approach discusses the use of TF-ICF (Term Frequency-Inverse Class Frequency), which calculates the popularity score for various categories instead of TF-IDF which calculates the popularity score of words. However, the results of the initial experiments indicated a poor performance of this approach as opposed to its counterpart [3]. TS-IDS (Term Size-Inverse Document Size) is also mentioned in a paper which is yet another modification of the TF-IDF algorithm [7]. Another adaptation of the TF-IDF algorithm proposed is W-TF IDF which uses a weighting method. On comparison with conventional TF and TF-IDF algorithm, this approach was found to be more efficient and stable [8]. However, many models have been proposed to attempt the problem of classification in a unique and unconventional approach.

A paper suggests a four-step model as a classifier. It procedurally parses data from XML forms, then training of the language model is conducted using which salient words are extracted, finally the last part being the rank assessment of extracted content. Thus, reducing the task of the classifier to extracting results only by

using methods like top-k selection technique [9]. Another approach suggests finding linguistic similarity using the cosine similarity function. The similarity of two fields in this approach is calculated as the weighted average of the similarity between their names, labels, and names versus labels [10]. Also, one of the proposed architectures in this paper consists of three modules: rule-based, dictionary-based, and conditional random fields-based (CRF) extractors. The rule-based extractor works based on regular expressions, whereas the dictionary-based extractor exploits the known entity list approach. The CRF model is employed to extract the undiscovered entities which have not been identified by the other two extractors. This module consists of four feature templates, namely, atomic features template, combination features template, marker features template, and semantic features template [11].

On similar lines, another paper suggests clustering of blocks based on the similarity in their appearances. The apparent similarity is calculated based on the three aspects; images, plain text, and link text. It is a one-pass algorithm. For the derivation of the result, the approach suggested is compared with DEPTA (Data Extraction based on Partial Tree Algorithm). It was observed that DEPTA causes misalignment in the same data record. The difficulty of misalignment is overcome by the proposed algorithm as it can easily distinguish these data items due to the difference in the fonts and positions [12]. One of the approaches suggested the development of a new query selection algorithm. However, the results indicated that the approach suggested proved to be unusable when significant ranked data sources are taken into consideration with a small return limit [13]. Another paper discusses the use of TF-IDF and BOW algorithms for the feature extraction for their classification techniques. Their classification consists of Naive Bayes, Support Vector Machine and Linear Regression. Results were compared for each classification methodology accompanied by one of the feature extraction algorithms [14].

The paper suggests use of Support Vector Machines (SVMs) for classification of content. A training set of 200 entries is used for training and active learning is used for classification. These entries are added to the dataset, and the trainer is rerun. The classifier gives a rough estimate of the distribution of the content but is also slightly biased toward classifying the content as illegal. The various approaches are as well analyzed to ensure that the approach employed is the most efficient. The result indicated that 60% of the content was found to be illegal [15]. Another paper suggests a page classifier module which uses Rainbow, which is a freely available Naive Bayes classifier. It is trained with the samples from the topic taxonomy of the Dmoz Directory. When a page is retrieved, it is assigned a probability P by the classifier, if P is greater than the decided threshold frequency, it is marked as relevant. The threshold probability considered in this paper is 0.5. The number of relevant forms retrieved is the measure for the effectiveness of the form classifier. The results suggest that the crawler implemented in the paper is more effective than the existing ones and it is planned to work on automation of evaluation of the quality of the harvested form in the future [16].

A paper discusses the implementation of the classification of a user channel from YouTube, using the YouTube API, as relevant or irrelevant in which the use of breadth-first search algorithm and shark search algorithm is proposed [17]. However,

a few papers suggested the use of more basic algorithms such as the support vector machines or SVM and Naive Bayes classification algorithm.

The use of a two-dimensional classification methodology designed within the scope of law enforcement that is known as TMM (Tor-use Motivation Model) is discussed in a paper. TMM achieves greater granularity and provides a richer labeling scheme by explicitly distinguishing site content from motivation. The paper demonstrates this robustness and flexibility through direct examples [2]. Another approach that is understudied in text classification is the word2vec algorithm. Still, as the approach is a distinct feature in the text classification method, we discuss it in our work.

The Word2vec algorithm takes every word from the text that is being classified and transforms it into a unique vector. Similar to a space vector, operations like addition, subtraction, and other kinds of manipulation can be done with these vectors. Word2vec incorporates vocabulary into a vector space at a high level. This is the consequence of building particular neural network-based algorithms designed to perform tasks such as auto-completion or detection of probable adjacent words in a document. As the neural network is trained through reading document after document, it learns the interpretation of the vocabulary in a way which can be interpreted in its hidden layer(s) to predict the most feasible missing words; the algorithm learns more about the interconnection that each of the terms in the vocabulary has with each other, based on the frequencies they come together with. These patterns end up getting encoded into a matrix that will be able to map any word in the vocabulary to a vector in much lower-dimensional vector space [18].

It was observed that the approaches mentioned above have their shortcomings and have not provided satisfactory time and space complexity. To overcome the existing issues, we suggest a novel approach for text classification, called the ELEMENT.

Table 1 depicts the use of feature extraction methodologies for the papers in the literature review. After the feature extraction, the extracted features are classified for a better understanding of the data.

The extracted features are a crucial part of classification and clustering methodology. If the features generated are incorrect, there could be a flawed classification of the data, which would further make ones indexing out of order. The feature extraction algorithms are overshadowed by the classification or clustering methodology. The point is the performance of classification and clustering strategies would be significantly affected by the feature extraction methodology. Thus, the use of a proper feature extraction methodology would be an idea for any classification or clustering approach; otherwise, cascading failures of the product would be likely.

3 Methodology

ELEMENT has been developed as a variation of TF-IDF, it has been modified to be able to run more efficiently without losing any accuracy. In this section, ELEMENT will be elaborated depicting the flow and conversion of data within the module itself.

Table 1 Papers and their feature extraction algorithm

Sr. No.	Paper name	TF-IDF	BOW	Word2Vec	Others	Other Algorithms
1	ATOL: A framework for automated analysis and categorization of the Darkweb Ecosystem	1	1	0	1	TF-ICF
2	Focused deep web entrance crawling by form feature classification	1	0	0	0	-
3	Generating queries to crawl hidden web using keyword sampling and random forest classifier	1	0	0	0	-
4	Practical Guides for Data Retrieval in Deep Web Crawling	0	0	0	1	TS-IDS
5	Deep web content monitoring	1	0	0	0	-
6	Vide: A vision-based approach for deep web data extraction	0	0	0	1	ViDE
7	Classifying illegal activities on TOR network based on web textual contents	1	1	0	0	-
8	Deep web classification based on domain feature text	1	0	0	0	W-TFIDF
9	Cybersecurity Named Entity Recognition using Multi-modal Ensemble Learning	0	0	0	1	CRF
10	TODWEB: Training-Less ontology-based deep web source classification	0	1	0	1	BOC
11	An interactive clustering-based approach to integrating source query interfaces on the deep web	0	0	0	1	Similarity Measure

(continued)

Table 1 (continued)

Sr. No.	Paper name	TF-IDF	BOW	Word2Vec	Others	Other Algorithms
12	Method of Deep Web Collection for Mobile Application Store Based on Category Keyword Searching	1	0	0	0	–
13	A crawler architecture for harvesting the clear, social, and dark web for iot-related cyber-threat intelligence	0	0	0	1	Multi-word Expression tokenization
14	E-FFC: an enhanced form-focused crawler for domain-specific deep web databases	1	0	0	0	–
15	Criminal motivation on the dark web: A categorisation model for law enforcement	0	0	0	1	TMM
16	Digital cryptomarkets	0	0	1	0	–

Understanding the methodology allows us to understand why ELEMENT was also chosen to be a part of the evaluation.

Three fundamental steps are used which lead to the creation of tags for a given webpage. The input to the module will be the textual data present on the website. The output to be given is a list of tags along with its frequencies in the text. As stated previously, the tags will give information regarding the website's content. The first fundamental step will be the conversion of all words into their lemmas to avoid the generation of duplicate tags. The next step is to remove stop words from the text as they are pointless in the generation of tags. Finally, the term frequencies are generated for each unique term and will be produced as the output.

3.1 Conversion to Word Lemma

A lemma is a canonical form of a set of words. Primarily, a lemma can be considered a generalized form for its set of words. All the set of words are entered under the dictionary entry of the lemma of those words.

An example for Word Lemma. ‘run’, ‘runs’, ‘ran’, and ‘running’ are all different forms of the word ‘run’. The word ‘run’; it is the lemma for this set of words. Similarly, ‘go’ is the lemma for ‘go’, ‘goes’, ‘going’, ‘went’, and ‘gone’. The lemma can be viewed as the chief of the principal parts.

Different forms of a word are considered as distinct in the calculation of Tag Frequency. For some websites, the most frequent word could even change. Lemmas can be used to get a definitive yet generic word. Hence, it is used as a low-level generalization technique allows conversion of all words into a general stable form. This is the first and the foremost step in ELEMENT's methodology, taking all the textual data and converting all words into their respective lemmas. This changed text with the lemmas will be sent over to the next step.

3.2 Removal of Stop Words

Stop words are the most common words in a given language. These words have no meaning by themselves. Words like 'the', 'a', 'an', 'is', and so on, are stop words. These words are mostly irrelevant as tags and to the tagging methodology.

As stated above, stop words are relatively irrelevant and should be removed. If included, their frequent nature will make them on top of the frequency list. This would lead to the actual content of a website being overshadowed by irrelevant text. Removing these stop words will result in not only better-quality tags but also save space and time hence increasing the accuracy to a great extent.

The stop words are used after the lemmas. As a result, the stop words are also generalized, which makes it easier to eliminate. The output of this step is sent for the tag-frequency calculation. The output at this stage would be generalized as per the lemmas and free of all stop words.

3.3 Calculation of Tag Frequency

Tag Frequency is the measure of the occurrence of a term in the website's textual data. It gives the raw count of the 'term' in the whole text. The frequency for each unique term is calculated. The terms with higher frequency count are the terms which indeed give insight about the website.

Tag frequency gives information related to a website in an economical method. The frequency helps with the search as well as it governs the Page Indexing of a webpage. This page Indexing allows ordering the webpages that are scraped by authors. One can think of Page Indexing as the order in which Google Search results appear. The higher frequency for a term means higher the relevance. It means the webpage will have a higher ranking if the words can be used as a search term as well as for a concise/summarized view of the data of webpage.

In the example, one can see that the most frequent data in the whole set of term-frequency does indeed relate to the actual topic of the data. The Deep web crawler can understand the actual data just by looking at these terms itself.

The Tag frequency is the final step of ELEMENT. In here, some words (terms) are ignored as they would be having next to no value in considering them as tags.

A list of such common words is created which consists of common adjectives like more, less as well as some pronouns like you, they and many other commonly used words which would not contribute anything as tags.

After the algorithm is done, the output will be a tag-frequency list. Each one of the tags will be associated with its own frequency which, as stated before, is the measure of that tag's influence on the webpage.

4 Evaluation

This section evaluates various Text feature extraction methodologies against each other. Given that the use case is unique for each individual an attempt has been made to keep the tests as generalized as possible, something that will be shown later too. The procedure and the results of this evaluation will be seen in this section.

4.1 Algorithms

1. Top-10

This is a Top-K approach, where the 10 most significant features are selected. In the results, the top-10 words will be considered (free of stop words).

2. Top-20

This is also a Top-K approach, where the 20 most significant features are selected.

3. Bag of Words

Bag of Words is a common and simple feature extraction method. In this algorithm, the sentence is considered as a bag containing the words present in it. The semantics relationship of the words within the sentence is ignored and the words and its frequency of occurrence is considered for fetching results.

4. TF-IDF (Term Frequency—Inverse Document Frequency)

TF-IDF is also a common and simple feature extraction method. It involves creation of a sparse matrix where the columns are the sentences in the input, and the rows are words in the sentence. The value for a (row, column) in the sparse matrix of bag of words is the TF-IDF value of the word (row) in the sentence (column). TF-IDF value gives the significance for the word for the text.

5. Word2Vec (A neural network-based approach)

This algorithm takes input data and produces a vector space as output. Word2Vec is a simplistic, two-layer neural network trained to revamp words' linguistic contexts that locates similar words close to one another. These vectors are stored as vocabulary for the model. The associated names of the features are used as the features extracted from the data.

4.2 Dataset

Next part will be regarding the chosen dataset and purpose for choosing the same. Authors have chosen the dataset ‘*DBpedia Classes—Hierarchical Taxonomy of Wikipedia article classes*’ from <https://www.kaggle.com/danofer/dbpedia-classes>.

The database DBpedia is a project to extract structured content from the information in Wikipedia. It is the data (after cleaning, kernel included) that provides taxonomic, hierarchical categories (“classes”) for 342,782 Wikipedia articles.

DBpedia dataset was selected since it provides a list of websites along with a list of tags for each website. So, all the algorithms can be tested on the website data and compared among themselves. The length of actual text content ranges from as low as 2 words and climbs up to more than 700 words. This wide variety of word count per entry will ensure proper working of the module.

Use of a Web Dataset like DBpedia, which has Wikipedia data of various domains, ensured proper testing and provided quality results. Due to the nature of the Web, the webpages themselves are structured by the use of HTML tags. When going from one website to another, be it Surface Web or Deep Web, the only thing that would change will just be the textual content as every other aspect of the webpage is bound to the structured representation of HTML. Thus, the use of such a diversified web dataset would likewise ensure the algorithm’s performance on the Web.

All items in the dataset were taken to be analyzed by ELEMENT. The formula for the calculation of accuracy is given below using an example:

$$\text{Total Number of Dataset Tags} = 3(1994, \text{ Mindoro}, \text{ Earthquake})$$

$$\text{Tags given by the ELEMENT} = 3(1994, \text{ Mindoro}, \text{ Earthquake})$$

$$\text{Accuracy per website} = 3/3 \times 100 = 100\%$$

Hence, ELEMENT is 100% accurate for this example. Similarly, this process was carried out for all the algorithms for each sampled item.

4.3 Comparative Analysis

Several comparisons are made among the methodologies. Each comparison criteria depicts a different aspect of the methodology. Authors chose to implement these algorithms using pre-built open-source libraries to ensure that they had the best performance. TF-IDF, Bag of Words were used from the sklearn library for Python. Word2Vec was from the gensim library for Python. The comparison was conducted on approximately 340,000 unique entries in the Dataset. All tests were conducted on the same device to minimize hardware factors as much as possible.

Tests have been carried out based upon 6 factors that hugely define and matter when it comes to Text feature Extraction. Each of these factors will be discussed and the results for all algorithms considered shown. Following are the factors:

1. Accuracy Measure

The Accuracy of each methodology is compared here, this is an important factor as it allows in determining if an algorithm has the best accuracy or not. Due to the lower accuracy values of Top 10 and Top 20, these methodologies are not considered in further comparisons.

2. Features Extracted

This factor is the count of the number of features extracted. This is important as the length of data increases, the features extracted by the methodologies should also increase. But it should not be such that useless features are also taken into account.

3. Space Complexity

This is the total amount of space taken while storing the data. For the calculation of space, the `__sizeof__()` method of python was used. It was recursively iterated for each part of the methodology. This gave an estimation of the size occupied by the methodology during the time of its execution.

4. Space Efficiency

This is measured to find out how efficiently the space is managed by each methodology. It is also estimated that how many resources (in terms of space) are utilized to calculate ONE feature.

5. Time Complexity

The execution time is calculated using the python function `time()`. Start time was recorded just before the function of the methodology was called. After the results are received from the function, the end time was recorded. This raw time difference is supposed to be the time required to execute the methodology.

6. Time Effectiveness

This is measured to find out how effectively time is consumed by each one of the methodologies. It is also used to estimate how much resource (in terms of time) is consumed to calculate ONE feature.

Tables 2 and 3 illustrate the same values. Accuracy was the most important factor during the evaluation, and since Top 10 and Top 20 both showed a lower level of accuracy, they were not considered for the further tests. Table 3 shows how each of the other factors weighs in for the different methodologies taken into account.

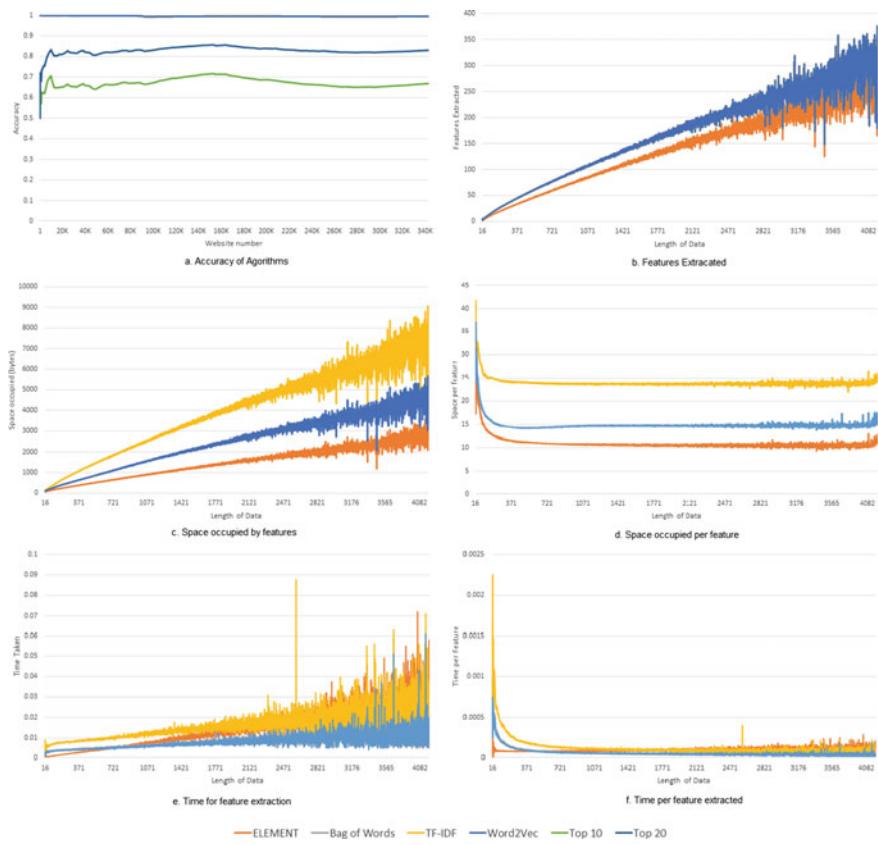
Figure 1 is a composition of all of these values in a graphical representation for a better understanding. It compiles all the 6 comparison factors that are mentioned

Table 2 Accuracy of each method

Methodology	ELEMENT	Top—10	Top—20	Bag of Words	TF-IDF	Word2Vec
Accuracy	0.994923	0.672234	0.828717	0.994923	0.994923	0.994923

Table 3 Feature extraction comparison

Methodology	ELEMENT	Bag of Words	TF-IDF	Word2Vec
Average number of Features extracted	51.0307309	65.3244899	65.3244899	65.3244899
Average space (in bytes)	563.771235	952.861205	1561.32519	952.861205
Average space per feature (in bytes)	11.7758695	14.9366035	24.2971151	14.9366035
Average time (in seconds)	0.00444135	0.00495388	0.00961431	0.00469996
Average time per feature (in milli-seconds)	0.00842903	0.01061370	0.02062940	0.00999023

**Fig. 1** Graphical representation of comparisons

earlier in the section. Part A shows the comparison of the accuracy of all methodologies considered, the moving average of the accuracy on the X-axis and webpage number on the Y-axis. As observed, the accuracy of ELEMENT, Bag of Words, TF-IDF, and Word2Vec, all these have nearly the same accuracy (~100%) and their lines overlap.

Part B shows the count of features extracted per webpage for each methodology. Here, Bag of Words, TF-IDF, and Word2Vec, all have an equal number of features extracted and overlap, ELEMENT shows a lower value here due to the removal of stop words in a more robust manner.

Part C is the comparison of space complexity of all methodologies having Length of Data on X-axis and Space Occupied on Y-axis. For the calculation of space, the `__sizeof__()` method of python was used and it was observed that Bag of Words and Word2Vec have a very similar space consumption.

The space occupied per feature for all the algorithms is represented in Part D of the figure. This is measured to find out how many resources (in terms of space) are utilized to calculate ONE feature. Similar to Part C, the values for Bag of Words and Word2Vec are very identical.

Part E is the time consumed for feature extraction by each methodology. On an average, the time consumed by ELEMENT, Bag of Words, and Word2Vec is similar and differs only by a factor of 0.0001 s. But, for consistent proper performance, Bag of Words, or Word2Vec would be ideal.

Finally, Part F demonstrates the time consumed per feature by each of the methodologies. This helps estimate how much time is consumed to calculate ONE feature. Again, Bag of Words, or Word2Vec would be ideal to use for a situation. Conclusions based on all this data have been elaborated in the next section.

4.4 Concluding Comparisons

Table 4 presents the color coding, and its meaning and Table 5 presents the grading to the methodologies using the colors. Evaluation of each of the methodologies is relative to the performance of other methodologies for the same evaluation criteria.

The above tables depict the accuracy and the evaluation of the methodologies respectively. One can clearly find that the methodology suggested by authors does farewell in most aspects of the comparisons.

Table 4 Color coding

Color				
Evaluation	Best	Great	Good	Average
				Below average

Table 5 Factors evaluation grading

	ELEMENT	Top - 10	Top - 20	Bag of Words	TF-IDF	Word2Vec
Accuracy						
Features		-	-			
Space		-	-			
Space/Feature		-	-			
Time		-	-			
Time/Feature		-	-			

In summary, ELEMENT gives better quality tags, consumes less space, and takes less time on an average, all while having the same accuracy as the popular methodologies of text feature extraction. These added benefits of superior complexities enable ELEMENT to be really useful in the case where the task is to extract textual features from a vast corpus of text.

5 Conclusion

This paper presents a basic yet constructive methodology for text mining. Use of Lemmas and elimination of stop words enables ELEMENT to get better accuracy of relevant tags. The concept of lemmas was used as a generalization method, whereas elimination of stop words as a cleaning and enriching approach. During the generation of tags, term-frequency pairs are ensured not to have some common words (common pronouns, adjectives, verbs). Considering, the nature of the text, all terms in the final iteration, regardless of the frequency associated with it are considered since all the tags will be useful. All these steps were crucial to getting such high accuracy. Although the methodology of the ELEMENT is useful, it is primitive, and much work can be done. The tag-frequency pair can be used as pre-processing for several tasks like classification or clustering as well as to summarize the textual data.

For the evaluation accuracy, each one of the methodologies considered in comparison is exceptionally significant. However, beyond the accuracy of an algorithm, the resource consumption, effectiveness and efficiency, all play a crucial role, especially if taking into account the massive amount of data in today's world. The methodology discussed in the paper, might not be perfect, yet is competitive and in a majority of the cases better than the commonly used approaches. Feature extraction by ELEMENT is a little low, but of better quality. And ELEMENT triumphs in terms of space complexity and space efficiency too. When it comes to being the

quickest, Word2Vec, and Bag of Words leave their mark, being the better options. Each of these methodologies is extremely competitive and with some fine tuning can easily fit into a particular use case.

References

1. Li, S., Chen, C., Luo, K., Song, B.: Review of deep web data extraction. In: 2019 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1068–1070. IEEE (2019)
2. Dalins, J., Wilson, C., Carman, M.: Criminal motivation on the dark web: a categorisation model for law enforcement. *Digital Investigation* **24**, 62–71 (2018)
3. Ghosh, S., Porras, P., Yegneswaran, V., Nitz, K., Das, A.: ATOL: A framework for automated analysis and categorization of the Darkweb Ecosystem. In: Workshops at the Thirty-First AAAI Conference on Artificial Intelligence (2017)
4. Noor, U., Rashid, Z., Rauf, A.: TODWEB: Training-Less ontology based deep web source classification. In: Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services, pp. 190–197 (2011)
5. Wang, L., Hawbani, A., Wang, X.: Focused deep web entrance crawling by form feature classification. In: International Conference on Big Data Computing and Communications, pp. 79–87. Springer, Cham (2015)
6. Khelghati, M.: Deep Web Content Monitoring. University of Twente (2016)
7. Xu, G., Wu, Z., Li, C., Yan, J., Yuan, J., Wang, Z., Wang, L.: Method of deep web collection for mobile application store based on category keyword searching. In: International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage, pp. 325–335. Springer, Cham (2019)
8. Wu, C., Qiang, B., Zou, X.: Deep web classification based on domain feature text. *Int. J. Adv. Comput. Technol.* **3**(6) (2011)
9. Koloveas, P., Chantzios, T., Tryfonopoulos, C., Skiadopoulos, S.: A crawler architecture for harvesting the clear, social, and dark web for IoT-related cyber-threat intelligence. In: 2019 IEEE World Congress on Services (SERVICES), Vol. 2642, pp. 3–8. IEEE (2019)
10. Wu, W., Yu, C., Doan, A., Meng, W.: An interactive clustering-based approach to integrating source query interfaces on the deep web. In: Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, pp. 95–106 (2004)
11. Yi, F., Jiang, B., Wang, L., Wu, J.: Cybersecurity Named Entity Recognition using Multi-modal Ensemble Learning. *IEEE Access* (2020)
12. Liu, W., Meng, X., Meng, W.: Vide: A vision-based approach for deep web data extraction. *IEEE Trans. Knowl. Data Eng.* **22**(3), 447–460 (2009)
13. Lu, J., Wang, Y.: Challenges in crawling the deep web. In: Big Data, pp. 105–128. Auerbach Publications (2016)
14. AlNabki, M.W., Fidalgo, E., Alegre, E., de Paz, I.: Classifying illegal activities on TOR network based on web textual contents. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 35–43 (2017)
15. Avarikioti, G., Brunner, R., Kiayias, A., Wattenhofer, R., Zindros, D.: Structure and content of the visible Darknet (2018). arXiv preprint [arXiv:1811.01348](https://arxiv.org/abs/1811.01348)
16. Barbosa, L., Freire, J.: Searching for hidden-web databases. In: WebDB, Vol. 5, pp. 1–6 (2005)
17. Agarwal, S., Sureka, A.: Topic-specific YouTube crawling to detect online radicalization. In: International Workshop on Databases in Networked Information Systems, pp. 133–151. Springer, Cham (2015)
18. Agrawala, T.S., Kumaraguru, P.: Digital cryptomarkets (2017)

Anomalies in Punjabi Language WordNet: An IndoWordNet Evaluation



Harjit Singh

Abstract The IndoWordNet is a multilingual WordNet for eighteen Indian languages. Punjabi language WordNet is a part of the IndoWordNet. The efforts made for the development of Punjabi language WordNet are undoubtedly appreciable, but various anomalies exist in this language resource. These anomalies were detected in Punjabi language WordNet, while using it for text processing. This paper presents the results of the evaluation, discusses the possible causes behind identified issues, and recommends necessary modifications. In this evaluation, the anomalies were identified by analyzing the IndoWordNet text files manually with the help of various word processors. To verify the results, these anomalies were also tested with online versions of the IndoWordNet available at www.cfilt.iitb.ac.in/indowordnet/ and www.tdil-dc.in/indowordnet/. Although, these web applications behave differently for similar inputs, but the existence of anomalies was verified by these online versions. The IndoWordNet is a structured language resource, but these anomalies violate the rules of its structure. It was tested for the purpose of improving its data accuracy.

Keywords NLP resource evaluation · Language resource evaluation · IndoWordNet · Punjabi WordNet · Natural language processing · Punjabi language resource

The original version of this chapter was revised: The incorrect Gurmukhi texts have now been updated. The correction to this chapter is available at https://doi.org/10.1007/978-981-16-2164-2_50

H. Singh (✉)
APS Neighbourhood Campus, Punjabi University, Patiala, India
e-mail: hj�t@live.com

1 Introduction

WordNets are very useful resources for research and development in Natural Language Processing (NLP) [1]. These are freely accessible rich resources of language vocabulary which are fairly accurate. These resources have been used for word sense disambiguation [2], information retrieval [3], Question Answering system [4], and many other NLP tasks. The first WordNet was developed for English language [5], named Princeton WordNet [6].

The EuroWordNet is a multilingual WordNet developed for twelve European languages [7]. IndoWordNet is also a multilingual WordNet developed for eighteen Indian languages [8]. The root of the IndoWordNet is Hindi language WordNet [9]. The WordNets for other Indian languages were generated by using expansion approach. The Punjabi WordNet is a part of IndoWordNet and was created from Hindi language WordNet using expansion approach [10]. Currently, the Punjabi WordNet, downloaded as a part of IndoWordNet, contains 36534 Synsets having 52793 words.

IndoWordNet is not only a research project; it represents the presence of Indian languages at global level. It is also linked with the Princeton WordNet which marks its importance [11]. It is supported by various agencies and ministries of Govt. of India [12]. This multilingual resource is a milestone for NLP research in India. Findings suggest that corrections should be made to achieve more accuracy in this language resource.

1.1 *Punjabi Language*

Punjabi is a language spoken by the people of Punjab state of India and related Punjabi community in other countries such as Pakistan, Canada, UK, US, and UAE [3, 4]. The total Punjabi speaking population counts to more than 100 million worldwide [5]. Punjabi community uses two types of scripts to write that are Gurmukhi and Shahmukhi [6]. This paper deals with Gurmukhi scripted Punjabi language.

Punjabi language belongs to Indo-Aryan language family which has its roots in Indian sub-continent. It is spoken by more than 100 million people across the world and is tenth most spoken language in the world [13, 14]. It is a native language of people of Punjab region, in India and Pakistan. Punjabi is eleventh most spoken language in India and is written in Gurmukhi script. In Pakistan, it is written in Shahmukhi script and is the most spoken language. In IndoWordNet, Gurmukhi script is used to generate Punjabi language WordNet. There are huge number of Punjabi immigrants in Canada, UK, US, Australia, and many other countries.

Punjabi is a syllabic language having 41 consonants, 9 vowel symbols, 2 nasal sound symbols, [15] and one double sound symbol. The Punjabi letters with their English sound equivalents are listed below:

Consonants:

ਚ (*)	ਐ (a)	ਏ (*)	ਐ (s)	ਯ (h)	
ਕ (k)	ਖ (kh)	ਗ (g)	ਘ (gh)	ਙ (ng)	
ਚ (ch)	ਛ (chh)	ਜ (j)	ਯ (jh)	ਝ (nj)	
ਟ (t)	ਠ (th)	ਡ (d)	ਢ (dh)	ਣ (n)	(1)
ਤ (t)	ਥ (th)	ਦ (d)	ਧ (dh)	ਨ (n)	
ਪ (p)	ਫ (f)	ਬ (b)	ਭ (bh)	ਮ (m)	
ਯ (y)	ਰ (r)	ਲ (l)	ਵ (v)	ੳ (r)	
ਸ (sh)	ਖ (kh)	ਗ (gh)	ਝ (z)	ਙ (ph)	ਝ

*These consonants are not used without vowel and acquire the sound of attached vowel.

Vowels are not used independently, rather are used with a consonant:

ਾਂ (ਾ)	ਿਂ (ਿ)	ਿਂ (ਿ)	ੋ (ੋ)	ੁਂ (ੁ)
ੇ (ੇ)	ੈ (ਾਈ)	ੈ (ਾਈ)	ੋ (ੋ)	ੁਂ (ੁ)

Nasal sound symbols are:

ਾਂ (ਾਨੀ)	ੈਂ (ਾਨੀ)
----------	----------

Double sound symbol:

ੂ

1.2 Synset Structure in IndoWordNet

In IndoWordNet data is stored in the form of Synsets. A Synset is a collection of synonymous words and is stored in a structure with fields ID, CAT, CONCEPT, EXAMPLE, and SYNSET-PUNJABI. Following example shows a Synset structure with ID 23:

ID :: 23

CAT :: ADJECTIVE

CONCEPT :: ਜੋ ਯੋਗ ਨਾ ਹੋਵੇ

EXAMPLE :: "ਪ੍ਰਬੰਧਕਾਂ ਨੇ ਅਯੋਗ ਵਿਅਕਤੀਆਂ ਨੂੰ ਸੰਸਥਾ ਵਿਚੋਂ ਕੱਚ ਦਿੱਤਾ"

SYNSET-PUNJABI :: ਅਯੋਗ, ਅਸਮਰਤੱਥ, ਨਾ_ਕਾਬਲ, ਅਨਾੜੀ, ਨਲਾਈਕ, ਨਲੈਕ

For reference, the CONCEPT and EXAMPLE are translated and SYNSET-PUNJABI is transliterated as follows:

ID :: 23

CAT :: ADJECTIVE

CONCEPT :: Someone who is not eligible

EXAMPLE :: “Organizers expelled **ayੋg** persons from the organization”
 SYNSET-PUNJABI :: ayੋg, asamaratha, nā_kabil, anārī, nala’ik, nalaik

ID is a unique identifier which starts with one and is incremented by one for every next Synset. CAT is the grammatical category of the words included in the Synset. CONCEPT provides short description or meaning of those words. EXAMPLE shows the usage of first word of Synset in a sentence. SYNSET-PUNJABI has the actual data in the form of all synonyms of the first word in the set.

2 Anomalies in Punjabi Language WordNet

Very limited number of digital resources are available for Punjabi language text processing. In this situation, the efforts made to develop the IndoWordNet are undoubtedly appreciable, but still there are various anomalies to be removed to make this rich language resource more accurate and useful. The rest of this section discusses the detected anomalies one by one.

2.1 *Prefixed Double Quotation Marks*

While testing the Synsets stored in Punjabi language WordNet, it was found that there are eleven words which are mistakenly prefixed with double quotation marks. There are no double quotation marks at the end of these words or at the end of these Synsets. These words are listed in Table 1 with their corresponding Synset IDs for reference. The double quotation marks appear as first character of these words which arises problems while using these words for text processing.

2.2 *Underscore Versus Space Versus Hyphen*

A collocation is a group of multiple words that are frequently used together. The Princeton WordNet documentation states that “a collocation is entered by joining the individual words with an underscore character” [16]. Although no specific documentation is found regarding the entry of multiword phrases in the IndoWordNet, but in a document related to “National Workshop on WordNet Creation” [17], it is stated that “while introducing such multi-words the lexicographer should separate component words by an underscore”.

There are 3361 collocations in Punjabi language WordNet that have been entered with space between individual words. Along with that there have been 12936 collocations entered with underscore character between individual words. There also exist 2300 collocations where hyphen has been used as word separator.

Table 1 List of words prefixed with double quotes

Sr. No.	Synset ID	Word as appear in Synset (Gurmukhi Script)	Transliteration (For Reference)
1.	547	"ਛੱਡਣੇ_ਯੋਗ	"Chadana_yoga
2.	1430	"ਵਿਵਹਾਰਕ	"vivahāraka
3.	7168	"ਬਾਰੂੰਵੀ	"bārū̄vī
4.	10985	"ਸੁਭਾਸ_ਚੰਦਰ_ਬੋਸ	"subhāśa_chandar_bōs
5.	11084	"ਜਦੋ_ਤੱਕ	"jadōṁ_tak
6.	28712	"ਨਿਸਚਲਾ_ਨਦੀ	"niśachalā_nadī
7.	28891	"ਚਾਕਸੂਕ	"cākaśuk
8.	30931	"ਅਰਯਮਾ	"arayamā
9.	32180	"ਕੱਲ ਲਈ ਕੁਝ ਨਾ ਰੱਖਣ ਵਾਲਾ	"kal'h laṭ kujh nā rakhan vālā
10.	32865	"ਆਤਮ ਦ੍ਰੋਹੀ	"ātam drōhī
11.	34473	"ਗੈਰ ਜ਼ਿਮੇਵਾਰ	"Gair zimēvār

While testing the Punjabi WordNet, long lists of all these words were generated. The variety of ways used to enter collocations has resulted in non-uniformity of data.

2.3 Question Marks

The evaluation of Punjabi language WordNet shows fourteen words suffixed with one or more question marks. Table 2 shows the list of these words with their corresponding Synset IDs. These question marks might not have been entered intentionally rather it appears that these might be the results of some translation procedure adopted in expansion approach. Anyway, these question marks provide a “raw text” appearance to such a structured language resource.

2.4 Grave Accent

In Punjabi language WordNet, a word at Synset ID 5887 was found having grave accent (‘) separated with a white space. The reason might be an unintentional pressing of the key while entering the word and due to its tiny size, it remained un-noticed. The same word in correct form exists at Synset ID 6189 resulting in redundancy. Table 3 shows both entries of the word.

Table 2 List of words suffixed with question mark(s)

Sr. No.	Synset ID	Word as appear in Synset (Gurmukhi Script)	For Reference
1.	2094	ਲੰਮ੍ਹੇ_ਵੇਲੇ_?	Langhē_vēlē_?
2.	4108	ਹੱਥ?????????	Hath????????
3.	5327	ਲੰਬੀ ਉਮਰ ਦੀ ਕਾਮਨਾ?	Lambī umar dī kāmanā?
4.	5930	ਵਣਮਨੁੱਖ_?	Vaṇamanukh_?
5.	6709	ਜਾਬਰ_??	Jābar_??
6.	7719	ਅਮ੍ਰਿਤ_ਵੇਲੇ?????????????	Amrit_vēlē?????????????
7.	26257	ਲੇਖਾ-ਜੋਖਾਵਿਭਾਗ?	Lēkhā-jōkhāvibhāg?
8.	26309	ਅਣਬੱਕਸੇਵਾ?	Anathakasēvā?
9.	26319	ਚੇਣਾ?	Chōṇā?
10.	27122	ਮੈਬਰ?	Maimbar?
11.	27231	ਨਖਵਿਸ਼ੁਕਿਰ_?	Nakhaviśrakir_?
12.	27337	ਸ੍ਰੇਣੀਕਰਨ?	Śrenīkaran?
13.	28330	ਘੁਮਣਾ_??	Ghumāṇā_??
14.	29712	ਸੰਯਕਤਰਾਸਟਰ_ਬਾਲ_ਨਿਧਿ?	Sanyukatarāśatār_bāl_nidhi?

Table 3 The word with and without accent grave

Sr. No.	Synset ID	Word as appear in Synset (Gurmukhi Script)	For Reference
1.	5887	ਜੌਮ̂ `	Jaum̂ `
2.	6189	ਜੌਮ̄	jauṁ

2.5 Abbreviations

The abbreviations have been entered with dots, but their format is not uniform. Various ways have been used to enter abbreviations. Some abbreviations have been entered with a dot after every letter while in others last dot has been missed. Synset ID 3586 in Table 4 shows the related word. Some abbreviations have an underscore between letters, while others do not. Synset ID 6357 in Table 4 shows the related word. At the time of the evaluation, some points were considered to determine the uniformity of abbreviations. An abbreviation was considered as correct, if it:

Table 4 Abbreviations that miss uniformity

Sr. No.	Synset ID	Word as appear in Synset (Gurmukhi)	For Reference (English)	Non-Uniformity
1.	3586	ਧੂ.ਐਫ	U.F	Last dot missing
2.	4976	ਐਲ.ਪੀ.ਜੀ	L.P.G	Last dot missing
3.	6213	ਟੀ.ਵੀ	T.V	Last dot missing
4.	6270	ਆਈ.ਟੀ	I.T	Last dot missing
5.	6357	ਐਚ.ਪੀ.	H. P	Underscore after dot
6.	10637	ਪੀ.ਏ	P.A	Last dot missing
7.	20713	ਯੂ._ਐਸ_ਡਾਲਰ	U._S_Dollar	Second and last dot missing, underscore after dot
8.	21096	ਯੂ._ਐਸ_ਆਰ	U._S_R	All dot missing, underscores used in between, misplaced dot
9.	22579	ਸੀ.ਐਮ	C.M	Last dot missing
10.	26878	ਡਬਲਿਊ.ਐਚ.ਓ	W.H.O	Last dot missing
11.	27202	ਪੀ.ਸੀ.ਐਮ	P.C.M	Last dot missing
12.	27401	ਸੀ.ਪੀ.ਸਨੋ	C.P.Snow	Underscore missing before word
13.	27402	ਆਰ.ਡਬਲਿਊ ਸਰਵਿਸ	R.W Service	Last dot missing, underscore missing before word
14.	27429	ਐਮ.ਆਰਕ	M.Arch	Last dot missing
15.	28906	ਆਰ.ਏਨ.ਏ	R.N.A	Last dot missing
16.	28907	ਡੀ.ਏਨ.ਏ	D.N.A	Last dot missing
17.	28907	ਡੀ.ਐਨ.ਏ	D.N.A	Last dot missing
18.	34587	ਡਬਲਿਊ._ਡਬਲਿਊ._ਡਬਲਿਊ	W.W.W	Last dot missing, underscore used in between
19.	34656	ਸੀ._ਬੀ._ਆਈ	C._B._I	Last dot missing, underscore used in between
20.	34717	ਏ._ਟੀ._ਐਮ	A._T._M	Last dot missing, underscore used in between
21.	34868	ਆਈ._ਸੀ.ਸੀ.ਆਈ	I._C.C.I	Last dot missing, underscore used in between
22.	35025	ਐਮ._ਡੀ	M._D.	First and last dot missing, underscore used in between, misplaced dot
23.	35321	ਆਰ._ਟੀ._ਆਈ	R._T._I	Second and last dot missing, underscore used in between
24.	35395	ਐਸ._ਪੀ	S._P	Last dot missing, underscore used in between
25.	35814	ਪੀ.ਐਚ.ਡੀ	P.H.D	Last dot missing
26.	36292	ਐਲ.ਸੀ.ਡੀ	L.C.D	Last dot missing
27.	36318	ਪੀ.ਐਮ.ਓ	P.M.O	Last dot missing

- a) has a dot after every abbreviated letter, e.g. ਐਲ.ਪੀ.ਜੀ. (L.P.G.) is correct.
- b) has a dot after the last letter, e.g. ਐਲ.ਪੀ.ਜੀ (L.P.G) is incorrect.
- c) does not have spaces in between, e.g. ਐਲ. ਪੀ. ਜੀ (L. P. G.) is incorrect.
- d) does not have underscore between abbreviated letters, e.g. ਐਲ._ਪੀ._ਜੀ. (L._P._G.) is incorrect.
- e) has underscore between abbreviated letter and the complete word, e.g. ਐਲ.ਪੀ._ਗੈਸ (L.P._Gas) is correct.

Considering the above points to maintain uniformity, there are eleven abbreviations which follow these points, but 27 abbreviations do not. Table 4 shows the list of abbreviations that miss uniformity.

2.6 Dot Characters

There are twenty words, excluding abbreviations, in Punjabi language WordNet where the dot character has either been misplaced or used incorrectly in place of a comma. It not only disturbs the format of this structured language resource but also concatenates two or more words together. Table 5 shows all the Synset IDs having dot used incorrectly.

2.7 Curly Braces

One Synset was found having curly braces used to enclose a part of Synset which makes the enclosed word inaccessible. Table 6 shows the Synset having curly braces used to separate its subset of words. But “set” should always be a set of elements separated with a specific separator, such as comma. As per the structure of IndoWordNet that has been followed for all other Synsets, this subset can be stored as a separate Synset and then can be linked to the related Synset ID.

The word numbers 4 and 5 shown in the comment column have been affected by the use of curly braces.

2.8 Parentheses

While evaluating the Punjabi WordNet, five Synsets were found having parentheses used to describe the word. These words appear like multiword expressions or collocations having words separated with space or underscore. As already discussed, there are 3361 collocations where spaces have been used to separate words.

Table 5 Synset IDs having incorrect use of dot

Sr. No.	Synset ID	Word as appear in Synset (Gurmukhi)	For Reference	Dot Used in Place of:
1.	1736	ਨੀਲਮ.ਨੀਲ_ਮਣੀ	Nilam.Nil_maṇī	Comma
2.	3255	ਅਸਮਰਥ.ਸਮਤਾਹਿਣ	Asamarath.Śamatāhīṇ	Comma
3.	5232	ਦੀਪਕ.ਬੜੀ	Dīpak.Batī	Comma
4.	6624	ਰੇਸ਼ਮੀ_ਕੱਪੜੇ.ਰੇਸ਼ਮੀ_ਬਸਤਰ	Rēśamī_kapāṭe.Rēśamī_basatar	Comma
5.	6713	ਪਰਸਪਰਿਕ.	Parasaparik.	Misplaced dot
6.	7167	ਬਰਫ.ਬਰਫ	Baraph.Baraf	Comma
7.	7604	.ਮਿਣਨਾ	.Miṇanā	Misplaced dot
8.	9632	ਪ੍ਰਬੰਧਕ.	Prabandhak.	Misplaced dot
9.	10735	ਨੈਥਣਾ.ਨੈਥਣਾ	Nathaṇā.Nathanā	Comma
10.	13133	ਸਹੀ_ਸਲਾਮਤ.ਦਹੁਸਤ	Sahī_salāmat.Darusat	Comma
11.	13348	ਜੜ.ਮੂਲ	Jar.Mūl	Comma
12.	15728	₹2.67	67.67	Comma
13.	22701	ਸਰਭ_ਉਪਨਿਸਦ.ਸਰਭ	Śarabh_upaniśad.Śarabh	Comma
14.	23070	ਐਖ.ਕਠਿਨ	Aukhē.Kaṭhin	Comma
15.	25812	ਫਿਰੋਜਪੁਰ_ਜ਼ਿਲ੍ਹਾ.ਫਿਰੋਜਪੁਰ_ਜ਼ਿਲ੍ਹਾ	Phirōjapur_zil'hā.Phirōzapur_zilā	Comma
16.	26024	ਹਾਜੀਪੁਰ.ਹਾਜੀਪੁਰ_ਸਹਿਰ	Hājīpur.Hājīpur_śahir	Comma
17.	26387	ਵਾਲਯਥਿਲਾਅ.ਵਾਲਯਥਿਲਾਅ_ਹਿਜੀ	Vālayakhila'a.Vālayakhila'a_rīśī	Comma
18.	27401	ਸੀ.ਪੀ.ਸਨੌ.ਚਾਰਲਸ ਪਰਸੀ ਸਨੌ	Sī.Pī.Sanō.Chāralas Parasī Sanō	Comma (last dot)
19.	35602	ਗੀਗਾਵਾਈ.ਗੀਗਾਵਾਈ	Gīgābā'īt.Gīgābāt	Comma
20.	35955	ਪ੍ਰਸਾਂ-ਪੱਤਰ.ਸਲਾਘਾ_ਪੱਤਰ	Praśānsā-patar.Śalāghā_patar	Comma

Table 6 Words enclosed in curly braces

Sr. No.	Synset ID	SYNSET-PUNJABI	For Reference	Comment
1.	7468	:: ਭੇਜਨ, ਖਾਣਾ, ਅੰਨ_ਗੁਹਿਣ, {ਪ੍ਰਸਾਦਾ_ਛੱਕਣਾ, ਲੱਗਰ_ਛੱਕਣਾ_(ਪਾਰਮਿਕ_ਖੇਤਰ_ਵਿਚ()	:: Bhōjana, khāṇā, ana_grahiṇā, {praśādā_chakaṇā, lagara_chakaṇā_(dhāramika_khētar_a_vica)}	Following the format of all other Synsets, this Synset has following five words: 1.ਭੇਜਨ 2.ਖਾਣਾ 3.ਅੰਨ_ਗੁਹਿਣ 4.ਪ੍ਰਸਾਦਾ_ਛੱਕਣਾ 5.ਲੱਗਰ_ਛੱਕਣਾ_(ਪਾਰਮਿਕ_ਖੇਤਰ_ਵਿਚ{})

Table 7 Words having parentheses used for description

Sr. No.	Synset ID	Word as appear in Synset (Gurmukhi Script)	For Reference
1.	4315	ਸੁਲੀ (ਘੜੀ ਵਾਲੀ)	Sūlī (gharī vālī)
2.	4322	ਕੁਣੀ (ਮੱਡੀ ਫਰਨ ਵਾਲੀ ਕੁਣੀ)	kundī (machī phaṇā vālī kunḍī)
3.	6284	ਵਿਵਸਥਾ(ਪ੍ਰਬੰਧ)	vivasathā(prabandh)
4.	7468	ਲੰਗਰ_ਛੱਕਣ_(ਪਾਰਮਿਕ_ਖੇਤਰ_ਵਿਚ)}	langar_chhakanā_(dhāramik_khētar_vich)}
5.	8958	ਦਾਲ ਕੇ ਬਾਈਆਂ ਗਾਰੀਆਂ (ਦਲਵਾਂ)	ḍhāl kē baīāਂ Tām gaTām (ḍhalavāṁ)

Table 8 Synsets having comments in English

Sr. No.	Synset ID	SYNSET-PUNJABI	For Reference
1.	34372	:: ਵਾਧੂ, ਅਤਿਰਿਕਤ* Pan India Synsets *	Vādhū, atirikata* Pan India Synsets*
2.	34374	:: ਬਿਨਾਂ, ਇਲਾਵਾ, ਸਿਵਾਏ, ਸਿਵਾ* Adverb record Synsets *	Bināṁ, ilāvā, sivā'ē, sivā *Adverb record Synsets*

Table 7 shows the Synsets having parentheses used to describe the word. The explanation, if required can be given in the “CONCEPT” field of the Synset instead of the SYNSET-PUNJABI field. If the word is a special word, then it can be stored as a separate Synset and then can be linked to the related Synset ID.

2.9 Misplaced Comments

Two Synsets were detected having attached comments enclosed in asterisk characters. Table 8 shows the Synsets with which the comments have been written in English language. It appears that these have not been written by mistake, instead entered intentionally. The suitable field for comments is CONCEPT field instead of SYNSET-PUNJABI field.

3 Testing and Results

After finding above anomalies in Punjabi WordNet directly from IndoWordNet files, the issues were also tested using online version of IndoWordNet available at www.cfilt.iitb.ac.in/indowordnet/ and www.tdl-dc.in/indowordnet/. These tests prove that the issues discussed above really exist and are making those affected words inaccessible or unusable. Both these web applications behave differently for

same set of inputs. Online access to IndoWordNet provided at www.cfilt.iitb.ac.in/indowordnet/ does not allow any special character in the input word. The “search” button does not work if the input word contains any special character. The search does not work even if a hyphen, an underscore character, or a dot character is used in the input word. The analysis shows that there are 38 abbreviations containing dot character and 12936 collocations containing underscore character and 2300 collocations containing hyphen. Obviously it was tried to search collocations by replacing the underscore with a white space, but the web application returns “Word _____ not found in WordNet” message. Same happens for hyphenated collocations. The search is successful only if the exact white spaced collocation is present in the Punjabi WordNet. It means that a total of $12936 + 2300 = 15236$ collocation and 38 abbreviations are inaccessible from this web location. Online access to IndoWordNet provided at www.tdil-dc.in/indowordnet/ allows any special character in the input word. But here search is successful if an exact match is found, i.e., we have to input the collocation with underscore or hyphen if it is stored like that. Surprisingly, it does not search for a collocation that contains white spaces; it returns the same word as suggestion to input. Table 9 shows the response when some collocation is entered with a space.

It is clear that API developed to find the words is unable to match the input with those words of Punjabi WordNet which are not following a uniform rule. For example, either all collocations should have underscore or all of them should have white space as word separator. There are a number of special characters which are embedded in words mistakenly.

Table 9 provides detailed list of all these issues when tested with online access to Punjabi WordNet provided through IndoWordNet available at two different web locations. The same input word related to an issue is entered with variations to test it for various input possibilities. The outputs shown are the exact “phrases” responded by the web applications for each possible input word.

There are 52793 words in Punjabi language WordNet out of which 18691 are found anomalous. Using this count the inaccuracy percentage of Punjabi language WordNet is calculated as shown in Table 10.

Table 9 List of issues tested with online access

Synset ID::Synset Sources: a) http://www.cfilt.iitb.ac.in/indowordnet/ b) http://tdil-dc.in/indowordnet/	Punjabi Word Input (Example)	Input Description	No. of Similar words	Online Access to Punjabi Wordnet through IndoWordnet available at		Reason for failure
				http://www.cfilt.iitb.ac.in/indowordnet/	http://tdil-dc.in/indowordnet/	
				Output (A)	Output (B)	
4 :: ਪਵਿੱਤਰ-ਸਥਾਨ, ਧਾਰਮਿਕਸਥਾਨ, ਪਵਿੱਤਰ-ਅਸਥਾਨ, ਪਵਿੱਤਰ-ਥਾਂ, ਪੁੰਨ-ਤ੍ਰਾਮੀ, ਪਾਵਨ_ਤ੍ਰਾਮੀ	ਪਾਵਨ_ਤ੍ਰਾਮੀ	Contains underscore	12936	Search not working	Search Successful	(A) does not search any special character
	ਪਾਵਨ ਤ੍ਰਾਮੀ	Omit underscore		Word ਪਾਵਨ ਤ੍ਰਾਮੀ not found in wordnet.	Word ਪਾਵਨ ਤ੍ਰਾਮੀ not found in wordnet.	(A) and (B) searches for exact match
5 :: ਸਿਰ ਮੰਦਰ, ਸਿਰ-ਮੰਦਰ, ਸਿਰਾਲਾ	ਸਿਰ ਮੰਦਰ	Contains space	3361	Search Successful	Word ਸਿਰ ਮੰਦਰ not found in wordnet. Suggested words: ਸਿਰ ਮੰਦਰ	(B) does not search a spaced collocation, but suggests same word
13604 :: ਸਨਾਤਕੀ, ਬੀ.ਏ., ਗੈਜ਼ੈਟਰਨ	ਬੀ.ਏ.	Contains dot	57	Search not working	Search Successful	(A) does not search any special character
7468 :: ਭੇਜਨ, ਖਾਣਾ, ਅਨੁ_ਗੁਹਿਣ, {ਪ੍ਰਸਾਦਾ_ਛੱਕਣਾ, ਲੰਗਰ_ਛੱਕਣਾ_(ਧਾਰਮਿਕ_ਖੇਤਰ_ਵਿਚ)}	{ਪ੍ਰਸਾਦਾ_ਛੱਕਣਾ	Contains curly brace	2	Search not working	Search Successful	(A) does not search any special character
	ਪ੍ਰਸਾਦਾ_ਛੱਕਣਾ	Omit curly brace		Search not working	Word ਪ੍ਰਸਾਦਾ_ਛੱਕਣਾ not found in wordnet. Suggested words: {ਪ੍ਰਸਾਦਾ_ਛੱਕਣਾ	(B) considers the { as part of the word
	ਪ੍ਰਸਾਦਾ_ਛੱਕਣਾ	Omit curly brace and underscore		Word ਪ੍ਰਸਾਦਾ_ਛੱਕਣਾ not found in wordnet.	Word ਪ੍ਰਸਾਦਾ_ਛੱਕਣਾ not found in wordnet.	(A) and (B) searches for exact match

14619 :: ਇਕੱਲਾ, 'ਕਲਾ, ਤਨਹਾ, ਏਕਾਕੀ, ਇਕੱਲਾ	'ਕਲਾ	Contains apostrophe	2	Search not working	Word 'ਕਲਾ not found in wordnet.	(B) does not search single quote
	ਕਲਾ	Omit apostrophe		Searches different word, not belonging to synset id: 14619	Searches different word, not belonging to synset id: 14619	(A) and (B) searches for exact match
547 :: "ਛੱਡਣ_ਯੋਗ, ਵਰਤਿਸਤ, ਤਿਆਗਣ_ਯੋਗ, ਪਰਤਿਆਗ	"ਛੱਡਣ_ਯੋਗ	Contains double quote	11	Search not working	Search Successful	(A) does not search any special character
	ਛੱਡਣ_ਯੋਗ	Omit double quote		Search not working	Word ਛੱਡਣ_ਯੋਗ not found in wordnet. Suggested words: "ਛੱਡਣ_ਯੋਗ	(B) considers the double quote as part of the word
2094 :: ਕਾਫੀ_ਸਮਾਂ_ਬੀਤਣ_ਤੇ, ਲੰਘੇ_ਵੇਲੇ_?	ਲੰਘੇ_ਵੇਲੇ_?	Contains question mark	14	Search not working	Word ਲੰਘੇ_ਵੇਲੇ_? not found in wordnet. Suggested words: ਲੰਘੇ_ਵੇਲੇ_?	(B) does not search question mark
	ਲੰਘੇ_ਵੇਲੇ_	Omit question mark		Search not working	Word ਲੰਘੇ_ਵੇਲੇ_ not found in wordnet. Suggested words: ਲੰਘੇ_ਵੇਲੇ_?	(B) considers the question mark as part of the word
	ਲੰਘੇ_ਵੇਲੇ	Omit question mark and underscore		Word ਲੰਘੇ_ਵੇਲੇ not found in wordnet.	Word ਲੰਘੇ not found in wordnet.	(B) searches for exact match

34372 :: ਵਾਧੂ, ਅਤਿਰਿਕਤ* Pan India Synsets *	ਅਤਿਰਿਕਤ	Contains an asterisk	2	ਅਤਿਰਿਕਤ not found in wordnet. Suggested words: ਅਤਿਰਿਕਤ/* Pan India Synsets */	Word ਅਤਿਰਿਕਤ not found in wordnet. Suggested words: ਅਤਿਰਿਕਤ/* Pan India Synsets */	(A) and (B) consider the comment as part of the word
	ਅਤਿਰਿਕਤ/* Pan India Synsets */	Input as suggested		Search not working	Word ਅਤਿਰਿਕਤ/* Pan India Synsets */ not found in wordnet. Suggested words: ਅਤਿਰਿਕਤ/* Pan India Synsets */	(A) does not search any special character (B) considers the comment as part of the word
4315 :: ਸੂਈ (ਘੜੀ ਵਾਲੀ), ਘੜੀ ਸੂਈ	ਸੂਈ (ਘੜੀ ਵਾਲੀ)	Contains parentheses	5	Search not working	Word ਸੂਈ (ਘੜੀ ਵਾਲੀ) not found in wordnet. Suggested words: ਸੂਈ (ਘੜੀ ਵਾਲੀ)	(B) does not search parentheses
	ਸੂਈ	Omit contents in parentheses		Searches different word, not belonging to synset id: 4315	Searches different word, not belonging to synset id: 4315	(A) and (B) searches for exact match
4 :: ਪਵਿੱਤਰ-ਸਥਾਨ, ਧਾਰਮਿਕਸਥਾਨ, ਪਵਿੱਤਰ-ਅਸਥਾਨ, ਪਵਿੱਤਰ-ਬਾਂ, ਪੁੰਨ-ਭਾਜੀ, ਪਾਵਨ_ਤ੍ਰਮੀ	ਪਵਿੱਤਰ-ਸਥਾਨ	Contains hyphen	2300	Search not working	Search Successful	(A) does not search any special character
	ਪਵਿੱਤਰ ਸਥਾਨ	Omit hyphen		Word ਪਵਿੱਤਰ ਸਥਾਨ not found in wordnet.	Word ਪਵਿੱਤਰ ਸਥਾਨ not found in wordnet.	(A) and (B) searches for exact match

5887 :: ਸੰ ' 6189 :: ਸੰ	Grave accent (‘) used after the synset id : 5887 with white space in between. Same word exist at synset id : 6189 correctly	1	Search Successful, but only synset id : 6189 is searched. Synset id: 5887 is not considered in the search	Search Successful, but only synset id : 6189 is searched. Synset id: 5887 is not considered in the search	(A) and (B) consider the word at Synset id : 5887 as invalid word.
Total Number of words with these issues:	18691				

Table 10 Accuracy and Inaccuracy percentage of Punjabi language WordNet

Total no. of words in the Punjabi WordNet	Number of anomalous words in the Punjabi WordNet	Accuracy percentage	Inaccuracy percentage
52793	18691	64.6	35.4%

4 Conclusions

Punjabi language WordNet is a part of the IndoWordNet which is a multilingual WordNet for eighteen Indian languages. It is an appreciable development of language resources which can be considered as a milestone for NLP research in India. However, Punjabi language WordNet has various anomalies which were detected by directly analyzing the Punjabi WordNet text files and then verified with online versions of IndoWordNet available at two different web locations. These anomalies include prefixed double quotation marks, missing collocations uniformity, suffixed question marks, misplaced grave accent, missing abbreviations uniformity, use of dot in place of comma, misused curly braces/parentheses, and misplaced comments. The paper provided a detailed list of anomalous words in separate tables along with respective Synset IDs for reference, except from the anomalous collocations which is a long list of more than 16000 words. The paper also presented the test results produced by online versions of the IndoWordNet. It is clear that the removal of these anomalies is a requirement for the improvement of Punjabi language WordNet, so that the future research in NLP can use this rich structured resource of Punjabi language vocabulary more accurately. This paper provided a detailed explanation of each and every issue and will be very helpful to NLP researchers as well as IndoWordNet development team. The aim of the research findings presented in this paper was to assist for the improvement of Punjabi language WordNet.

References

1. Luk, S.K., Knight, K.: Building a largescale knowledge base for machine translation. AAAI 94, 773–778 (1994)
2. Reddy, M., Manish Sinha, P.B.: An approach towards construction and application of multilingual indo-wordnet. In: 3rd Global Wordnet Conference (GWC 06), Jeju Island, Korea (2006)
3. Verdejo, F., Chugur, I., Julio Gonzalo, J.C.: Indexing with wordnet synsets can improve text retrieval. arXiv preprint cmplg (1998)
4. Harabagiu, S., Pasca, M.: The informative role of wordnet in open-domain question answering. In: Proceedings of NAACL-01 Workshop on WordNet and Other Lexical Resources, pp. 138–143 (2001)
5. Fellbaum, C.: WordNet.: Wiley Online Library (1998)
6. George, A., Fellbaum, R., Gross, C., Miller, D., Miller, K.J.: Introduction to wordnet: An on-line lexical database. Int. J. Lexicography 3(4), 235–244 (1990)
7. Vossen, P.: EuroWordNet: a multilingual database for information retrieval. In: DELOS, Zurich, pp. 5–7 (1997)
8. Bhattacharyya, P.: Indowordnet. In: Lexical Resources Engineering Conference 2010 (LREC 2010), Malta, May 2010
9. Chakrabarti, D., Pande, P., Dipak Narayan, P.B.: An experience in building the indo wordnet-a wordnet for hindi. In: First International Conference on Global WordNet, Mysore, India (2002)
10. Sharma, R.K., Ashish Narang, P.K.: Development of Punjabi WordNet. CSI Trans. ICT 1(4), 349–354 (2013)
11. Patel, K., Diptesh Kanodia, P.B.: India language Wordnets and their linkages with princeton WordNet. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018), Miyazaki, Japan (2018)
12. Bhattacharyya, P., Pawar J.D., Sekhar Dash, N. (Eds.): The WordNet in Indian Languages. Springer (2017)
13. Bhatia, T.K.: Major regional languages. In: Kachru, Y., Sridhar, S.N., Kachru, B.B. (Eds.) Languages in South Asia. Cambridge University Press, pp. 121–131 (2008)
14. Abbas Malik, M.G.: Punjabi machine transliteration. In: Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, Sydney, pp. 1137–1144 (2006)
15. Sharma, R.K., Preet, S., Bhatia, P., Kaur, R.: Punjabi WordNet Relations and Categorization of Synsets. In: 3rd national workshop on IndoWordNet under the aegis of the 8th international conference on natural language processing (ICON 2010), Kharagpur, India (2010)
16. Princeton WordNet. [Online]. <https://wordnet.princeton.edu/documentation/wninput5wn>
17. National Workshop on WordNet Creation. [Online]. www.cfilt.iitb.ac.in/coimbatore-amrita-university-wordnet-workshop.pdf

Supervised Machine Learning Strategies for Investigation of Weird Pattern Formulation from Large Volume Data Using Quantum Computing



Mukta Nivelkar and S. G. Bhirud

Abstract Quantum machine learning [QML] is a new formulation on quantum hardware platform will try to achieve more enhanced data analysis and prediction which classical computer will not be able to generate. Classical computers are having computational limitations in terms of large volume data processing. Quantum machine are not to replace classical machines but quantum computers will solve operational difficulties of classical machines in terms of computational time. Quantum machine learning accelerates the supervised, unsupervised, and reinforcement learning methods. Classical ML methods such as SVM, PCA, Clustering, Neural networks are giving promising results but classical machines are inadequate to perform certain computations. Proposed Quantum machine learning would result in complex and weird patterns. Quantum support vector machine(QSVM) is a method used in supervised learning for classification and regression. QSVM uses high-dimensional feature space possibly on infinite dimension called as enhanced feature space for generating hyperplane. This hyperplane will classify non-linear and complex data on multi-class domain to achieve improved accuracy in less computational time. This paper investigates various strategies for quantum enhanced machine learning algorithm in supervised learning method.

Keywords Quantum computing · Machine learning · Qubit · Quantum machine learning · Support vector machine

1 Introduction

Objective is investigation of methods for modeling and simulation of quantum machine learning algorithm. QML would be time efficient and more reliable. Quantum Computing will activate the classical algorithms by processing huge amount of data. Quantum and classical computers are compared on operational structures.

M. Nivelkar (✉) · S. G. Bhirud

Veermata Jijabai Technological Institute, Matunga, Mumbai, India

URL: <http://vjti.ac.in>

Quantum computing promises better performance of real-time data processing. Some real-time dataset processing will be time consuming on supercomputers too. Quantum computing is not meant for replacing the classical machines but this new standard of computing will be special purpose computing. Quantum computing research is growing in the various fields such as machine learning, cryptography, security, and many more.

1.1 Need of Quantum Machine Learning

Quantum Machine learning will perform complex analysis and prediction in terms of weird pattern generation. Complex and high-dimensional pattern generation will be not possible on classical machine at some extent. Quantum ML will speed up the task of big data processing to achieve accuracy more than traditional approach of ML. Proposing quantum machine learning algorithm: a new era of computing. Quantum computer can analyze vast amount of data and predict certain result in very less time in comparison with traditional computers such as digital and high performance computers. These machines are still taking longer time to predict accurate analysis. By proposing quantum machine learning algorithm, the model which will predict pattern in very short time with excellent accuracy. Quantum machine learning will help to generate and model system which will give faster result on classical machine [1–3].

1.2 Quantum Supervised Learning

Classical SVM is well-established methods in supervised learning which we can use for classification and regression as well. SVM can classify objects in the nth-dimensional feature space (N is the number of features). Objective is to find a hyperplane in nth-dimensional space that distinctly classifies the data points. Typical SVM method is used for binary data classification such as dataset which has two classes for classification. The two dimensions (i.e., dim1 and dim2) are not sufficient to classify data because dimension space is very small where linear data will support but if data is nonlinearly separable then it will not be able to classify with small dimensional space. Solution for such classification problem is Quantum representation of data and algorithm. Higher dimension allows the data to represent on more enhanced feature space and depths. The third dimension (i.e., dim3) allows the computer to understand the difference. To project the obj1 and obj2 into a higher dimensional space. Slide thin sheet between the two to separate them instead it will be “Line” which will classify on two classes which are linearly separable. Data on linearly and non-linearly separable classes. Proposed SVM solution for multiclass separation of data on higher dimension feature space where it will process and create more than two hyperplanes for classification. Not only supervised learning, but quantum com-

puting can also work better in unsupervised as well as reinforcement learning too which has been mentioned in future scope. Quantum machine learning for following algorithm defined as:

Quantum clustering method: When the data is represented in very large dimension space, it is very difficult to perform clustering with a classical computer. The use of Quantum Computer is a very good solution.

Quantum support vector machine: This type comes from classification and regression and it will find the hyperplane that separates many data points that are represented in very high-dimensional space is so difficult on a classical computer. On a quantum computer, it can be solved extremely efficiently.

Quantum PCA: The goal of this algorithm is to find the proper axes along which to group this data. This is something that takes $O(N^2)$ on a classical computer. But in quantum version you can do it exponentially faster.

Feature selection and topology: This is a method for finding the topological features of data. This problem of finding the eigenvectors and eigenvalues of some huge, high-dimensional matrix.

Quantum neural networks: Exciting breakthroughs may soon bring real quantum neural networks, specifically deep learning neural networks, to reality. Many research papers have shown remarkable results in quantum deep learning [4].

1.3 Modeling Strategies in Quantum Space

1. SVM is classification and regression method for binary data classification on classical machine which will be defined on quantum techniques.
2. Data representation will be on higher dimension and enhanced feature space.
3. Clarity on data visualization on higher dimensional space will be more.
4. SVM single hyperplane and multiple hyperplane generation will be proposed.
5. Hyperplane generation will be on random manner based upon classes available in dataset.
6. Time required for selection and generations of vector will be minimized.
7. Big dataset will be used accordingly classification of data will be done.

1.4 Aim and Objective

To investigate and Estimate the performance of Quantum machine learning algorithm on high dimension data for improved analysis and prediction. Quantum machine learning algorithm promises better and faster performance over time compare to

classical methods. Measuring the performance of quantum supervised and unsupervised method over classical machine learning for betterment of next generation of machine learning.

Following objectives are listed based upon problem statement and gap identification.

1. Study of classical supervised and unsupervised learning.
2. Study of SVM in supervised learning for classification and regression.
3. Propose mathematical model for quantum SVM.
4. Study of dataset and conversion of classical dataset into quantum dataset.
5. Study of feasible programming tools for quantum programming.
6. Selection of quantum simulator and execution of algorithm.
7. Comparison of time complexity wrt. classical and quantum computing.

2 Quantum Machine Learning Algorithm

Quantum machine learning accelerates the supervised, unsupervised, and reinforcement learning methods. Classical ML methods such as SVM, PCA, Clustering, NN are giving promising results but classical machines are inadequate to perform certain computations. Proposed Quantum machine learning would result in complex and weird patterns. Quantum support vector machine (QSVM) is a method used in supervised learning for classification and regression. QSVM uses high-dimensional feature space possibly on infinite dimension called as enhanced feature space for generating hyperplane [5]. This hyperplane will classify non-linear and complex data on multiclass domain to achieve improved accuracy in less computational time [6, 7]. If data is linearly Separable.

1. 2D hyperplane and 3D hyperplane

If data is Non-linearly Separable. We can classify data on high-dimensional feature space that means data represented on a very large scale.

1. Multiple hyperplanes which are going to classify data on Multiclass space [More than two] using SVM.
2. Classical Binary SVM[Which is restricted to only two Classes].

3 Features of Quantum Machine Learning

Following points are proposed to enhanced machine learning

1. Classical machine learning has computational complexities on classical machines in terms of complex pattern generation on reduced computational time.
2. Weird pattern generation which classical computing cannot do.

3. Quantum classification and clustering for enhanced prediction on higher dimensions of data.
4. Quantum efficiency and scalability on a more powerful computational standard.
5. Data and information processing to achieve more computational speed.
6. Quantum data analytics for real time and faster result generation with accurate analysis and prediction for big data processing.

4 Comparison of ML and QML Result

Table 1 shows the time required for quantum svm to perform on classical and quantum machine and comparison is given. Time is dependent on various parameters such as hardware and software which have been used for demonstration, and also time will vary based upon various factors used. ML algorithm is tested for the following points on classical and quantum platform. In this paper experimental results are tested on small dataset.

Dataset used: Iris Dataset

Characteristics: The dataset contains a set of 150 records under various attributes—sepal length, sepal width, petal length, petal width, and species. Dataset has three classes (Iris setosa, Iris virginica, and Iris versicolor). Quantum cloud experience is achieved through IBM Quantum Experience. IBM quantum hardware is accessible through cloud for authorized user [8].

1. Dataset training time
2. Support vector calculation time
3. Classification and prediction time
4. Total execution time.

4.1 SVM Variations

- (a) Hard-Margin SVM

Input dataset with M instances where

$$x \in R^N$$

$y \in \{+1, -1\}$ Hyperplane classifies each instance correctly. Idea of hard-margin SVM in two dimensional dataset.

Table 1 Estimated Time for Quantum SVM Practical Implementation

Model	Training time(sec)	SV finding time(sec)	Prediction time (sec)	Total ET(sec)	Accuracy
Classical SVM	0.140260	0.144761	0.278502	0.563523	98.18%
Quantum SVM	0.007368	0.008460	0.008709	0.024537	Hardware dependent

This type will classify linearly separable dataset which will be having only two classes.

Hard-margin SVM mathematically represents as $y_i(w_i^T x_i + b_i) \geq 1 \quad \forall i = 1 \dots M$

(b) Non-linear SVM and Kernel Trick

Non-linear separation would transform input feature space x as z and $x = \phi(z)$

by mapping to higher dimensional space of R^N .

In this type data is non-linearly separable and represented on higher dimensional space. $\text{argmin} \frac{1}{2} \sum_{i=1}^M \sum_{j=1}^M \alpha_i \alpha_j y_i z_i^T z_j + \sum_{i=1}^M \alpha_i$

where $\sum_{i=1}^M y_i \alpha_i = 0$

(c) Soft Margin SVM

Unlike, hard-margin SVM, soft margin SVM allows the data to be classified incorrectly.

Where ξ_i represents violation if x_i is at wrong side of margin

$y_i(w_i^T x_i + b_i) < 1$

$$\text{arg min } \frac{1}{2} \omega^T \omega + c \cdot \sum_{i=1}^M \xi_i$$

$$y_i(w_i^T x_i + b_i) \geq 1 - \xi_i \quad \forall i = 1 \dots M$$

where $\xi_i \geq 0 \quad \forall i = 1 \dots M$.

5 Future Work

Future work proposed in the direction of unsupervised and reinforcement machine learning methods. Future scope will contribute quantum implementation of machine learning in supervised, unsupervised, and reinforcement learning. Quantum implementation of machine learning highlights the logic of every machine learning method. Proposed future task will compare time complexity in quantum and classical methods of machine learning.

6 Conclusion

The quantum computer makes more enhancement on various domains that are currently running on to the classical computational standards. Quantum machines can prove new level of computational power in conclusion. Current study of research includes the understanding concept of qubit, superposition, quantum gate structure, etc. Further research study requires to know about quantum algorithms study and analysis over classical algorithms in terms of quantum mechanism, tools, and various

mathematical concept. Now the task of implementation is going to make changes over last few years of research in this domain because platforms in terms of circuit-based quantum machines are available on recent days. Even few vendors have given complete free access to their quantum services to the world so that people can explore their ideas and come up with a healthier solution in quantum domain. Quantum machine learning will definitely be going to contribute more in terms of clear and accurate data analysis and prediction in terms of voluminous data processing. Quantum mechanism such as superposition and entanglement will be used for quantum enhanced machine learning modeling and design. Quantum supervised, unsupervised, and reinforcement learning will supercharge classical machine learning in future.

References

1. Ablayev, F., Ablayev, M., Huang, J.Z., Khadiev, K., Salikhova, N., Wu, D.: On quantum methods for machine learning problems part I: quantum tools, big data mining analytics, vol. 3, no. 1, pp. 41–55 (March 2020). <https://doi.org/10.26599/BDMA.2019.9020016>. ISSN 2096-0654 ll04/06ll
2. Ablayev, F., Ablayev, M., Huang, J.Z., Khadiev, K., Salikhova, N., Wu, D.: On quantum methods for machine learning problems part II: Quantum classification algorithms, big data mining anlytics, vol. 3, no. 1, pp. 56–67 (March 2020). <https://doi.org/10.26599/BDMA.2019.9020018>. ISSN 2096-0654 ll05/06ll
3. Li, Y., (Senior Member, IEEE), Tian, M., Liu, G., Peng, C., Jiao, L. (Fellow, IEEE): Learning-Based Quantum Robust Control: Algorithm, Applications, and Experiments. <https://doi.org/10.1109/ACCESS.2020.2970105>
4. Nguyen, N.H., Behrman, E.C., Moustafa, M.A., Steck, J.E.: Benchmarking neural networks for quantum computations. In: IEEE Trans. Neural Netw. Learn. Syst
5. Willsch, D., Willsch, M., Raedt, H.D., Michielsen, K.: Support vector machines on the D-Wave quantum annealer. Elsevier (2018). www.elsevier.com/locate/cpc
6. Rebentrost, P., Mohseni, M., Lloyd1, S.: Quantum Support Vector Machine for Big Data Classification In: 2Google Research, Venice, California 90291, USA (2014). (Received 12 February 2014; published 25 September 2014)
7. Tod, M.: Quantum vs Classical survival of the ttest SVM A venture into Support Vector Machines using Qiskit In: School of Computer Science, vol. XX (2019)
8. IBM Quantum Experience. <https://quantum-computing.ibm.com/>
9. Dong, D., Member, IEEE, Chen, C., Member, IEEE, Li, H., Senior Member, IEEE, Tarn, T.-J., Fellow, IEEE: Quantum reinforcement learning. IEEE Trans. Syst. Man Cybern. Part B **38**(5) (2008)
10. Imran, Ahmad, S., Kim, D.H.: Quantum GIS based descriptive and predictive data analysis for effective planning of waste management. Received January 29, 2020, Accepted March 2, 2020, date of publication March 6, 2020, date of current version March 17, 2020. <https://doi.org/10.1109/ACCESS.2020.2979015>. (Department of Computer Engineering, Jeju National University, Jeju 63243, South Korea)
11. Fanizza, M., Mari, A., Giovannetti, V.: Optimal universal learning machines for quantum state discrimination. IEEE Trans. Inf. Theory **65**(9) (2019). <https://doi.org/10.1109/ACCESS.2020.2979015>
12. Nawaz, S.J., (Senior Member, IEEE), Sharma, S.K., (Senior Member, IEEE), Wyne, S., (Senior Member, IEEE), Patwary, M.N., (Senior Member, IEEE), Asaduzzaman, M. (Member, IEEE): Quantum machine learning for 6g communication networks: state-of-the-art and vision for

- the future. Received March 12, 2019, Accepted April 2, 2019, date of publication April 4, 2019, date of current version April 17, 2019. <https://doi.org/10.1109/ACCESS.2019.2909490>. (Department of Computer Engineering, Jeju National University, Jeju 63243, South Korea)
- 13. Dong, D., Senior Member, IEEE, Xing, X., Ma, H., Chen, C., Member, IEEE, Liu, Z., Rabitz, H.: Learning-based quantum robust control: algorithm, applications, and experiments. *IEEE Trans. Cybern.*
 - 14. Li, Y., (Senior Member, IEEE), Tian, M., Liu, G., Peng, C., Jiao, L., (Fellow, IEEE): Learning-Based Quantum Robust Control: Algorithm, Applications, and Experiments

A Comparative Analysis of Intelligent Classifiers for Seizure Detection Using EEG Signals



Arshdeep Singh, Debargho Basak, Upamanyu Das, Priya Chugh, and Jyoti Yadav

Abstract This paper presents a non-patient-specific methodology to offer a comparative analysis of the epileptic seizure prediction techniques using various machine learning classifiers based on the features extracted from electroencephalogram (EEG) signals. This methodology can be divided into subsequent stages of channel selection, feature extraction, feature selection, and prediction phase. The channel selection was implemented by employing the Boruta algorithm. The best performing channels were chosen. In feature extraction we investigated three fundamental roads: extracting statistical features from the raw EEG signals (mean variance, mean skewness, and mean kurtosis), performing empirical mode decomposition on raw EEG data to generate intrinsic mode functions and extracting statistical features from the generated intrinsic mode functions, and calculating the power spectral density of all channels partitioned into five frequency bands. These extracted features were then tested for their efficacy using the Boruta algorithm, resulting in the most desirable ones being selected. At last, in the prediction phase, the resultant dataset is tried on different machine learning classifiers: support vector machines (SVM), random forest, K-nearest neighbors (KNN), and LSTM. The sensitivities were tabulated to offer a comparative analysis.

Keywords EEG · Seizure prediction · Channel selection · Boruta · Statistical features

1 Introduction

Epilepsy is one of the most well-known neurological disorders, troubling roughly 1% of the total populace and affects around 1 million Indians every year. It is extremely uncertain in nature. It results in the cerebrum action becoming irregular, causing seizures or bouts of unordinary conduct, sensations, and occasionally leads to a loss of awareness. Epilepsy may occur because of a hereditary problem or due to a gained

A. Singh · D. Basak (✉) · U. Das · P. Chugh · J. Yadav
Netaji Subhas University of Technology, Dwarka Sector-3, Dwarka, Delhi 110078, India
e-mail: debarghob.ic.17@nsit.net.in

cerebrum injury, for example, an injury or stroke. What makes the comprehensive investigation of epilepsy considerably more troublesome is the way that most patients just spend a minuscule measure of their time encountering a seizure and show no clinical prodromes of their illness during the time between the seizures (inter-ictal span). Along these lines, to date, there is no monetarily accessible gadget to anticipate the onset of seizures. Thus, building up a robust and sustainable procedure to anticipate epileptic seizures has been the essential focal point of analysts in this field for an extensive stretch of time. During the 1970s, Viglione and Walsh [1] spearheaded the research to determine the link between seizures and EEG recordings. From that point forward, numerous studies have been done intending to anticipate epileptic seizures depending on the EEG information. Our study aims at building up a powerful and sustainable seizure forecast calculation. Seizure prediction refers to predicting/forecasting/anticipating the occurrence of epileptic seizures, typically by tracking the brain's electrical activity by means of electroencephalographic (EEG) recordings. The early expectation of seizures will enable doctors and specialists to shelter their patients from the endangerment brought about by epileptic seizures. The proficiency of any seizure predicting algorithm is characterized by its sensitivity, accuracy, and number of false alarms produced.

The entirety of the seizure expectation models has certain basic highlights. A large portion of them has two essential advances. First, every one of them attempts to identify and separate EEG-based measures over the long haul, portraying various phases of the epilepsy cycle. In such a manner, they utilize a moving window study wherein a straight or nonlinear portraying measure is determined from a window of EEG information with a predefined length [2]. The length of the time window is picked so that there is a sensible compromise between surmised stationarity of the EEG signal and an adequate number of information focused to portray the EEG elements. The subsequent advancement is recognizing and characterizing the measures into preictal and ictal states. Different paths have been explored in this literature enveloping different strategies for EEG seizure forecast with primary roads being time-area and the wavelet approach which contain signal deterioration and sign changes. Ibrahim et al. [3] used likelihood thickness capacities (PDFs) alongside predefined forecast and bogus caution likelihood limits for predicting seizures. They attained a maximum accuracy of 93.55% with a false-alarm rate of 0.074/h. The authors of [4] presented a seizure prediction algorithm depending on the interaction between pairs of EEG signals, and this algorithm attained a sensitivity of 60%. Wang et al. [5] presented a novel self-adaptation method that effectively integrates reinforcement learning, online monitoring, and adaptive control theory. They achieved an accuracy of 71.34%. Usman et al. [6] used empirical mode decomposition (EMD) for preprocessing phase and have extracted time and frequency domain features for training a prediction model, and this algorithm realized an average sensitivity of 92.23% and specificity of 93.38%. Li et al. [7] investigated the utilization of spike wave releases and spike rate (morphological tasks) alongside averaging channels for EEG seizure expectation. They achieved a sensitivity of 75.8% and a false-alarm rate of 0.09/h (Table 1).

Table 1 A comparison of various seizure prediction algorithms

Algorithm	Characteristic attributes	Performance
Ibrahim et al. [3]	Probability density functions (PDFs) along with predefined prediction and false-alarm probability thresholds	Maximum accuracy of 93.55% with a false alarm rate of 0.074/h
Usman et al. [6]	Empirical mode decomposition (EMD)	Average sensitivity of 92.23% and specificity of 93.38%
Wang et al. [5]	Lyapunov exponent, correlation dimension, Hurst exponent, and entropy features	Sensitivity of 73%, and specificity of 67%
Chiang et al. [8]	Nonlinear independence, cross-correlation, difference of Lyapunov exponents and phase locking	Sensitivity of 74.2%
Schelter et al. [4]	Interaction between pairs of EEG signals	Sensitivity of 60%
Li et al. [7]	Morphological features	Sensitivity of 75.8% and a false-alarm rate of 0.09/h
Wang et al. [2]	Distance values to measure similarity	Accuracy of 70%

Versions of wavelet transform have also been used for EEG seizure prediction. Wang et al. [5] exploited Lyapunov exponent, correlation dimension, Hurst exponent, and entropy features in the wavelet domain for seizure prediction. They achieved an average sensitivity of 73%, and a specificity of 67%. Chiang et al. [8] developed a wavelet-based seizure prediction algorithm adopting nonlinear independence, cross-correlation, the difference of Lyapunov exponents, and phase locking. This method realized a 74.2% sensitivity on the CHB-MIT database.

One of the biggest challenges that confront researchers while exploring seizure prediction is that of EEG channel selection. This is often a crucial step in any biocomputational process associated with EEG signals because it permits to tame numerous shortcomings, which include limited channel capacity, low computational power, and overall model performance through compressing the signal. Boruta [9] is a feature selection algorithm. It's a wrapper constructed on the random forest package in which classification is performed by voting of multiple unbiased weak classifiers—decision trees. Boruta has been established in past research to give top-notch results.

This paper proposes an approach based on the Boruta [9] study for EEG channel selection, and a comparative analysis of various machine learning seizure prediction algorithms is performed based on statistical features. The foremost pretense of the paper is to employ a complex method, i.e., Boruta to help properly clean the EEG data. The proposed operating framework of channel selection allows us not only to attain better results in the subsequent stages of features extraction and testing but also greatly reduces memory usage. Further, it allows us to offer an alternative course to contribute to the development of intelligent devices for non-patient-specific seizure prediction.

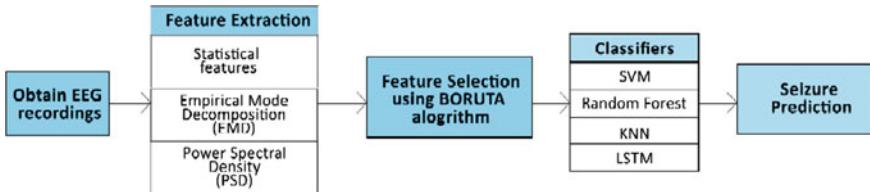


Fig. 1 Flowchart for proposed seizure prediction method

This paper is structured as follows. The suggested approach is presented in Sect. 2. In Sect. 3, the experimental findings are provided and we conclude the paper in Sect. 4 (Fig. 1).

2 Methodology

2.1 Dataset

The CHB-MIT dataset [10] has been used for this research project. The dataset contains seizure data of 22 pediatric patients from the Children's Hospital of Boston. The data is present as edf files with 23 channels of EEG signal information. Five of the 22 subjects were selected for this study. The chosen patients are patient no. 1, 8, 11, 14, and 20, the same as Ibrahim et al. [3].

2.2 Channel Selection

To reduce noise and computational load, channel selection was performed. The Boruta algorithm [9] was applied to the 23 channel seizure data of the selected patients. The best performing channels were selected.

The Boruta algorithm makes a randomized copy of the provided dataset and merges the two. It then builds the classifier on the extended dataset. A number of random forest runs are conducted, with the replicated features randomized before each run, and the importance of all the features is computed. A feature is deemed important for a run if its importance is higher than the maximum importance of all randomized features. A two-sided equality test is performed for each feature to select a certain feature, with the null hypothesis being that the importance of a feature is equal to the maximum importance of the random features. When the number of hits is higher than the expected value, a feature is considered important and is rejected when the number of hits is lower than the expected value. This procedure is performed for a predefined number of iterations, or until all features are either rejected or accepted.

2.3 Feature Extraction

Three avenues were explored to extract features from the selected channels—extracting statistical features from the raw EEG signals [11], performing empirical mode decomposition on raw EEG data to generate intrinsic mode functions and extracting statistical features from the generated intrinsic mode functions, and calculating the power spectral density of the 23 channels divided into five frequency bands—Delta (0.5–4 Hz), Theta (4–8 Hz), Alpha (8–13 Hz), Beta (13–30 Hz), and Gamma (36–90 Hz) (Table 2).

Statistical Features from Raw EEG Data

Statistical features [12], namely variance, kurtosis, and skewness, were extracted from the raw EEG signal.

Variance

Variance is the measure of spread of a particular dataset. The greater the spread of the data, the higher is the variance in relation to the mean.

In short, it can be defined as a measure of variability.

$$s^2 = \frac{\sum (xi - x)^2}{(n - 1)} \quad (1)$$

Kurtosis

Kurtosis is a statistical measure that illustrates the extent to which data points cluster in a frequency distribution's tails or peaks.

$$k = \frac{\mu_4}{\sigma^4} \quad (2)$$

Skewness

Skewness measures the level of dissymmetry/distortion in a standard distribution. It gauges this deviation with the help of a random variable from a symmetric distribution.

Table 2 List of extracted features

Time domain features	Frequency domain features
Kurtosis	Power spectral density
Skewness	
Variance	
Hjorth mobility	
Petrosian fractal dimension	
Hjorth complexity	

$$\text{skew} = \frac{\sum_i^N (x_i - \bar{x})^3}{(N - 1) * \sigma^3} \quad (3)$$

Other statistical features, such as Petrosian fractal dimension and the Hjorth parameters [12], were also extracted from the data.

Petrosian Fractal Dimension

The Petrosian fractal dimension is a simple way to estimate the fractal dimension of a finite sequence that converts the data to a binary representation before evaluating the time series' fractal dimension.

This metric gives us a measure of the complexity of our signal.

Hjorth Parameters

In the analysis of EEG signals, Hjorth parameters are the measures of statistical properties used in signal processing in the time domain. These parameters are widely used as statistical property indicators. We used two of three Hjorth parameters in our research mobility and complexity.

The mobility metric reflects the mean frequency or the proportion of the power spectrum standard deviation

$$\text{Mobility} = \sqrt{\frac{\text{var}\left(\frac{dy(t)}{dt}\right)}{\text{var}(y(t))}} \quad (4)$$

The shift in frequency is expressed by the Complexity parameter. This parameter compares the resemblance of the signal against a pure sine wave, where if the signal is quite similar, the value converges to 1.

$$\text{Complexity} = \frac{\text{Mobility}\left(\frac{dy(t)}{dt}\right)}{\text{Mobility}(y(t))} \quad (5)$$

Empirical Mode Decomposition

The method of empirical mode decomposition (EMD) [6] decomposes a large dataset into a finite number of elements. For the original signal, these components form a complete and almost orthogonal basis.

An averaging filter was used to generate a surrogate channel from the 23 original channels in the raw EEG data. EMD was performed to decompose the surrogate channel into its oscillatory components, also known as intrinsic mode functions (IMFs). Statistical moments like mean, standard deviation, kurtosis, and skewness were then calculated from the selected IMFs.

Power Spectral Density

The signal's power spectral density (PSD) defines the power present in the signal per unit frequency, as a function of frequency. To complement the time-domain features extracted power spectral density was chosen as a frequency feature to be extracted. The EEG signal power spectral density was calculated by first taking the signal's fast Fourier transform and then dividing the FFT squared amplitude obtained by the FFT bin width. The PSD values of the five frequency bands [13], namely—Delta (0.5–4 Hz), Theta (4–8 Hz), Alpha (8–13 Hz), Beta (13–30 Hz), and Gamma (36–90 Hz) are then used as features.

2.4 Feature Selection

Following the process of feature extraction, feature selection was also performed with the help of the Boruta algorithm [9]. It is important to note that the application of a feature selection/ranking algorithm, in most cases, increases the efficacy of results with dimensionality reduction. However, since our model involves only eight total features, we chose to implement a feature selection algorithm which ranked features based on their correlation with seizure occurrence in contrast with the dimensionality reduction algorithm which handles datasets comprising a large number of dimensions. After the conclusion of this process, the best performing features were selected and the rest were excluded/dropped from the dataset. This process allowed us to effectively streamline the data into a form that conforms with fruitful results.

2.5 Performance Metrics

To better understand our hypothesis, let us first look at the performance metrics we used to evaluate the effectiveness of our models (Table 3).

We used four performance metrics. These were sensitivity, precision, F_1 score, and Cohen's kappa coefficient. Sensitivity is the proportion of real positive cases expected to be positive (or true positive). This suggests that there would be another proportion of real positive events, which will be wrongly predicted as negative (i.e., false negative); on the other hand, precision is the ratio of positive observations (true positives) correctly predicted to the total positive observations predicted (true and

Table 3 A typical confusion matrix

Prediction/Reference	1	0
1	True positive (TP)	False positive (FP)
0	False negative (FN)	True negative (TN)

false positives). Furthermore, the F_1 score is a measure of a model's accuracy on a dataset. The F_1 score is a way of combining the sensitivity and precision of the model, and it is defined as the harmonic mean of the model's precision and recall.

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6)$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (7)$$

$$F_1 = \frac{2 * \text{sensitivity} * \text{precision}}{\text{sensitivity} + \text{precision}} \quad (8)$$

Lastly, Cohen's kappa coefficient (k) is a statistic which for qualitative objects measures inter-rater agreement. It is accepted as a more robust measure than simple percent agreement calculation since k compensates for agreements occurring by chance. A simpler way to believe this is often that Cohen's kappa may be a quantitative measure of the trustworthiness of two raters that are rating the same thing. Cohen's kappa coefficient lies in the range $[0, 1]$.

$$k = \frac{p_o - p_e}{1 - p_e} \quad (9)$$

where

- k = Cohen's kappa coefficient
- p_o = relative consensus between the raters.
- p_e = The hypothetical expectation of an agreement on chance.

These variables can be computed as follows:

$$\text{Total} = \text{TP} + \text{TN} + \text{FP} + \text{FN} \quad (10)$$

$$p_o = \frac{\text{TP} + \text{TN}}{\text{Total}} \quad (11)$$

$$p_e = p_{\text{yes}} + p_{\text{no}} \quad (12)$$

$$p_{\text{yes}} = \frac{\text{TP} + \text{FP}}{\text{Total}} * \frac{\text{TP} + \text{FN}}{\text{Total}} \quad (13)$$

$$p_{\text{no}} = \frac{\text{FN} + \text{TN}}{\text{Total}} * \frac{\text{FP} + \text{TN}}{\text{Total}} \quad (14)$$

The use of precision instead of specificity is an important feature of our study. This is unique to our project as precision is not often used as a performance metric

due to the fact that it does not use the negative entries in the matrix. However, as our proposed method is non-patient-specific and only focuses on predicting seizures, i.e., positive class (true positives and false positives) the identification of the negative class (false negatives and true negatives) is not of utmost importance. Nevertheless, to optimize our performance metrics even more, we evaluated the Cohen's kappa coefficient to boost the reliability of our results.

3 Results

Following the steps mentioned in the previous section, namely channel selection, feature extraction, and feature selection, we finally reach the testing phase of our research. Experiments have been conducted on five subjects from the CHB-MIT database [10] (subjects 1, 8, 11, 14, 20) with 148.6133 h of EEG data including a total of 31 seizures.

The following box plot depicts the outset of our research in which we apply the Boruta algorithm on the PSD values of each channel to obtain the best performing channels. The selected channels are FP1-F7, FP1-F3, FP2-F4, FP2-F8, T8-P8, and FT9-FT10 (Fig. 2).

Following this, features were extracted from the six selected channels, mentioned above. The extracted features were variance, skewness, kurtosis, Petrosian fractal dimension, Hjorth mobility, Hjorth complexity, and power spectral density.

The extracted features were again fed into the Boruta algorithm to check for their respective efficacy in regard to future classifier performance (Fig. 3).

Variance and power spectral density (PSD) performed the best among the statistics features, hence they were included. In PSD the Alpha and Theta band values showed a promising correlation with the occurrence of seizures. When it comes to the other

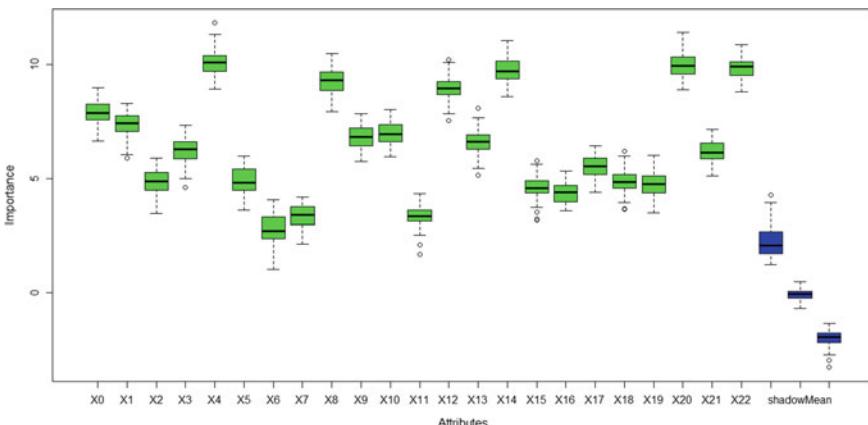


Fig. 2 Channel importance in predicting seizures as determined by the Boruta algorithm

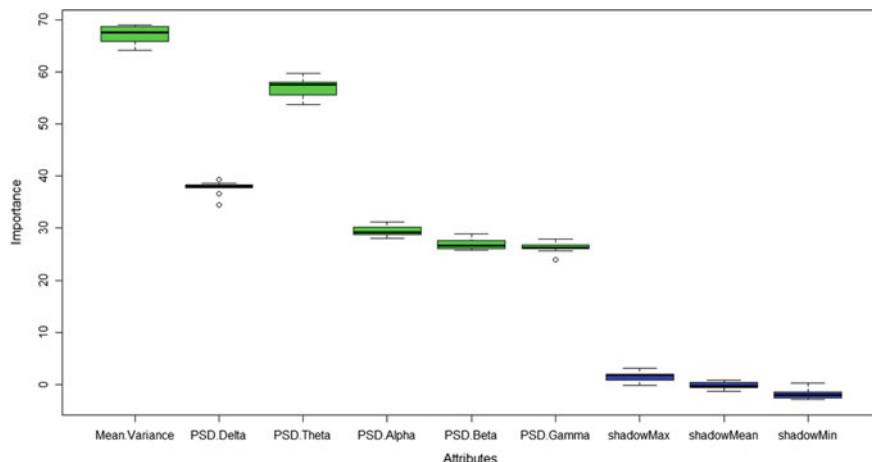


Fig. 3 Boxplot of feature importance

features, i.e., kurtosis, skewness, Petrosian fractal dimensions and Hjorth parameters did not contribute effectively to the results and hence were excluded. The features extracted from the intrinsic mode function values showed no significant increase in the precision or sensitivity of the model when included. Thus, they were excluded as well during the compilation of results.

The resultant EEG data after preprocessing is tested on four machine learning classifiers to predict epileptic seizures, which are as follows: K-nearest neighbors (KNN), support vector machines (SVM), long short-term memory (LSTM), and random forest. After applying all the models, the performance of every single algorithm was compiled and tabulated, to offer a comparative analysis, i.e., to assess the best performing classifier (Table 4).

Table 4 List of hyperparameters of machine learning classifiers

Classifier	Hyperparameters			
SVM	Soft margin constant (C): 20	Kernel: Radial basis function (RBF)		
Random forest	Nearest estimators: 500			
KNN	Distance metric: Manhattan	Nearest neighbors: 5	Weights: distance	
LSTM	Optimizer: Adam	Loss function: binary cross entropy	Layer 1: LSTM	Layer 2: Dense
	Input layer activation function: ReLu	Output layer activation function: Sigmoid		

In reference to the performance metrics discussed above, a model comparison is done on their respective prediction results.

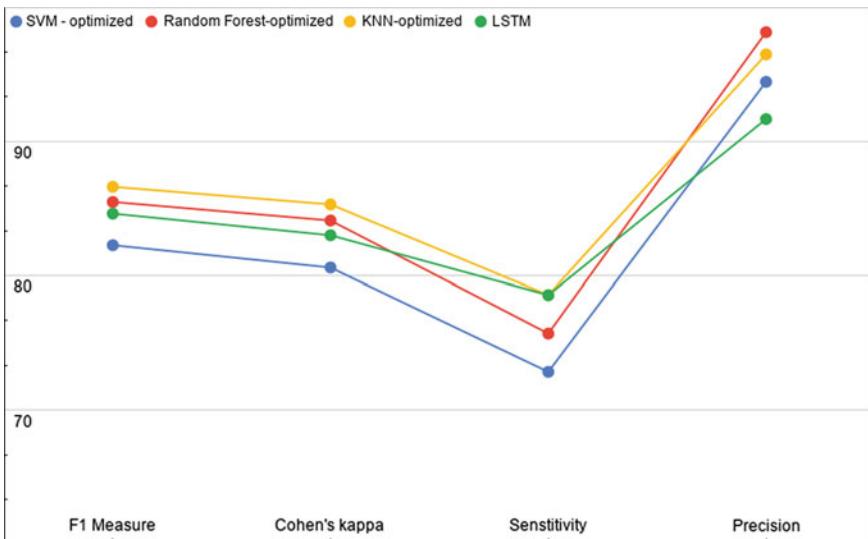
- **SVM**—We implemented a support vector machine with an RBF kernel. We performed a grid search for model optimization and tuned the punishment parameter (C) to 20. The evaluated performance metrics were:
 - Sensitivity = 72.8%
 - Precision = 94.4%
 - F_1 score = 82.2%
 - Cohen's kappa = 0.806
- **Random Forest**—The random forest gave the optimal prediction at 500 estimators on performing model optimization. The evaluated performance metrics were:
 - Sensitivity = 75.7%
 - Precision = 98.14%
 - F_1 score = 85.48%
 - Cohen's kappa = 0.841
- **KNN**—We implemented a K-nearest neighbors classifier to predict our seizures. On performing parameter tuning the optimal classification is obtained at five nearest neighbors, Manhattan distance, and weights varying by distance. The evaluated metrics were:
 - Sensitivity = 78.57%
 - Precision = 96.49%
 - F_1 score = 86.61%
 - Cohen's kappa = 0.853
- **LSTM**—The implemented neural network is an application of a recurrent neural network with an architecture capable of learning from past experiences. The model was trained for 100 epochs to obtain the results. The evaluated metrics were:
 - Sensitivity = 78.57%
 - Precision = 91.66%
 - F_1 score = 84.61%
 - Cohen's kappa = 0.83

The performance of all four machine learning classifiers is tabulated above in Table 5. Figure 4 shows a line plot comparing the performance of all classifiers. It is clear from the results that out of all classifiers K-nearest neighbor gives the best results as it performs better in regards to sensitivity, F_1 score, and Cohen's kappa. Hence KNN is the best classifier in this algorithm.

We estimated an average epileptic seizure prediction time of 6.009 s, while a maximum prediction time of 23.59 s was observed. The average sensitivity was 76.425% and the highest sensitivity was 78.57%. The average combined accuracy (F_1 measure) is computed as 84.62% with the maximum combined accuracy being

Table 5 Comparison of machine learning classifiers' performance classifier

	Training time (s)	Testing time (s)	Sensitivity (%)	Precision (%)	F_1 measure (%)	Cohen's kappa
SVM	0.0378	0.0039	72.85	94.44	82.25	0.806
Random forest	1.5996	0.3211	75.71	98.14	85.48	0.841
KNN	0.0045	0.1219	78.57	96.49	86.61	0.853
LSTM	29.64	23.59	78.57	91.66	84.61	0.83

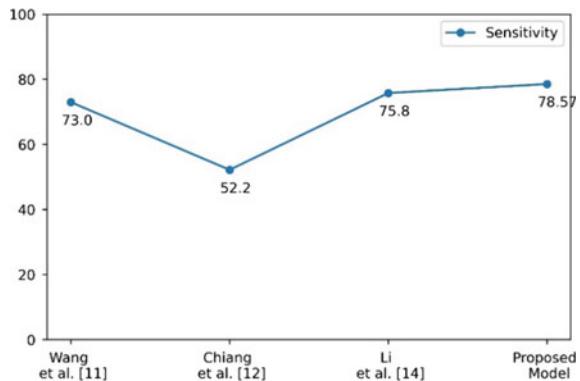
**Fig. 4** Comparison of classifier performance

86.14%. Maximum Cohen's kappa coefficient attained is 0.853, while the average Cohen's kappa coefficient is 0.8325. Table 6 and Fig. 5 demonstrate a juxtaposition between the results produced by our proposed model and those generated by other

Table 6 Model Comparison with previous literature

Model	Dataset	EEG signal type	Maximum sensitivity (%)
Wang et al. [5]	Long-term continuous intracranial EEG dataset	Intracranial EEG	73
Chiang et al. [8]	CHB-MIT + NTUH	Scalp EEG	52.2
	Freiburg EEG Database	Intracranial EEG	74.2
Li et al. [7]	Freiburg EEG Database	Intracranial EEG	75.8
Proposed model	CHB-MIT	Scalp EEG	78.57

Fig. 5 Model comparison with previous literature



models. It is fairly conspicuous from the juxtaposition that our model achieves a better result in reference to sensitivity and combined accuracy as compared to Chiang et al. [8], Wang et al. [5] and Li et al. [7].

4 Conclusion

The CHB-MIT database, which was obtained by placing electrodes on the scalp of pediatric subjects to predict seizures, was used in this study. We proposed an EEG channel selection algorithm based on Boruta [9] study, which helped us tackle various problems like low-channel capacity and lack of computational power. We proposed a feature extraction and feature selection algorithm that not only enhanced model performance by meticulously preprocessing the EEG data but also helped design a non-patient-specific approach to seizure prediction. Compared to other models, the suggested solution when tested on the dataset showed better results in terms of sensitivity and combined accuracy. Upon the implementation of the proposed model on the dataset, we achieved a maximum of 78.57% sensitivity coupled with an average prediction time of 6.009 s. The usage of the Boruta algorithm for channel and feature selection improves the model on a fundamental scale and sets up for major future improvements. Boruta eliminates the need to manually decide on a threshold metric for channel and feature selection. It takes into account multi-variable relationships, and can easily deal with the random nature of EEG signal on its own, thus paving the way for a robust and independent model in the future. Furthermore, our approach will allow doctors and medical staff to better understand the behavioral patterns of epileptic seizures in humans, i.e., it will help unlock the inner workings of the synapses and circuits that trigger seizures. This can buy doctors the time they need to effectively treat patients with the help of medication, before the onset of the seizure.

In future work, better preprocessing of the EEG signals can definitely facilitate sensitivity of seizure prediction to rise. One of the ways this can be achieved is by focusing on the EEG signals produced by the frontal lobe. In our research, we

found that out of the six best-performing channels, five have electrodes placed on the frontal lobe. In addition to this, another aspect which can be focused upon is the performance of extracted statistical features. In our research only variance and power spectral density showed a positive correlation to the occurrence of seizures; however, the rest of extracted features performed admirably in previous literature. The only difference is that our model was non-patient-specific as compared to other works. Thus, the effectiveness of various statistical features vis-à-vis non-patient-specific algorithms can be explored. Another path that could be explored is by pivoting the research around signal modalities, i.e., the preictal and postictal states of a human brain can be studied with greater scrutiny to understand the intricacies of the seizure triggering mechanism in the brain [14]. Preprocessing can also be enhanced by including adaptive windows [15]. Furthermore, one can also explore the possibility of developing an online, real-time model to predict seizures as this would be the more accurate and effective in preventing seizures.

References

1. Viglione, S.S., Walsh, G.O.: Proceedings: epileptic seizure prediction. *Electroencephalogr. Clin. Neurophysiol.* **39**(4), 435–436 (1975)
2. Wang, S., Chaovallitwongse, W. A., Wong, S.: A novel reinforce-ment learning framework for online adaptive seizure prediction. In: 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 499–504. IEEE (2010, December)
3. Ibrahim, F., El-Gindy, S.A.E., El-Dolil, S.M., El-Fishawy, A.S., El-Rabaie, E.S.M., Dessouky, M.I.,... & Abd El-Samie, F.E.: A statistical framework for EEG channel selection and seizure prediction on mobile. *Int. J. Speech Technol.* **22**(1), 191–203 (2019)
4. Schelter, B., Feldwisch-Drentrup, H., Ihle, M., Schulze-Bonhage, A., Timmer, J.: Seizure prediction in epilepsy: From circadian concepts via probabilistic forecasting to statistical evaluation. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 1624–1627. IEEE (2011, September)
5. Wang, S., Chaovallitwongse, W.A., Wong, S.: Online seizure prediction using an adaptive learning approach. *IEEE Trans. Knowl. Data Eng.* **25**(12), 2854–2866 (2013)
6. Usman, S.M., Usman, M., Fong, S.: Epileptic seizures prediction using machine learning methods. *Comput. Math. Methods Med.* **2017** (2017)
7. Li, S., Zhou, W., Yuan, Q., Liu, Y.: Seizure prediction using spike rate of intracranial EEG. *IEEE Trans. Neur. Sys. Rehabil. Eng.* **21**(6), 880–886 (2013)
8. Chiang, C.Y., Chang, N.F., Chen, T.C., Chen, H.H., Chen, L.G.: Seizure prediction based on classification of EEG synchronization patterns with on-line retraining and post-processing scheme. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 7564–7569, September 2011. IEEE (2011)
9. Kursa, M. B., Jankowski, A., Rudnicki, W. R. (2010). Boruta—a system for feature selection. *Fundam. Inform.* **101**(4), 271–285
10. Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G.,... & Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101**(23), e215–e220 (2000)
11. Bao, F.S., Liu, X., Zhang, C.: PyEEG: An open source python module for EEG/MEG feature extraction. *Comput Intell Neurosci* **2011**, 1–7 (2011)
12. Temko A., Thomas E., Marnane W., Lightbody G., Boylan G.: EEG-based neonatal seizure detection with support vector machines, *Clin. Neurophysiol.* **122**(3), 464–473 (2011)

13. Blanco, S., Garay, A., Coulombe, D.: Comparison of frequency bands using spectral entropy for epileptic seizure prediction. *ISRN Neurol.* **2013** (2013)
14. Nagaraj, V., Lee, S., Krook-Magnuson, E., Soltesz, I., Benquet, P., Irazoqui, P., Net-off, T.: The future of seizure prediction and intervention: Closing the loop. *J. clin. neurophysiol. Official Publication Am. Electroencephalogr. Society* **32**(3), 194 (2015)
15. Lan, K., Fong, S., Song, W., Vasilakos, A.V., Millham, R.C.: Self-adaptive pre-processing methodology for big data stream mining in internet of things environmental sensor monitoring. *Symmetry* **9**(10), 244 (2017)
16. Shaw, R.N., Saikia, A.: Deep learning and its importance for early signature of neuronal disorders. In: 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, pp. 1–5 (2018). <https://doi.org/10.1109/ccaa.2018.8777527>

Adaptive Fuzzy Logic Models for the Prediction of Compressive Strength of Sustainable Concrete



Chakshu Garg, Aman Namdeo, Abhishek Singhal, Priyanka Singh[✉], Rabindra Nath Shaw, and Ankush Ghosh

Abstract Sustainable growth has encouraged the utilisation of waste materials in conventional concrete for replacement. This study focuses on concrete produced by partially replacing cement and sand with waste products, like slag and fly ash. The application of these materials in concrete lowers the global energy demand and also saves money on the verge of depletion. These materials provide increased mechanical and durability properties as well as a wide range of advantages, including decreased strain on natural resources, and a lower carbon footprint. Experimental work pertaining to concrete contributes to the waste of resources, time and money. Over the last four decades, the development of methods for seeking optimal mixing proportions has been the focus of research. Several researchers have worked in recent years to establish reliable concrete models of compressive force prediction. The prediction of compressive strength of concrete is therefore an active research area. An alternative approach that used machine learning has recently gained momentum in the field of civil engineering. Machine learning is a soft computing mechanism that embodies the characteristics of the human brain, learns from prior circumstances and adapts without any restrictions to new environments. In this research work, a model has been proposed to predict the compressive strength of concrete comprising slag and fly ash as partial substitutes. The first section encompasses a brief summary of the works done by different researchers in this field, and the factors affecting the compressive strength of concrete. The next segment elaborates upon fuzzy logic

C. Garg · A. Singhal

Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University, Noida, Uttar Pradesh, India

A. Namdeo · P. Singh

Department of Civil Engineering, Amity School of Engineering and Technology, Amity University, Noida, Uttar Pradesh, India

R. N. Shaw

Department of Electrical, Electronics and Communication Engineering, Galgotias University, Greater Noida, India

e-mail: r.n.s@ieee.org

A. Ghosh (✉)

School of Engineering and Applied Sciences, The Neotia University, Sarisha West Bengal, India

and the proposed model used to predict the compressive strength of different design mixes. Thereafter the results are compared and evaluated. The objective of this study is to develop a model that can be deployed to predict the compressive strength of different types of concrete mixes.

Keywords Fuzzy logic · Sustainable concrete · Compressive strength · Machine learning · Soft computing

1 Introduction

Development and growth in infrastructure go hand in hand. India is a developing country, so there is a great demand for concrete which is required to be matched up. Cement is the principal constituent employed for producing concrete. It is a finely ground, non-metallic and an inorganic powder which hardens and sets on mixing with water. Though cement is used abundantly in the construction industry, the cement manufacturing industry is under critical observation these days due to the massive emissions of carbon dioxide. To mitigate this problem we need to produce concrete which is eco-friendly. But, the process of manufacturing sustainable concrete is a massive challenge. There aren't any limpid lines established to utilise a specific type of substance as a partial substitute. Neither are the ratios of the design mix known at large nor are the mechanical properties. Moreover, some waste substances are accessible, others aren't. Some are appropriate for certain applications whilst others are a misfit. Hence extensive research and analysis of the substances that can be used as a partial substitute of cement and sand is needed. The properties of the substitutes in concrete have to be examined and the appropriate ratios for the design mix are to be deduced. Thereafter, we need to deploy machine learning models to determine the mechanical properties of concrete. This is required because for every design mix there are distinct mechanical properties. Deriving these mechanical properties experimentally requires abundant resources and time. Therefore, there is a dire need to utilise computational methods. In this work, a fuzzy model to predict the compressive strength of concrete has been proposed.

2 Related Work

Different studies done by researchers have elaborated upon the usage of slag and fly ash in concrete.

Pulverised fuel ash or fly ash is derived from the combustion of coal. The production of fuel ash has increased many folds in nearly three decades. But the disposal of fuel ash is a problem to ponder on. The produced bottom ash contaminates the atmosphere and water bodies. Pulverised fuel ash has a specific gravity of 2.09 g/cm^3 ,

specific surface of $3355 \text{ cm}^2/\text{g}$ and loss on ignition value of 1.66 [1]. The addition of fly ash to concrete improves its workability, ultimate strength and durability, simultaneously reducing water demand, heat of hydration and permeability. Concrete containing fly ash has low early age strength [2, 3]. The British Standards and the Indian Standards [3] support utmost 35% replacement to be efficient. According to Oner et al. [1], fly ash can be used up to 40% as a replacement of cement, whereas British Standards Institution [4] establishes a usage of 15–20% to be optimum. Bendapudi [5] and Malhotra [6] studied extensively about the usage of fly ash in concrete and claimed 50 and 56% to be the adequate replacement level, respectively.

The process of manufacturing iron in blast furnaces results in a by-product called ground granulate blast furnace slag. Slag, floating on the iron in the furnace, is quenched in lots of water so that the cementitious characteristics of the slag is optimised. GGBFS is white in colour, has a specific gravity of 2.9, a bulk density of 1200 kg/m^3 and a fineness of $350 \text{ m}^2/\text{kg}$. The use of GGFBS not only improves workability, chloride ingress resistance and sulphate attack resistance but also the risk posed by thermal cracking is reduced. Cement can be replaced by GGBS directly by 30% to an extent of 85%. Generally, it is replaced from 40 to 50% but when early age strength is required 20–40% replacement level is preferred. While structures like sewage plants require higher durability, 50–70% cement is preferred to be substituted, 66–80% is the optimum replacement level to resist sulphate attacks or chloride ingress [7, 8].

Many researchers have also depicted the usage of different machine learning models to predict the compressive strength of concrete.

For estimating the compressive strength of high-performance concrete using computational methods, different researchers opted for different techniques.

Back-propagation neural network (BPNN) technique and decisions tree algorithm are used to predict the compressive strength of high-performance concrete. It is plied with a dataset encompassing 300 instances and utilised RMSE, MAE and R values as a measure of performance. The inputs given to the prediction model were water, coarse aggregates, fine aggregates, cement, superplasticiser, fly ash, blast furnace slag and curing age. It was established that the decision tree model manifested better outcomes.

Some researchers plied three different machine learning algorithms to derive the compressive strength of concrete. They administered back-propagation neural nets, support vector machine (SVM) and decision tree to acquire the desired results. The dataset was leveraged from UCI constituting 1030 instances. It comprised eight input parameters to predict the compressive strength. It was witnessed yet again that decision tree demonstrated preferable results [9, 10].

While Chithra et al. [11] deployed back-propagation neural networks (BPNN), Behnoor et al. [12] and Han et al. [13] put to use decision trees for the task of estimation.

Yuan et al. used an adaptive network-based fuzzy inference system (ANFIS) to estimate the compressive strength of concrete containing cement, slag, fly ash, water, fine aggregate, coarse aggregate and superplasticiser [14].

3 Fuzzy Logic

Fuzzy set theory was given by Zadeh in 1965. It is a technique to deal with vague, imprecise and uncertain data [15, 16]. Therefore, a model developed using fuzzy logic can be deployed if the data is not available or is available in insufficient quantities [17–19]. It is advantageous to use fuzzy logic systems over other soft computing techniques as it barely relies on historical values.

Vague statements and imprecise data act as an input for the fuzzy logic system and output is produced in the form of decisions as depicted in Fig. 1.

Input is mapped to the output with the help of four modules fuzzification, rule base or the knowledge base, inference engine and the defuzzification, as shown in Fig. 2 [20–22].

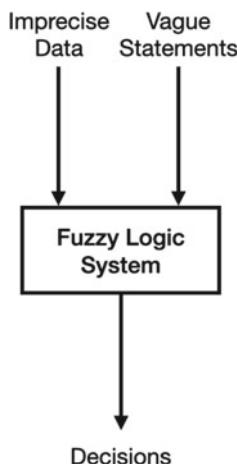


Fig. 1 Fuzzy logic system

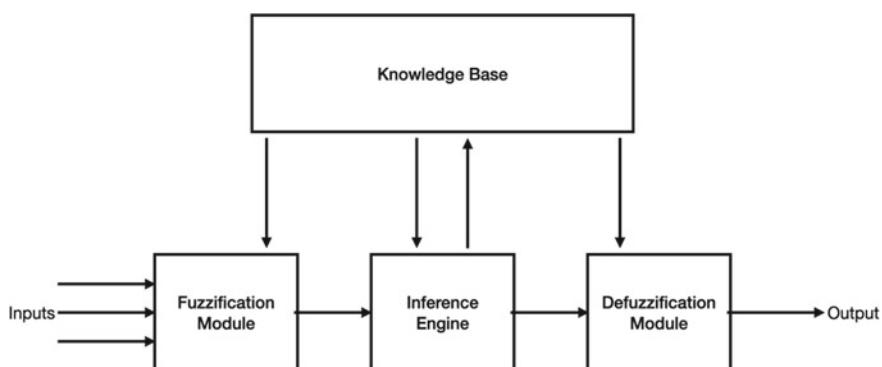


Fig. 2 Fuzzy model

Initially, the crisp sets are covered into fuzzy sets with the help of fuzzification. Next, the inference engine processes the fuzzy values based upon the knowledge base. Thereafter, the processed data is transformed into crisp sets by the process of defuzzification to obtain the output.

4 Proposed Fuzzy Model

The model proposed considers seven parameters to determine the compressive strength of 28-day concrete. Cement, slag, fly ash, water, superplasticiser, coarse aggregate and fine aggregate are the variables that impact the compressive strength of concrete as shown in Fig. 3.

These parameters were categorised into different fuzzy sets as mentioned in Table 1 and the output was classified as very low, low, moderate, high and very high. To fuzzify these variables, three different membership functions, namely triangular membership function (trimf), trapezoidal membership function (trapmf), Gaussian membership function (gaussmf), were used. Subsequently, a rule base containing 86 rules was fired to acquire the desired results using MATLAB as shown in Fig. 4. Furthermore, some rules proposed are presented in Table 2.

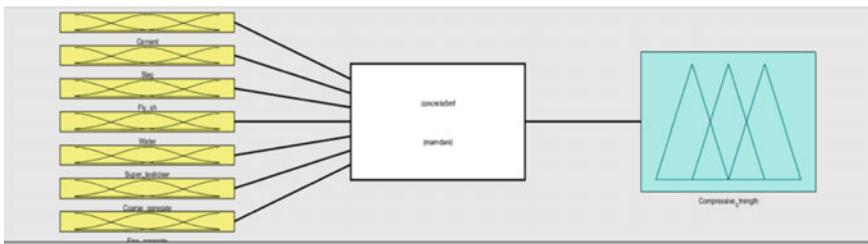


Fig. 3 Proposed model with 7 inputs, 1 output, and 86 rules

Table 1 Fuzzy sets

Parameter	Fuzzy set
Cement	Very low, low, moderate, high, very high
Slag	Low, moderate, high, very high
Fly ash	Very low, low, moderate, high, very high
Water	Low, moderate, high
Superplasticiser	Low, moderate, high
Coarse aggregate (CA)	Very low, low, moderate, high, very high
Fine aggregate (FA)	Low, moderate, high, very high

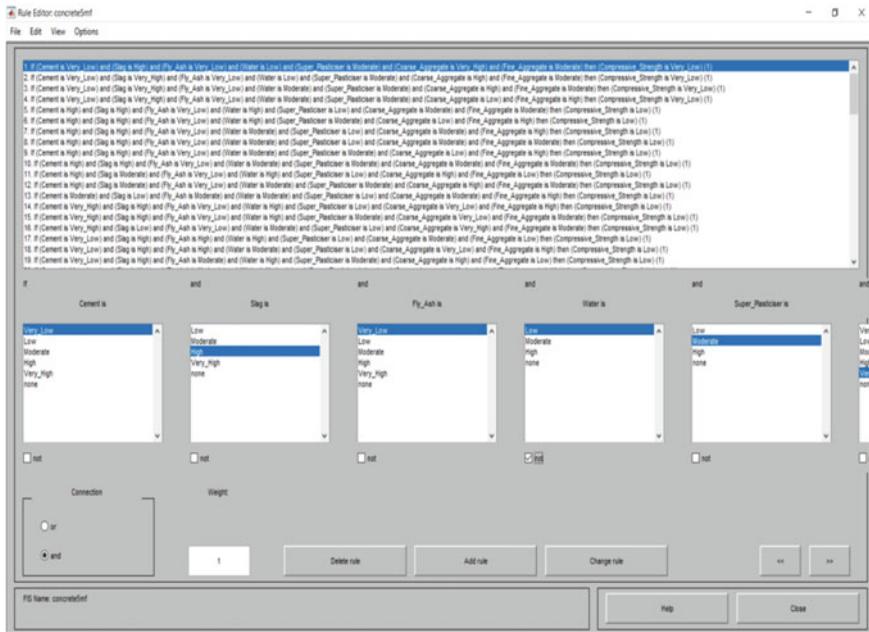


Fig. 4 Rule base

Table 2 Some proposed rules

Cement	Slag	Fly ash	Water	Superplasticiser	Coarse aggregate	Fine aggregate	Compressive strength
Very low	High	Very low	Low	Moderate	Very high	Moderate	Low
Very low	Very high	Very low	Low	Moderate	High	Moderate	Very low
Very low	Very high	Very low	Moderate	Moderate	High	Moderate	Very low
Very low	Very high	Very low	Moderate	Moderate	Low	High	Very low

Initially, the triangular membership function (trimf) was put to use for fuzzification of all the parameters as seen in Fig. 5. Next, a rule base was fired and the estimated compressive strength for different combinations of input parameters was observed. The rule viewer depicted in Fig. 6 shows the compressive strength of concrete for the input (cement: 154, slag: 112, fly ash: 144, water: 220, superplasticiser: 10, coarse aggregate: 923, fine aggregate: 658) to be 26.2.

Next, the trapezoidal membership function (trapmf) was played for fuzzification as shown in Fig. 7. The rule base was fired, and using a rule viewer as seen Fig. 8,

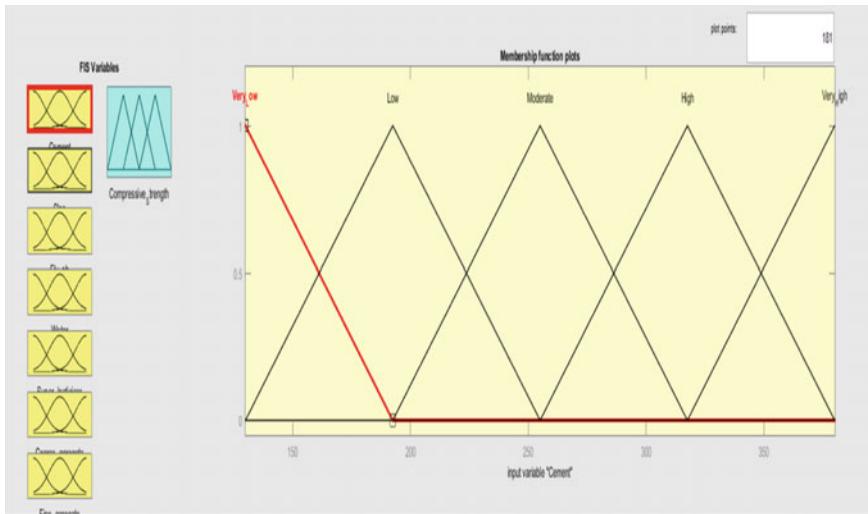


Fig. 5 Triangular membership function

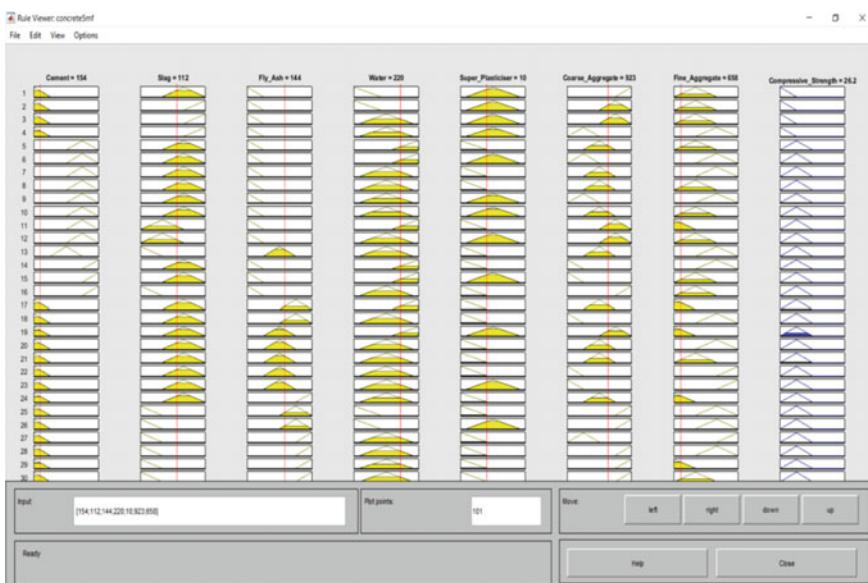


Fig. 6 Rule viewer for triangular membership function

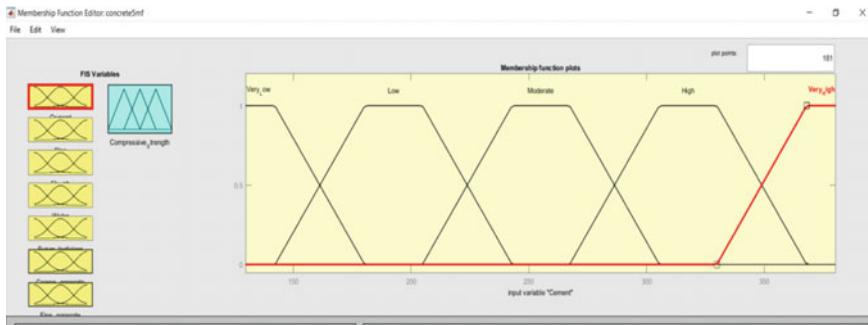


Fig. 7 Trapezoidal membership function

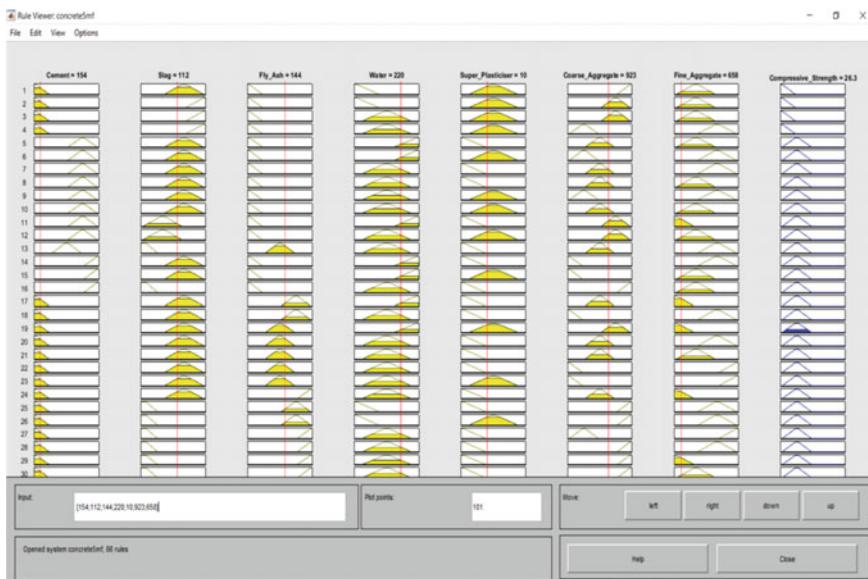


Fig. 8 Rule viewer for trapezoidal membership function

the compressive strength of concrete (28 days) was observed to be 26.3 for the same set of input values.

Thereafter, fuzzification was done using the Gaussian membership function (gaussmf) as depicted in Fig. 9 and the compressive strength was noted to be 26.5 from the rule viewer shown in Fig. 10.

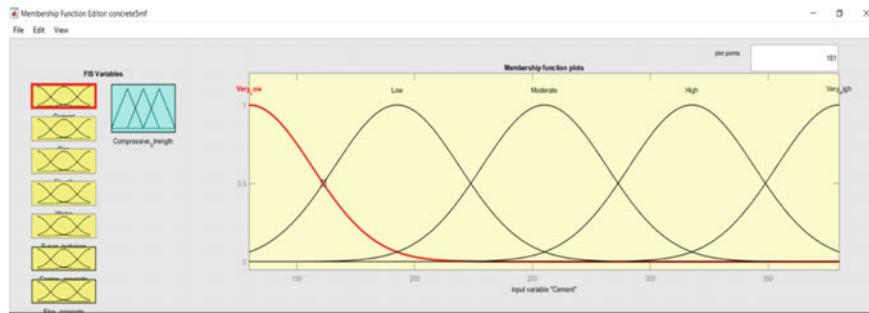


Fig. 9 Gaussian membership function

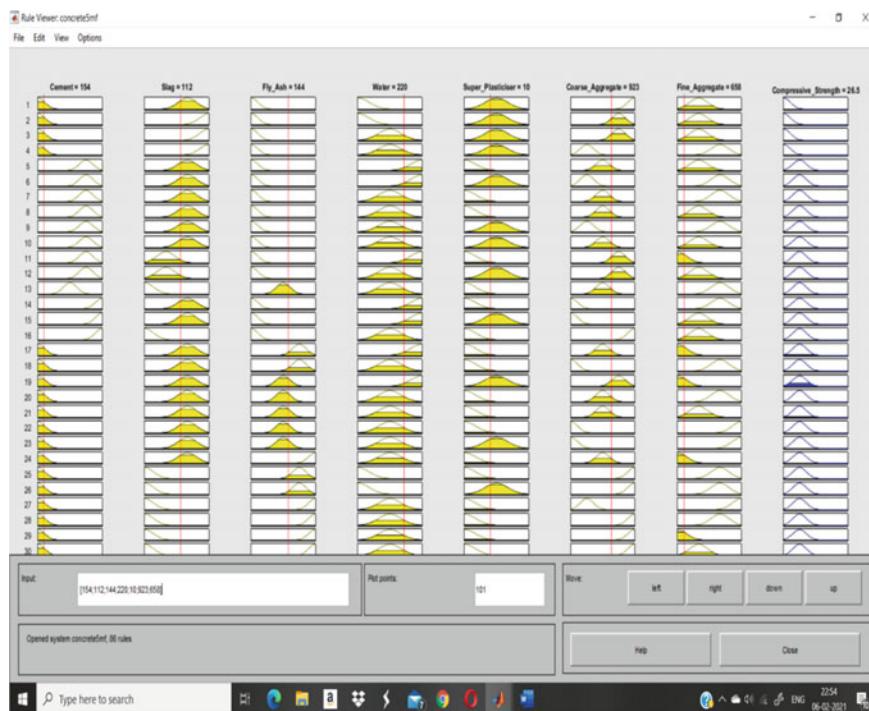


Fig. 10 Rule viewer for Gaussian membership function

5 Results and Discussions

This paper discusses the model proposed to predict the compressive strength of concrete using fuzzy logic. The compressive strength of concrete is dependent on the quantity of cement, slag, fly ash, water, superplasticiser, coarse aggregate and fine aggregate it possesses. The proposed model is efficient and is dependent on all these parameters.

The work also comprehends the output in the case of three different membership functions, namely triangular membership function (trimf), trapezoidal membership function (trapmf) and Gaussian membership function (gaussmf). The predicted value for all the three membership functions is compared with the experimental value and depicted in Table 3. The mean absolute error for all three cases has also been calculated. It is seen that 3.98, 3.37 and 3.75 are the mean absolute errors estimated for the proposed fuzzy model using triangular membership function (trimf), trapezoidal membership function (trapmf) and Gaussian membership function (gaussmf) for fuzzification, respectively.

6 Conclusions

Cement production and sand extraction lead to many environmental issues. For sustainability, there is a need to encourage the usage of green concrete. The production of green concrete requires the usage of different waste materials like slag and fly ash. The incorporation of these materials requires extensive research. Also, the experimental work pertaining to concrete requires a huge amount of resources, time and expertise. Hence, a need of deploying soft computing techniques to evaluate the compressive strength of different types of concrete mixes was the need of the hour. This work proposed a fuzzy model that can be put to use for determining the compressive strength of concrete containing slag and fly ash as partial substitutes of cement. This will help all the stakeholders to use these substances extensively.

Furthermore, similar models can be utilised for different types of concretes using varied waste materials that can be used as partial substitutes of cement or sand. Fuzzy models can also be developed for predicting the mechanical strength of recycled aggregate concrete, high-performance concrete, self-compacting concrete and many more so that the goal of sustainable development could be attained.

Table 3 Comparison of different member functions

Experimental value	Value obtained using Trimpf	Value obtained using Trapmf	Value obtained using Gaussmf	Absolute error using Trimpf	Absolute error using Trapmf	Absolute error using Gaussmf
41.81	43.1	43.1	42.9	3.08538627122698	3.08538627122698	2.60703181057163
42.08	43.1	43.1	42.9	2.42395437262357	2.42395437262357	1.94866920152091
26.82	26.2	26.3	26.5	2.3117076808352	1.93885160328113	1.19313944817301
25.21	26.2	26.3	28	3.92701309004363	4.32368107893693	11.067036890123
38.86	37.5	37.5	37.5	3.49974266598044	3.49974266598044	3.49974266598044
36.59	37.5	37.5	36.5	2.48701831101394	2.48701831101394	0.245968843946433
38.46	40.8	39.9	39.3	6.08424336973479	3.74414976599064	2.18408736349454
26.02	26.3	26.3	26.6	1.076095311299	1.076095311299	2.22905457340507
28.03	26.3	26.3	27.5	6.17195861576882	6.17195861576882	1.89083125222975
28.29	26.2	26.3	27	7.38776952986921	7.03428773418169	4.55991516436903
49.3	46.5	48.8	44.4	5.67951318458418	1.01419878296146	9.93914807302231
36.19	37.5	37.5	37.5	3.6197844708483	3.6197844708483	3.6197844708483
Mean absolute error				3.97951557281901	3.36825908200941	3.74870081314037

References

1. Oner, A., Akyuz, S., Yildiz, R.: An experimental study on strength development of concrete containing fly ash and optimum usage of fly ash in concrete. *Cem. Concr. Res.* **35**(6), 1165–1171 (2005)
2. Hemalatha, T., Ramaswamy, A.: A review on fly ash characteristics—towards promoting high volume utilization in developing sustainable concrete. *J. Clean. Prod.* **147**, 546–559 (2017). <https://doi.org/10.1016/j.jclepro.2017.01.114>
3. Singh, P., Shah, N.D.: An experimental investigation on sustainable concrete with flyash and steel fibers. *Int J Civil Eng Technol* **9**(6), 1131–1140 (2018)
4. British Standards Institution, 1997a. BS 5328: Part 1, Guide to specifying concrete. British Standards Institution, 1997b. BS 8110: Part I, Structural use of concrete: code of practice for design and construction. IS-1489, 2000. IS 1489 (Part I): 1991 Portland-Pozzolana Cement specification. Indian Standards, India. Amendment no. 3
5. Bendapudi, S.C.K.: Contribution of fly ash to the properties of mortar and concrete. *Int. J. Earth Sci. Eng.* **4**(6 SPL), 1017–1023 (2011)
6. Malhotra, V.M.: Durability of concrete incorporating high-volume of low-calcium (ASTM Class F) fly ash. *Cement Concr. Compos.* **12**, 271–277 (1990)
7. Langley, W., Carette, C., Malhotra, V.: Structural concrete incorporating high volumes of ASTM Class F fly ash. *ACI Mater. J.* **86**, 507–514 (1989)
8. Singh, P.R., Goel, A., Thakur, S., Shah, N.D.: An experimental approach to investigate effect of steel fibers on tensile and flexural strength of fly ash concrete. *Int. J. Sci. Eng. Appl. Sci. (IJSEAS)* **2**(5), 384–392 (2016)
9. Suresh, D., Nagaraju, K.: Ground Granulated Blast Slag (GGBS) in concrete—a review. *IOSR J. Mech. Civil Eng. (IOSR-JMCE)* **12**(4), Ver. VI, 76–82, July–August 2015
10. Singh, P.R., Shah, N.D.: Impact of coal combustion fly ash used as a binder in pavement. *Civ. Eng. Environ. Tech.* **1**, 57–60 (2014)
11. Chithra, S., Kumar, S.R.R.S., Chinnaraju, K., Ashmita, F.A.: A comparative study on the compressive strength prediction models for high performance concrete containing nano silica and copper slag using regression analysis and artificial neural networks. *Constr. Build. Mater.* **114**, 528–535 (2016). <https://doi.org/10.1016/j.conbuildmat.2016.03.214>
12. Behnood, A., Behnood, V., Modiri, M., Esat, K.: Prediction of the compressive strength of normal and high-performance concretes using M5P model tree algorithm. *Constr. Build. Mater.* **142**, 199–207 (2017). <https://doi.org/10.1016/j.conbuildmat.2017.03.061>
13. Han, Q., Gui, C., Xu, J., Lacidogna, G.: A generalized method to predict the compressive strength of high-performance concrete by improved random forest algorithm. *Constr. Build. Mater.* (2019). <https://doi.org/10.1016/j.conbuildmat.2019.07.315>
14. Yuan, Z., Wang, L., Ji, X.: Advances in engineering software prediction of concrete compressive strength: research on hybrid models genetic based algorithms and ANFIS. *67*, 156–163 (2014). <https://doi.org/10.1016/j.advengsoft.2013.09.004>
15. Singh, P., Khaskil, P.: Prediction of compressive strength of green concrete with admixtures using neural networks. In: 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, India, pp. 714–717 (2020). <https://doi.org/10.1109/gucon48875.2020.9231230>
16. Chou, J., Ph, D., Chiu, C., Ph, D., Farfoura, M., Al-taharwa, I.: Optimizing the prediction accuracy of concrete compressive strength based on a comparison of data-mining. *Techniques* **25**, 242–253 (2011). [https://doi.org/10.1061/\(ASCE\)CP.1943-5487](https://doi.org/10.1061/(ASCE)CP.1943-5487)
17. Deepa, C., Sathiya Kumari, K., Sudha, V.P.: Prediction of the compressive strength of high performance concrete mix using tree based modeling. *Int. J. Comput. Appl.* **6**, 18–24 (2010). <https://doi.org/10.5120/1076-1406>
18. Mandal, S., Biswas, S., Balas, V.E., Shaw, R.N., Ghosh, A.: Motion prediction for autonomous vehicles from Lyft dataset using deep learning. In: 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, pp. 768–773 (2020). <https://doi.org/10.1109/iccca49541.2020.9250790>

19. Zadeh, L.A.: Fuzzy logic **21**(4), 83–93 (1988). <https://doi.org/10.1109/2.53>
20. Novák, V., Perfilieva, I., Dvořák, A.: Insight into fuzzy modeling what is fuzzy modeling, 3–10. <https://doi.org/10.1002/9781119193210.ch1>
21. Mandal, S., Balas, V.E., Shaw, R.N., Ghosh, A.: Prediction analysis of idiopathic pulmonary fibrosis progression from OSIC dataset. In: 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, India, pp. 861–865 (2020). <https://doi.org/10.1109/gucon48875.2020.9231239>
22. Liu, Z., Li, H.-X.: A probabilistic fuzzy logic system for modeling and control, **13**(6), 0–859 (2005). <https://doi.org/10.1109/tfuzz.2005.859326>

Unique Action Identifier by Using Magnetometer, Accelerometer and Gyroscope: KNN Approach



Prajyot Palimkar, Varnica Bajaj, Arpan Kumar Mal, Rabindra Nath Shaw, and Ankush Ghosh

Abstract In today's world, where technology is advancing every single day, new methodologies are being developed, and are brought in everyday use making our lives simpler, faster, safer, and powerful. Similarly, Human Activity Recognition (HAR) is getting more popular with all the revolutions made in the technologies. Sensor Network Technology is used in industrial applications, smart homes and system. A massive amount of data can be obtained from these sensors which are linked to the human body. Recognition of Human Activities using these sensors, and wearable technologies has been actively studied. Behavior Recognition seeks to distinguish one or more people's activities and goals through a collection of observations on the actions and environmental conditions of the person. Health surveillance, aged treatment, and plenty of other domains can be used to automatically understand the behavioral context. An existing dataset consisting of 10 subjects (5 females, 5 males) is being used in the paper, which incorporates both young and old volunteers between 19 and 60 years of old with weights ranging from 55 to 85 kg. The dataset reflects motion data collected when subjects are engaged in 11 separate (static and dynamic) smart home activities: computer usage (1 min), telephone conversation (1 min), vacuum cleaning (1 min), book reading (1 min), TV watching (1 min), ironing (1 min), walking (1 min), exercise (1 min), cooking (1 min), drinking (20 times), hair brushing (1 min) (20 times). Most of the activities are similar because of the multi sensor environment which makes it more difficult. Using three tri axial IMU (inertial measurement unit), Magnetometer, Accelerometer, Gyroscope sensors attached to

P. Palimkar · A. K. Mal · A. Ghosh (✉)

School of Engineering and Applied Science, The Neotia University, Kolkata, West Bengal 743368, India

V. Bajaj

School of Computer Science and Engineering, Galgotias University, Greater Noida, Uttar Pradesh 203201, India

R. N. Shaw

Department of Electrical Electronics & Communication Engineering, Galgotias University, Greater Noida, India

e-mail: r.n.s@ieee.org

the subject area of the hand, chest, and thigh, using Machine Learning we introduced a better model to prognosticate the human activity. We have applied various machine learning classification algorithms like Random Forest Classifier, K-Nearest Neighbors, Decision Tree, Multi-layer Perceptron Classifier, Extra Tree Classifier, Ensemble Extra Trees Classifier, Label Propagation and Label Spreading. The experimental results are tabulated and analyzed, and might be effectively accustomed to recognize human activities in terms of efficiency and accuracy.

Keywords Human activity recognition · KNN · Classification algorithm · Sensors · Magnetometer · Accelerometer · Gyroscope

1 Introduction

Nowadays, technologies are updating very fast that we can't think of. Today, if we imagine anything, at the subsequent moment that imagination turns into a brand new technology. Some modern sensors are behind this evolution of technology. Some of the trendy sensors are GPS sensor, light sensor, finger print sensor, proximity sensor etc. These sorts of all sensors are often found in every advance system. This kind of powerful sensor had been deployed within the systems to play various task in a very skillful manner. One among those tasks is recognizing human activities (like- using computer, drinking, reading books, walking and many more) by placing some wearable sensors (like magnetometer, accelerometer, gyroscope) in contact with human body and utilizing the data that's taken by the sensor. For detecting the human activity machine learning algorithms are employed. This kind of detecting system can be employed in various places like, in smart homes, in hospitals and many more. This system is especially useful for senior citizens. By this system one can also monitor the health condition of senior citizens.

Artificial Intelligence (AI) has such a large amount of applications (like- machine vision, Speech recognition and many more). Machine learning is one of these applications. This application provides systems the ability to learn by itself and improve the skill from previous experience. For this ability no complex programming is required. By the use of this application, computer programs are developing. After developing, system can access data and use data and able to learn automatically. Similarly, during this research, by the assistance of machine learning, system can able to learn itself by monitoring the human activities again and again. For this machine needs to be trained like human beings.

In this world where a snug & healthy life style is a common thing, here it's very important to keep watch on every movements of a person continuously. Detection of human activities is extremely important for this survey, because it is necessary to keep watching the daily routine of a person, as an example whether the person is doing exercise, or walking or ironing or reading book or using computer and many more. The objective of this research is to keep a check on the daily home activity of a human. In this research, it is required monitor, how an individual behaves while using

a computer, reading books, drinking, cooking, watching a TV, phone conversation, vacuum cleaning, brushing hair.

All information of this research is discussed briefly below: In Sect. 2 literature review is mentioned. Description of dataset is discussed in Sect. 3. In Sect. 4 overview of whole system and applied methodology is explained. Proposed model is discussed in Sect. 5. Final output, results and accuracies of the research is explained in Sect. 6. In Sect. 7 conclusion part of the research is mentioned. References, used in research is mentioned at last.

2 Literature Review

The wearable devices are implanted with variable sensors, and are affordable too. And these sensors have helped, and made in more easy and compatible by opening different methods in the areas of data mining and data analytics. The presence of these sensors has motivated individuals to perform various tasks, analyze at the data, works thereon accordingly. One of the tasks of this hot topic that has been the topic of study i.e. the movement detection of the individuals by recognizing, detecting and classifying the activities of human which is termed as movement recognition or activity recognition. Accelerometer is one of the common sensors which is used in such kinds of activities because it gives data with better accuracy [1]. Gyroscope is additionally also one of the most effective sensor as it gives the accurate position in all three axes. Aside from this Magnetometer is playing a vital role in explaining the strength of magnetic field. In the field of activity detection, different authors have adapted different reasonable techniques, and different frameworks to classify the complex patterns of human activities. Each study requires different sensors at different positions, and is tested against the various machine learning algorithms by producing varied results for each study. Anguita et al. presented “Human Activity Recognition on Smartphones Using a Multiclass Hardware Friendly Support Vector Machine”. At the point where these sensors are linked to the subject body they check all the activities and physiological signs continuously. He provided a system for understanding the human physical activities. As these cell phones are constrained as far as vitality and processing power, equipment benevolent methodology is proposed for multi-class characterization. This strategy adjusts the standard Support Vector Machine (SVM) and endeavors fixed-point number juggling for computational cost decrease. An examination with the customary SVM demonstrates a noteworthy improvement as far as computational expenses while keeping up comparative precision, which can add to grow increasingly maintainable frameworks for AmI [2]. Varkey et al. presented a “Human motion recognition using a wearable wireless based system” [3]. A. Jalal et al. presented hierarchical features of human motions and using Linear Support Vector Machine (LSVM) for human behavior classification [4]. Wu et al. introduced a wearable system using EMG and IMU sensors for recognizing American Sign Language (ASL) in real-time [5]. Ms. S. Roobini et al. used the deep learning approach to recognize and understand the human motions. And also compared

Convolutional Neural Network (CNN) with Long-Short Term Memory (LSTM) and Recurrent Neural Network (RNN) with Long-Short Term Memory, finally proved that Recurrent Neural Network with Long Short Term Memory (RNNLSTM) provides better accuracy with lower mean absolute percentage error. Hence they suggested RNNLSTM will be used to reduce the human loss of lives in recognizing the activities of human in real world [6].

This Table 1 shows the comparative analysis of the studies that have been carried out by different authors which uses sensors in the wearable devices for recognizing the human activities.

Algorithm Used

SVM-Support Vector Machine, **MP**-Multilevel Perception, ***DT**-Decision Tables, **DT**-Decision Trees (C4.5), **KNN**-K-Nearest Neighbors, **NB**-Naive Bayes, **AdaBoost**-Adaptive Boosting, **HMM**-Hidden Markov models, **NN**-Neural Networks, **LR**-Logistic Regression, **RB**-Rule Based Classifiers, **BN**-Bayes Net, **BFT**-Best-First Tree, **KS**-K-Star, **C-Tree**-Conditional Inference Trees, **RF**-Random Forest, **ETC**- Extra Tree Classifier, ***EET**-Ensemble Extra Tree, **LP**-Label Propagation, **LS**-Label Spreading.

Activities Performed

A1-Sitting, **A2**-Standing, **A3**-Walking, **A4**-Lying Down, **A5**-Stepping Down, **A6**-Climbing Up, **A7**-Dancing, **A8**-Stairs-Down, **A9**-Running, **A10**-Stairs-up, **A11**-Brushing, **A12**-Vacuum Cleaning, **A13**-Driving, **A14**-Cycling, **A15**-Inactive, **A16**-Cooking, **A17**-Medication, **A18** - Sweeping, **A19**-Washing Hands, **A20**-Watering Plants, **A21**-Biking, **A22**-Tai Chi Movements, **A23**-Hammering, **A24**-Screwing, **A25**-Spanner Using, **A26**-Power Drill, **A27**-Using Computer, **A28**-Phone Conversation, **A29**-Reading Book, **A30**-Watching TV, **A31**-Ironing, **A32**-Exercise, **A33**-Drinking, **A34**-Brushing Bair, **A35**-Waist-Bends Forward, **A36**- Frontal Elevation of Arms, **A37**-Knees Bending (Crouching), **A38**- Jogging, **A39**- Jumping Front &Back, **A40**-Brisk Walking, **A41**-Jump-Turn-Twist, **A42**-Tango, **A43**-Violent Motion, **A44**-Slow Walk, **A45**-Fast Walk, **A46**- Sit Ups.

3 Data Description

The dataset used in this particular research is obtained from [19] by performing 11 different activities by 10 subjects (5 females, 5 male) using three tri-axial IMU (Inertial Measurement Unit) sensors attached to the subject area of the hand, chest, and thigh. The dataset reflects activity data collected when subjects are engaged in 11 separate (static and dynamic) smart home activities: computer usage (1 min), telephone conversation (1 min), vacuum cleaning (1 min), book reading (1 min), TV watching (1 min), ironing (1 min), walking (1 min), exercise (1 min), cooking (1 min), drinking (20 times), hair brushing (1 min) (20 times). Both young and old

Table 1 Comparative analysis of different works performed by various authors

Reference ID, author name and year of publication	Sensor/s used	Activities performed	Algorithm used	Accuracy (%)
Proposed model	Magnetometer, Accelerometer, Gyroscope	A27, A28, A12, A29, A30, A31, A3, A32, A16, A33, A34	RF, KNN, DT, MP, ETC, *EET, LP, LS	98.56
Randhawa et al. [7]	Magnetometer, Accelerometer, Gyroscope	A2, A3, A40, A41, A42, A43	DT, SVM	85.9
D'souza et al. [8]	Accelerometer, Gyroscope	A39, A9, A38, A14, A37, A36, A35, A10, A4, A3, A1, A2	KNN, NB, SVM, C-Tree, J48 and RF	97.15
Ronao and Cho [9]	Accelerometer	A35, A3	DL	94.79
Baya et al. [10]	Accelerometer	A7, A8, A10, A9, A44, A45	MP, SVM	91.15
Shoaib et al. [11]	Magnetometer, Accelerometer, Gyroscope	A1, A2, A3, A5, A6, A9	NB, SVM, NN, LR, KNN, RB, DT	94.5
Anguita et al. [12]	Accelerometer, Gyroscope	A1, A2, A4, A3, A6, A5	SVM	96
Mannini et al. [13]	Accelerometer	A2, A3, A1, A4, A8, A10, A9, A14	HMM	97.4
Rosati et al. [14]	Magnetometer, Accelerometer Gyroscope	A35, A3	KNN, NN, SVM, DT	97.1
Krishnan et al. [15]	Accelerometer	A4, A2, A1, A9, A14, A3, A8, A10	SVM, AdaBoost	95.35
Tapia et al. [16]	Accelerometer	A3, A1, A2, A4, A9, A8, A10, A14	DT	94.6
Kunze et al. [17]	Accelerometer	A22	KNN	86
Kumar et al. [18]	Accelerometer	A2, A3, A10, A5, A6, A46, A12, A11	*DT, DT (C4.5), KNN, SVM, NB	84

candidates are considered for this study aged between 19 and 60 with weights ranging between 55 and 85 kgs. are involved.

The dataset contains 220 inertial data sequences from three inertial measurement unit sensors carried by the body with a variable time period (between 45 and 60 s). We used 10 subjects who conducted tasks repetitively in nature to train the device. The dataset consists of 28 attributes three for every magnetometer, accelerometer and

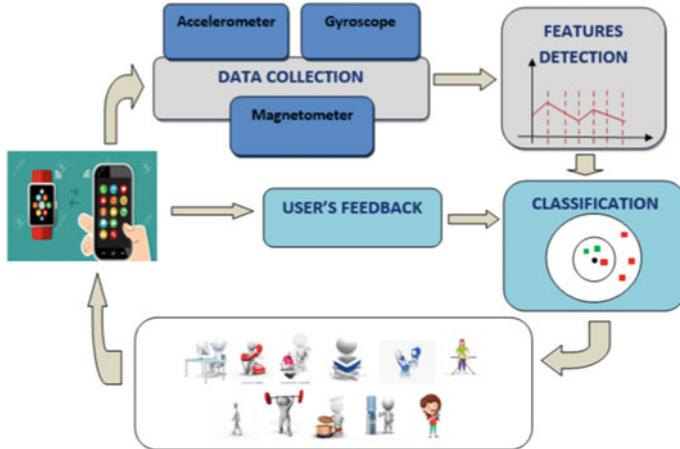


Fig. 1 Pictorial architecture of the model

gyroscope sensors, and one feature is activity performed which depends on all these 27 attributes. We used a variation of repetitive and passive activity from each activity set for testing sets. Figure 1 shows the architecture or the proposed block model for the human activity recognition system. Table 2 depicts the number of records that have been collected for each of the 10 subjects (Graph 1).

4 System Overview

Machine Learning is conquering the globe of engineering, and technologies day by day, and becoming a really big thing [19, 20]. In this study, we have proposed a system that uses wearable devices, and different machine learning algorithms are implemented in the available data. The aim of this research is to recognize various human activities from the values obtained from various sensors which are placed on person's body. In this section, we have discussed about the architecture of the model, and also the complete framework is shown within the Fig. 2. [21]

4.1 Data Collection and Processing

For the recognition and prediction of human activities the IM-WSHA dataset is collected from AIR University [22]. It contains 10 CSV files of 10 peoples which were considered in which half were males and other were females. Each Comma Separated Value (CSV) file contains all the 11 activities performed by them. The

Table 2 Description of dataset

Subject => ↓ activities	1	2	3	4	5	6	7	8	9	10	Activity total
Using computer	1192	1213	1191	1207	1159	1220	1237	1276	1160	1246	12101
Phone conversation	1208	1217	1221	1259	1222	1210	1217	1181	1171	1271	12177
Vacuum cleaning	966	1112	1135	950	1008	971	904	939	1041	895	9921
Readme book	1573	1589	1347	1707	1570	1604	1599	1631	1581	1698	15899
Watchine book	1110	1095	1072	1029	1102	1070	1138	1098	1087	995	10796
Ironing	1189	1311	1201	1206	1217	1186	1166	1150	1284	1198	12108
Waiting	1217	1141	1255	1215	1270	1301	1241	1196	1112	1200	12148
Exercise	1233	1211	1222	113B	1214	1166	1213	1228	1254	1305	12184
Cooking	1227	1372	1155	1454	1311	1276	1255	1436	1296	1154	12936
Drinking	1141	1027	a? 2	927	1037	1025	1096	1042	1150	919	10236
Brushing hair	550	559	1155	685	529	362	502	166	211	486	5205
Instances total	12606	12847	12826	12777	12639	12391	12568	12343	12347	12367	125711

dataset contains values of various sensors in all three axes while performing versatile activities [23].

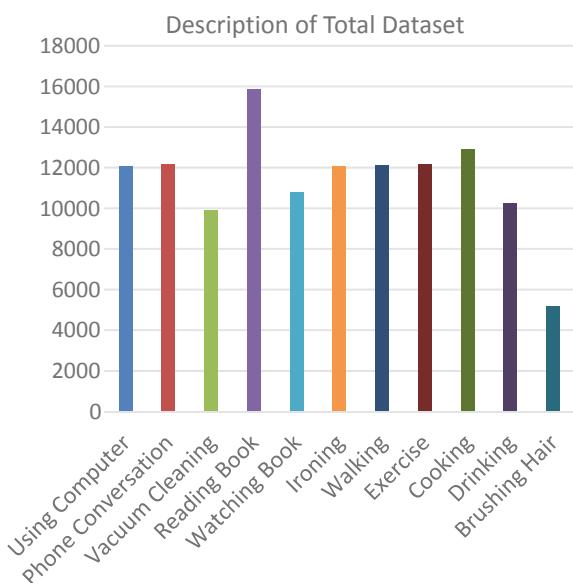
4.2 Pre-processing

For using this dataset, it has to be pre-processed so, the data of all candidates are combined in a single CSV file. Then, same dataset is splitted into training and testing dataset. Training dataset is used to establish a model so it is the 70% of total dataset. So, rest is used in testing dataset for validating the performance of the model.

4.3 Building Model

The above dataset is now fed for training and testing to different machine learning classification algorithms like Random Forest Classifier, K- Nearest Neighbors, Multi-layer Perceptron Classifier, Ensemble Extra Trees Classifier, Label Spreading and many more.

Graph 1 Description of dataset



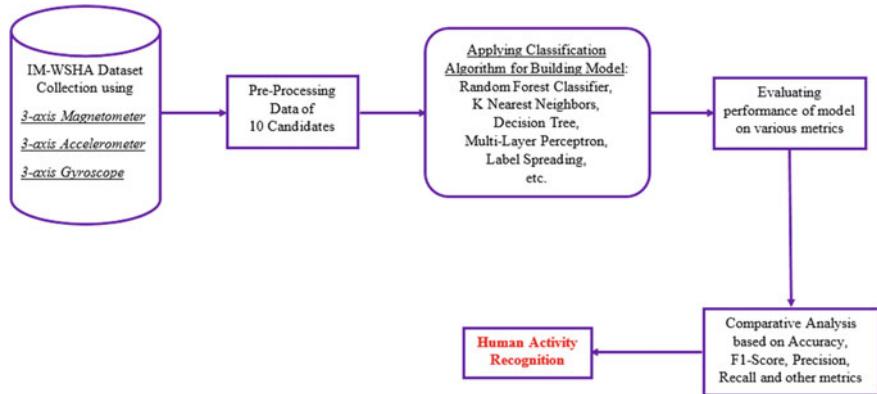


Fig. 2 Proposed methodology flowchart

4.4 Evaluation

The evaluation of various model will be done by considering several performance metrics like Accuracy, Precision, F1-Score, Recall and many others which could be obtained using Confusion Matrix. As confusion matrix shows the working of model, so it is the most important thing while considering performance of the model. Figure 3 shows the confusion matrix and some of the metrics related to it.

		PREDICTED CLASS		Recall / Sensitivity
		Positive (+ve)	Negative (-ve)	
ACTUAL CLASS	Positive (+ve)	True Positive (TP)	False Negative (FN) (Type II Error)	$\frac{TP}{(TP + FN)}$
	Negative (-ve)	False Positive (FP) (Type I Error)	True Negative (TN)	$\frac{TN}{(FP + TN)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predicted Value $\frac{TN}{(FN + TN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

Fig. 3 Confusion matrix with related metrics

4.5 A Subsection Sample

Based on the performance metrics comparative analysis are going to be done and accordingly several hyper parameters are going to be tuned to get the perfect model. So, *GridSearchCV()* chooses the different hyper parameters for optimization and provides the result which have best accuracy.

4.6 Results

By doing comparative analysis, it would be easy to elucidate which model is best for prediction of human activity. So, the model which will be best would be considered for deploying in smart watches and smart phones. Using this model, it will be better to recognize the activity performed by the human.

5 Proposed Model

Modeling

Covariance: The linear association between the two variables is measured by covariance i.e. it measures how one variable is inter related with the other one, and on what factors it depends. In simple words, it is the relation that shows how two variables are associated with one another. When using covariance in algorithm, it measures the proportion of change in two random variables. Variables which have covariance are called as correlated variables, and the ones which have zero covariance are known as uncorrelated variables.

The formula for the calculation of covariance is:

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$$

where, μ_X is mean of X and μ_Y is mean of Y, σ_{XY} denotes the covariance between variable X and Y.

Correlation: Both covariance and correlation are similar as both of them measure the deviation of random variables from the expected values.

On considering X and Y, two random variables, ρ_{XY} denotes correlation between them μ_X and μ_Y are their mean, and σ_X and σ_Y are their standard deviations, and their relation is shown below:

$$\rho_{XY} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

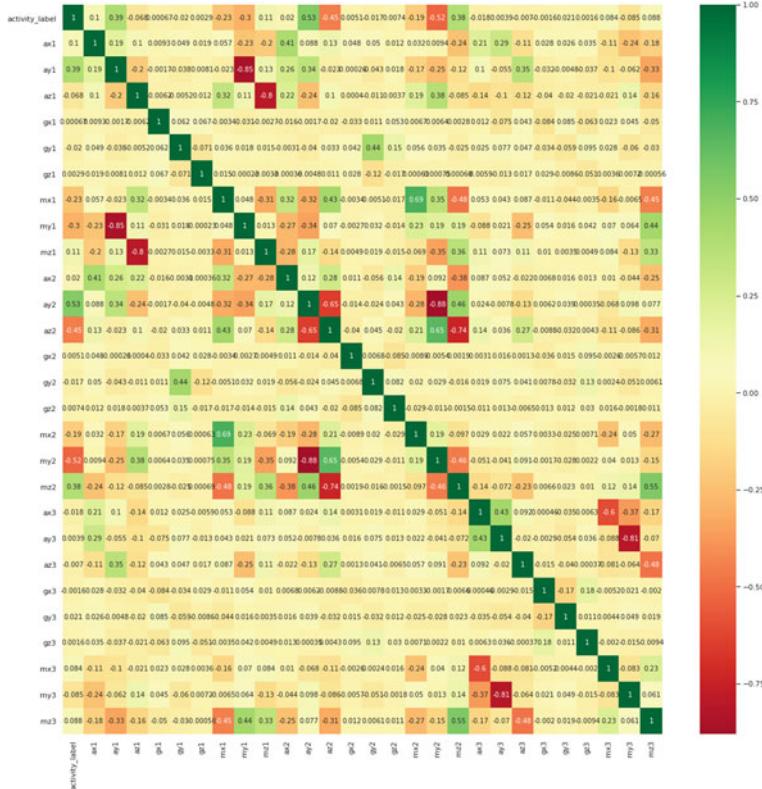


Fig. 4 Correlation between different features

Correlation is dimensionless, and covariance has units. And if on finding the value of the correlation is 0 that means there's no correlation between the two variables whereas -1 shows the perfect negative linear correlation and $+1$ depicts the perfect positive linear correlation.

Collinearity: In this one feature of the variable is very similar or related to the other feature of the variable i.e. one variable can be predicted by observing the other variables feature, and therefore the accuracy of the prediction is high. The perfect collinear exists when two variables have exact linear relationship with each other.

The Correlation of between different features are shown in Fig. 4.

5.1 Random Forest

Like K-NN, Random Forest is also a Supervised Learning algorithm, but unlike K-NN, Random Forest is used in Classification as well as in Regression problems,

but mainly it is focused on the problems associated with Classification. It created different training samples, from the training dataset provided, and then it constructs a decision tree for each training sample created, and at that time voting is done to search out the most effective predicted solution, and therefore the predicted result which has gained more votes would be the ultimate prediction result. It overcomes the problem of overfitting. And it maintains and possess a good accuracy of the dataset.

5.2 *K-Nearest Neighbors*

K-Nearest Neighbor (K-NN) is considered to be a part of Supervised Learning methodology in Machine Learning, and is one of the easiest algorithms too. It is mostly utilized in the Classification, because it stores the dataset given, and performs the operation on the dataset at the time of Classification, and it can be concluded from here that it doesn't learn anything automatically. The K-NN algorithm considers old data and accordingly treats the new data points to classify it in the particular classes. It doesn't make any kind of assumptions of the given data. Within the training period of K-NN, it stores the data, and then checks the similarities between the data set, and then will give an output which is analogous to the dataset provided.

5.3 *Decision Tree*

A decision tree is a representation of a tree like structure where it has nodes, and branches, and each node consists of an attribute, and branch depicts the test outcome. They distinguish which are the important field for prediction and classification of the given dataset. They handle both typed of the value i.e. categorical and continuous variables.

5.4 *Multi-layer Perceptron Classifier*

Multi-layer Perceptron Classifier is one of the foremost important algorithm in the field neural network. It works on making models simpler by performing difficult task like HAR. It created more models that reflects the working of human brain, and the way it is going to analyze, and perform solutions to all the difficulties, and challenges by creating realistic models of brains so that we can work on difficult models and tasks. It is performed or created in such a simplest way that it learns everything, and offers the most effective output from the given training set.

5.5 *Extra Tree Classifier*

Extra Tree Classifier is an ensemble machine learning algorithm which is also called Extremely Randomized Trees because the features and splits are selected in random way. Extra Trees perform same as Decision Tree and Random Forest but there's difference in their performances like Extra Tree Classifier shows low variance as compared to Random Forest. But Decision Trees comparatively shows higher variance as compared to other two.

5.6 *Ensemble Extra Trees Classifier*

It is kind of Random Forest, but differs while creating the decision trees. The decision tree constructed in this is created out of the training dataset that has been provided. There at each node of the decision tree a random sample of k characteristics is provided from which the decision tree gets an option as well as the idea to split divide the dataset which on the basis of the most feature using some mathematical parameters.

5.7 *Label Propagation*

Label Propagation is proposed by Xiaojin Zhu and Zoubin Ghahramani [20] in their report in 2002. As the name suggests, Label Propagation Algorithm it's the iterative algorithm, the labels are assigned to the unlabeled points as it predicts their labels by propagating the labels across the dataset. During this a little subset of data points having labels is taken into account and then throughout the algorithm it is propagated to unlabeled points.

5.8 *Label Spreading*

Label Spreading is one of the semi supervised learning model, that is for training it uses both labelled and unlabeled data. In 2003 Zhou et al. had proposed this algorithm in their one of the paper. Spreading activation networks is the experimental psychological technique from which Label Spreading is been inspired. Label Propagation is also almost similar to Label Spreading but distinguishing feature of Label Spreading is that, it uses affinity matrix which is based on normalized Laplacian graph and soft clamping across the labels.

$$L = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$$

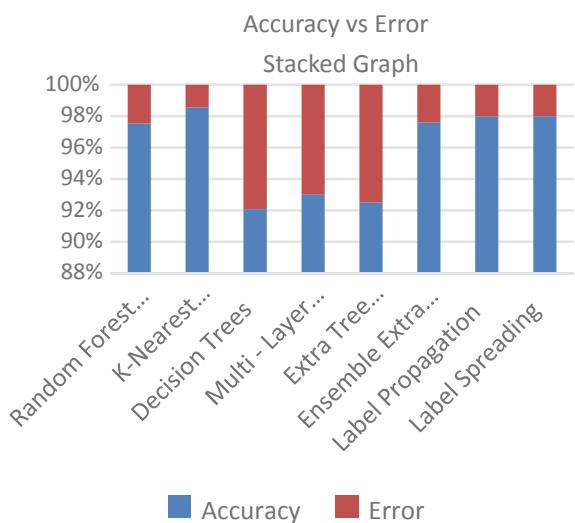
6 Results

Accuracy and error graph makes it easy for better understanding of working model (Table 3 and Graph 2).

Table 3 Accuracy and error of algorithm

Algorithm used	Training accuracy	Testing accuracy	Error
Random Forest Classifier	100	97.5341	2.4659
K-Nearest Neighbors	100	98.5602	1.4398
Decision Trees	100	92.0878	7.9122
Multi-layer Perceptron	94.9021	93.0026	6.9974
Extra Tree Classifier	95.8669	92.5147	7.4853
Ensemble Extra Trees	100	97.5765	2.4235
Label Propagation	100	97.9769	2.0231
Label Spreading	100	97.9769	2.0231

Graph 2 Accuracy versus error stacked graph



6.1 Classification Report

Classification Report is that report which contains the precision, recall, F1 Score and support of the model. It is used for the simple comparison of various model to pick the better for deploying in the final system. It's a numerical data within which values of various metrics are calculated supported True Positive, False Positive, True Negative and False Negative. The Classification Report of different algorithms that were used are shown below.

Random Forest Classifier

	Precision	Recall	F1-score	Support
Using computer	0.98	0.98	0.98	3635
Phone conversation	0.99	0.97	0.98	3585
Vacuum cleaning	0.95	0.99	0.97	2996
Reading book	0.99	0.97	0.98	4707
Watching TV	0.99	0.99	0.99	3223
Ironing	0.98	0.97	0.98	3604
Walking	0.95	0.96	0.95	3669
Exercise	0.96	0.97	0.97	3730
Cooking	0.98	0.97	0.98	3939
Drinking	0.97	0.99	0.98	3040
Brushing hair	0.98	0.99	0.98	1586
Accuracy			0.98	37714
Macro avg	0.98	0.98	0.98	37714
Weighted avg	0.98	0.98	0.98	37714

K-Nearest Neighbors

	Precision	Recall	F1-score	Support
Using computer	0.98	0.98	0.98	3635
Phone conversation	0.98	0.98	0.98	3585
Vacuum cleaning	0.99	0.99	0.99	2996
Reading book	1.00	0.99	1.00	4707
Watching TV	0.99	0.99	0.99	3223
Ironing	0.99	0.99	0.99	3604
Walking	0.99	0.99	0.99	3669
Exercise	0.99	0.99	0.99	3730
Cooking	0.98	0.98	0.98	3939
Drinking	0.97	0.97	0.97	3040

(continued)

(continued)

	Precision	Recall	F1-score	Support
Brushing hair	0.96	0.97	0.97	1586
Accuracy			0.99	37714
Macro avg	0.98	0.98	0.98	37714
Weighted avg	0.99	0.99	0.99	37714

Decision Tree

	Precision	Recall	F1-score	Support
Using computer	0.95	0.94	0.94	3635
Phone conversation	0.94	0.94	0.94	3585
Vacuum Cleaning	0.89	0.87	0.88	2996
Reading book	0.91	0.91	0.91	4707
Watching TV	0.96	0.96	0.96	3223
Ironing	0.94	0.94	0.94	3604
Walking	0.89	0.90	0.90	3669
Exercise	0.91	0.91	0.91	3730
Cooking	0.93	0.93	0.93	3939
Drinking	0.90	0.91	0.91	3040
Brushing hair	0.92	0.92	0.92	1586
Accuracy			0.92	37714
Macro avg	0.92	0.9	0.92	37714
Weighted avg	0.92	0.92	0.92	37714

Multi-layer Perceptron Classifier

	Precision	Recall	F1-score	Support
Using computer	0.96	0.96	0.96	3635
Phone conversation	0.97	0.95	0.96	3585
Vacuum cleaning	0.92	0.96	0.89	2996
Reading book	0.90	0.93	0.92	4707
Watching TV	0.96	0.97	0.97	3223
Ironing	0.96	0.95	0.95	3604
Walking	0.89	0.93	0.91	3669
Exercise	0.93	0.91	0.92	3730
Cooking	0.94	0.91	0.92	3939
Drinking	0.89	0.94	0.92	3040
Brushing hair	0.94	0.92	0.93	1586
Accuracy			0.93	37714

(continued)

(continued)

	Precision	Recall	F1-score	Support
Macro avg	0.93	0.93	0.93	37714
Weighted avg	0.93	0.93	0.93	37714

Extra Tree Classifier

	Precision	Recall	F1-score	Support
Using computer	0.95	0.96	0.95	3635
Phone conversation	0.98	0.92	0.95	3585
Vacuum cleaning	0.83	0.94	0.88	2996
Reading book	0.96	0.86	0.90	4707
Watching TV	0.96	0.96	0.96	3223
Ironing	0.96	0.92	0.94	3604
Walking	0.86	0.92	0.89	3669
Exercise	0.90	0.93	0.92	3730
Cooking	0.93	0.92	0.92	3939
Drinking	0.90	0.95	0.92	3040
Brushing hair	0.95	0.92	0.93	1586
Accuracy			0.93	37714
Macro avg	0.93	0.93	0.93	37714
Weighted avg	0.93	0.93	0.93	37714

Ensemble Extra Trees Classifier

	Precision	Recall	F1-score	Support
Using computer	0.98	0.99	0.98	3635
Phone conversation	1.00	0.97	0.98	3585
Vacuum cleaning	0.95	0.99	0.97	2996
Reading book	1.00	0.96	0.98	4707
Watching TV	0.99	0.99	0.99	3223
Ironing	0.98	0.97	0.97	3604
Walking	0.94	0.96	0.95	3669
Exercise	0.96	0.97	0.97	3730
Cooking	0.98	0.97	0.98	3939
Drinking	0.97	0.99	0.98	3040
Brushing hair	0.98	0.98	0.98	1586
Accuracy			0.98	37714

(continued)

(continued)

	Precision	Recall	F1-score	Support
Macro avg	0.98	0.98	0.98	37714
Weighted avg	0.98	0.98	0.98	37714

Label Propagation

	Precision	Recall	F1-score	Support
Using computer	0.98	0.98	0.98	3635
Phone conversation	0.98	0.98	0.98	3585
Vacuum cleaning	0.99	0.99	0.99	2996
Reading book	0.99	0.99	0.99	4707
Watching TV	0.99	0.99	0.99	3223
Ironing	0.98	0.98	0.98	3604
Walking	0.98	0.98	0.98	3669
Exercise	0.98	0.98	0.98	3730
Cooking	0.97	0.98	0.97	3939
Drinking	0.96	0.95	0.96	3040
Brushing hair	0.95	0.96	0.96	1586
Accuracy			0.98	37714
Macro avg	0.98	0.98	0.98	37714
Weighted avg	0.98	0.98	0.98	37714

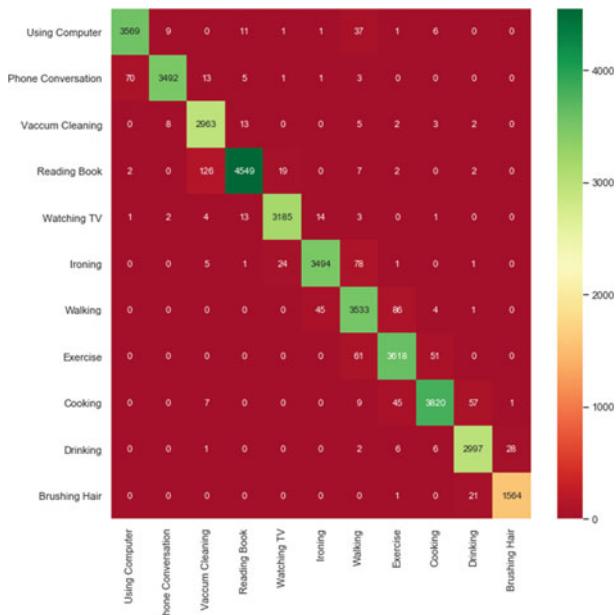
Label Spreading

	Precision	Recall	F1-score	Support
Using computer	0.98	0.98	0.98	3635
Phone conversation	0.98	0.98	0.98	3585
Vacuum cleaning	0.99	0.99	0.99	2996
Reading book	0.99	0.99	0.99	4707
Watching TV	0.99	0.99	0.99	3223
Ironing	0.98	0.98	0.98	3604
Walking	0.98	0.98	0.98	3669
Exercise	0.98	0.98	0.98	3730
Cooking	0.97	0.98	0.97	3939
Drinking	0.96	0.95	0.96	3040
Brushing hair	0.95	0.96	0.96	1586
Accuracy			0.98	37714
Macro avg	0.98	0.98	0.98	37714
Weighted avg	0.98	0.98	0.98	37714

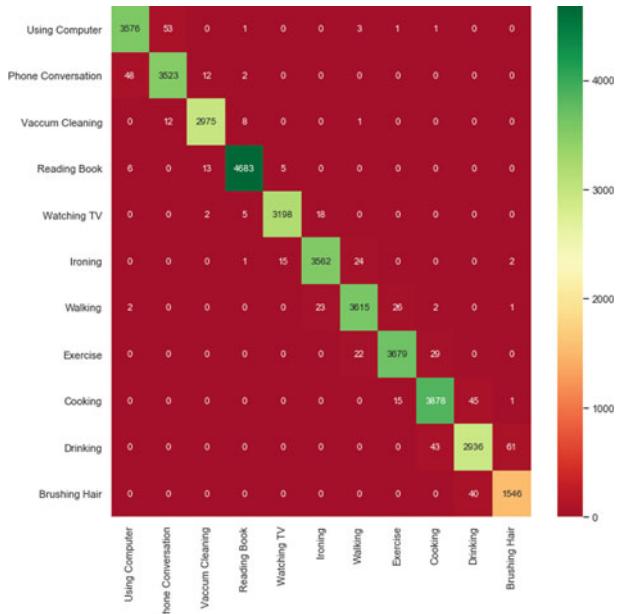
6.2 Confusion Matrix

It is a $M \times M$ matrix used for evaluating the performance of the model within which M is the number of target classes. It clearly shows how many instances were predicted to be true and also it guides where the model went wrong in predicting. Basically, it gives us values of True Positive, False Positive, True Negative and False Negative. It's foremost important for evaluating the performance of any model because considering only accuracy are often misleading when the instances present in each class isn't uniformly distributed and also when there's not a binary class within the datasets. In this False Positive depicts Type I error and False Negative depicts Type II error.

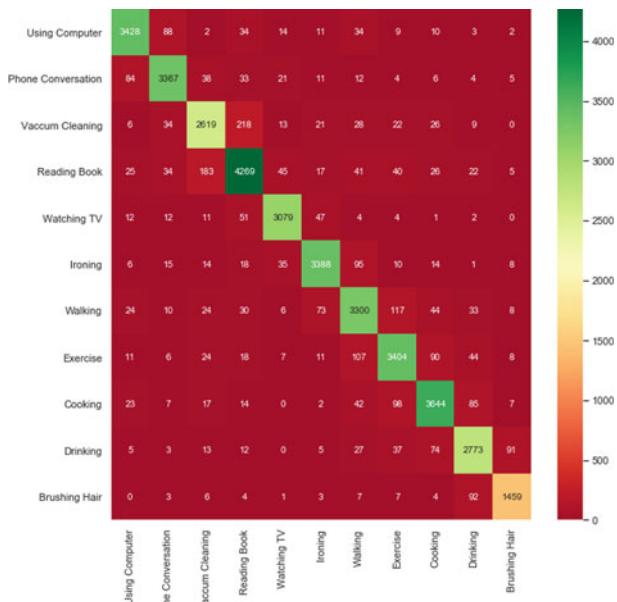
Random Forest Classifier



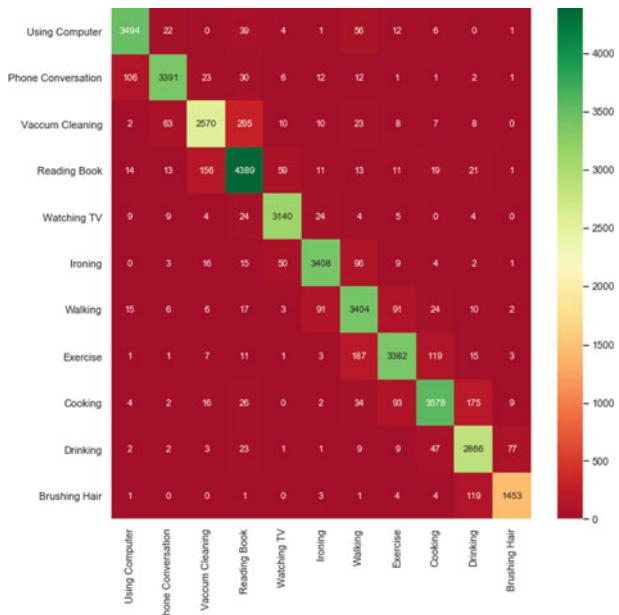
K-Nearest Neighbors



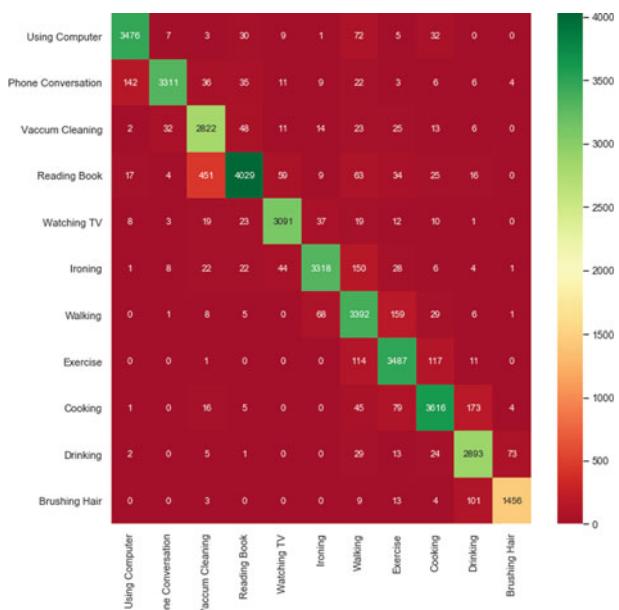
Decision Tree



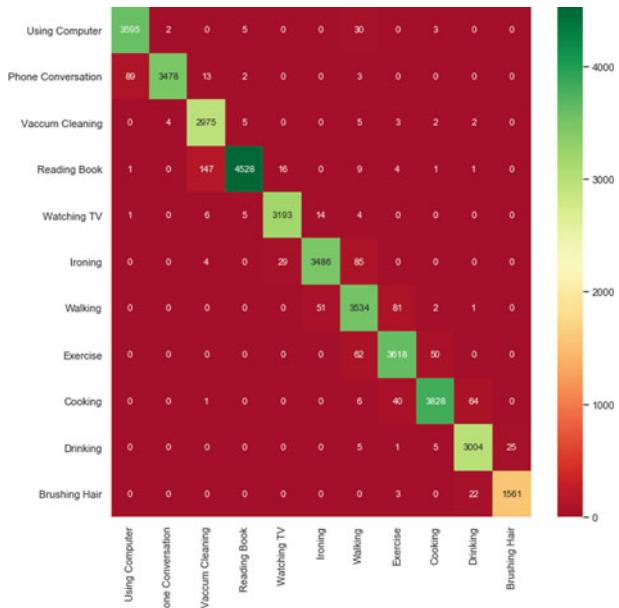
Multi-layer Perceptron Classifier



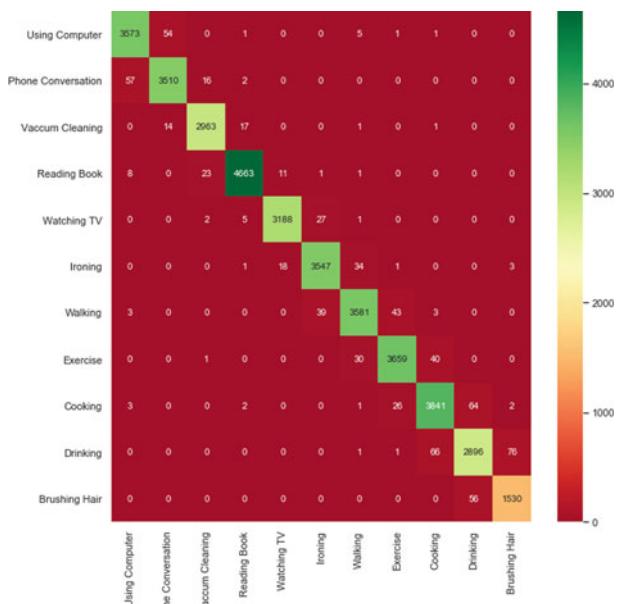
Extra Tree Classifier



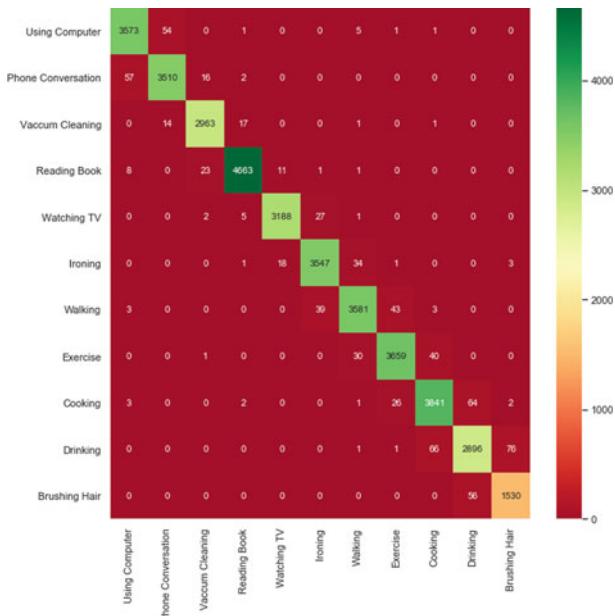
Ensemble Extra Trees Classifier



Label Propagation



Label Spreading



The dataset that was utilized during this research has 11 categories of activities performed by the subjects. For recognition of these activities various machine learning algorithms were tried but among them few had given satisfactory results. Like Random Forest Classifier gave accuracy of 97.53% with error of 2.47%, Decision Trees with accuracy of 92.09% and with error of 7.91%, Multi-layer Perceptron with accuracy of 93% and with error of 7%, Extra Tree Classifier with accuracy of 92.51% and with error of 7.49%, Ensemble Extra Trees giving accuracy of 97.58% and resulting error of 2.42%, Label Propagation and Label Spreading both are giving accuracy of 97.98% thus resulting error of 2.02%.

Among all of these classification algorithms applied KNN had given highest good accuracy of 98.56% with error of just 1.44%. KNN worked well as compared to other because as the dataset have 11 classes and KNN works on principle of Nearest Neighbors so it easily classifies the activities within different classes. As KNN doesn't learn automatically itself which helps us because it stores the information which it had received and classifies the further data just by finding the similarity between previous inputs got. In our model classification report had showed that KNN had classified 'Reading Book' activity very well because there have been many instances of this activity as compared to other ones.

7 Conclusion

The objective of this research is to classify the various human activity performed based on the various values of sensors attached on their body. With the utilization of machine learning the dataset containing the data of values of varied sensors were fed to the classification algorithm like Random Forest Classifier, K-Nearest Neighbors, Decision Tree, Multi-layer Perceptron Classifier, Extra Tree Classifier, Ensemble Extra Trees Classifier, Label Propagation and Label Spreading for recognition of human activity. By applying of these algorithms, a comparative analysis was done supported on accuracy and other performance metrics and it showed that, among of these algorithms used KNN had showed comparatively better results because it showed accuracy of 99% and it classifies the activities better as compared to other algorithms.

This research may be further extended by optimizing the various hyper parameters to boost the accuracy of the model. And also further study can be extended by considering the precise location of the subject to avoid any misclassifications. But for now KNN had given us better result with only one as ‘number of neighbor’ with ‘minkowski’ metric having ‘p’ value equal to one and leaf size of 30. We believe Multi-layer Perceptron Model and Random Forest Classifier can give us more accuracy up to 100%.

Further, this model can be implemented to smartphones, smart watches and smart gadgets and may be used on real time basis, i.e. after getting values from sensors it should immediately predict the activity performed by the person.

References

1. Kwapisz, Jennifer R., Weiss, Gary M., Moore, Samuel A.: Activity recognition using cell phone accelerometers. ACM SIGKDD Explor. Newslett. **12**(2), 74–82 (2010)
2. Anguita, D., Ghio, A., Oneto, L., Parra, X., Reyes-Ortiz, J.L.: Human activity recognition on smartphones using a multiclass hardware friendly support vector machine. Springer International Workshop on Ambient Assisted Living Lecture notes in Computer Science, vol. 7657, pp. 216–223 (2012)
3. Varkey, J., Pompili, D., Walls, T.: Human motion recognition using a wireless sensor-based wearable system. In: Proceedings of Ubiquitous Computing, pp. 897–910 (2012)
4. Jalal, A., Khan, M.A., Hasan, A.S.: Wearable sensor based human behavior understanding and recognition in daily life for smart environments. In: Proceedings of IEEE Conference on FIT (2018)
5. Wu, J., Sun, L., Jafari, R.: A wearable system for recognizing American sign language in real-time using IMU and surface EMG sensors. In: Proceedings of IEEE Journal of Biomedical and Health Informatics, pp. 1281–1290 (2016)
6. Roobini, S., FenilaNaomi, J.: Smartphone sensor based human activity recognition using deep learning models. Int. J. Recent Technol. Eng. **8**(1), ISSN: 2277-3878
7. Randhawa, P., Shanthagiri, V., Kumar, A., Yadav, V.: Human activity detection using machine learning methods from wearable sensors (2020). <https://doi.org/10.1108/sr-02-2020-0027>
8. D’souza, W.T., Kavitha, R.: Human activity recognition using accelerometer and gyroscope sensors (2017). <https://doi.org/10.21817/ijet/2017/v9i2/170902134>

9. Ronao, C.A., Cho, S.-B.: Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Syst. Appl.* **59**, 235–244 (2016)
10. Bayat, A., Pomplun, M., Tran, D.A.: A study on human activity recognition using accelerometer data from smartphones. In: The 11th International Conference on Mobile Systems and Pervasive Computing (MobiSPC-2014)
11. Shoaib, M., Scholten, H., Havinga, P.J.M.: Towards physical activity recognition using smartphone sensors. In: 2013 IEEE 10th International Conference on Ubiquitous Intelligence & Computing and 2013 IEEE 10th International Conference on Autonomic & Trusted Computing, pp. 80–87 (2013)
12. Anguita, D., Ghio, A., Oneto, L., Parra, X., Reyes-Ortiz, J.L.: A public domain dataset for human activity recognition using smartphone's. In: ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) (2013)
13. Mannini, A., Sabatini, A.M.: Machine learning methods for classifying human physical activity from on body accelerometers. *Sensors* **2010**(10), 1154–1175 (2010)
14. Rosati, S., Balestra, G., Knaflitz, M.: Comparison of different sets of features for human activity recognition by wearable sensors. *Sensors* **18**(12), 4189 (2018)
15. Krishnan, N.C., Panchanathan, S.: Analysis of low resolution accelerometer data for continuous human activity recognition. ICASSP (2008)
16. Tapia, E.M., Intille, S.S. et al.: Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor. In: Proceedings of the 2007 11th IEEE International Symposium on Wearable Computers, 1–4 (2007)
17. Kunze, K., Barry, M., Heinz, E.A., Lukowicz, P., Majoe, D., Gutknecht, J.: Towards recognizing Tai Chi—an initial experiment using wearable sensors (2006)
18. Kumar, M., Shenbagaraman, V.M., Ghosh, A.: Predictive data analysis for energy management of a smart factory leading to sustainability. In: Favorskaya, M.N., Mekhilef, S., Pandey, R.K., Singh, N. (eds.) Innovations in Electrical and Electronic Engineering. Springer, pp. 765–773 [ISBN 978-981-15-4691-4] (2020)
19. Ravi, N., Dandekar, N., Mysore, P., Littman, M.L.: Activity recognition from accelerometer data. In: IAAI'05 Proceedings of the 17th Conference on Innovative Applications of Artificial Intelligence, Vol. 3, pp. 1541–1546, Pittsburg, Pennsylvania (2005)
20. Mandal, S., Biswas, S., Balas, V.E., Shaw, R.N., Ghosh, A.: Motion prediction for autonomous vehicles from Lyft dataset using deep learning. In: 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, pp. 768–773 (2020). <https://doi.org/10.1109/iccca49541.2020.9250790>
21. Mandal, S., Balas, V.E., Shaw, R.N., Ghosh, A.: Prediction analysis of idiopathic pulmonary fibrosis progression from OSIC dataset. In: 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, India, pp. 861–865 (2020). <https://doi.org/10.1109/gucon48875.2020.9231239>
22. AIR University Dataset: Intelligent Media—Wearable Smart Home Activities (IM-WSHA) Dataset
23. Zhu, X., Ghahramani, Z.: Learning from labeled and unlabeled data with label propagation. School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, Technical Report CMU-CALD-02-107 (2002)

Achieving Maximum Sum Spectral Efficiency with Channel Estimation



Ashu Taneja, Ankita Rana, and Nitin Saluja

Abstract The major challenges faced by the mobile users are the poor quality of communication and the fast battery drainage. The varying channel characteristics in wireless communication is the main reason for huge signal drops and poor signal quality. The channel estimation plays an important role and is crucial for determining the accuracy of the system. In this paper, the two channel estimation techniques, namely least square (LS) and minimum mean square error (MMSE) are analyzed for multi-antenna communication scenario. The different practical channel scenarios are considered to evaluate the performance of the proposed system in terms of average sum spectral efficiency (SE). It is observed that the maximum average sum SE is obtained with MMSE channel estimation for Rician fading channels. The impact of varying channel statistics with spatial correlation is also evaluated for different pilot reuse factors. Further, the comparison of different receive combiners based on the channel estimates for signal detection is performed in terms of computational complexity.

Keywords 5G · Wireless communication · Channel estimation · Spatial correlation · Pilot reuse

1 Introduction

The fifth generation (5G) wireless technology aims to provide ultra-reliable low latency communication (URLLC) scenarios for mobile users. The techniques of

A. Taneja

Chitkara University School of Engineering and Technology, Chitkara University, Himachal Pradesh, India

e-mail: ashu.taneja@chitkarauniversity.edu.in

A. Rana (✉) · N. Saluja

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

N. Saluja

e-mail: nitin.saluja@chitkarauniversity.edu.in

multiple-input-multiple-output (MIMO) and massive MIMO have widely been adopted in the modern wireless communication systems. These multiple-antenna techniques provide the advantages of high spectral efficiency, energy efficiency and network reliability [1–4]. In URLLC scenarios, the major challenge is to obtain accurate and real-time channel information which is not known a priori. In massive MIMO system, channel is estimated by means of uplink pilot signals which are transmitted along with the uplink data signals to obtain the channel state information (CSI) [5]. The base station estimates the channel responses of its users in a single-cell scenario and also the channel estimates of the interfering users from neighbouring cells in case of multi-cell scenario. In literature, the papers are proposed [6–8] that considers perfect CSI available at the transmitter. Due to errors in estimation, the perfect CSI is almost impossible. [9] has considered the impact of imperfect CSI on the performance of multi-user multi-antenna wireless communication systems. [10] has presented two channel estimation algorithms for massive MIMO system considering both subarray for transceivers and scatters. The complexity and mean square error (MSE) performance of both algorithms are also compared. [11] has proposed direction-of-arrival (DOA) based channel estimation in which the channel information is divided into DOA information and channel gain information. The DOA information is estimated using angle rotation technique and channel gain information is estimated using training signals. Due to pilot reuse, there is a problem of pilot contamination in massive MIMO systems which require accurate channel estimates [12]. The proposed channel estimation scheme reduces the overhead due to inter-cell large scale fading to zero. [13] considers a multi-user MIMO-OFDM system and presents an iterative channel estimation algorithm such that the pilot contamination is suppressed. [14] has obtained channel estimates for OFDM system based on angle-of-arrival (AoA) information for applications in high speed rails. [15] has proposed a low-complexity discrete fourier transform (DFT) based channel estimation for OFDM systems while noise elimination (NE) based DFT scheme is proposed in [16]. [17] has used the concept of steering vector for processing after the pilot-aided estimation such that the overall system performance is improved but with a small computational cost. [18] jointly estimates the channel parameters viz., DOA, fading coefficient and delay thereby reducing the complexity of the computations. The problem of downlink channel estimation is overcome in [19] with a low-complexity algorithm for frequency division duplex (FDD) massive MIMO systems. [20, 21] propose channel estimation techniques for massive MIMO systems with hybrid precoding for random and frequency selective channel scenarios.

Very few papers have considered the impact of spatial channel correlation into account. The massive MIMO system with spatially uncorrelated Rayleigh fading channels are considered in [22] while line-of-sight (LoS) propagation in [23]. But the practical channels are well explained by Rician fading model with a LoS component and a non-line-of-sight (NLoS) component [24]. Massive MIMO system with Rician fading model is analysed in [25–27] for a single-cell scenario. An analysis of multi-cell scenario is performed in [28–30] such that within one cell Rician channel fading is considered and among adjacent neighbouring cells Rayleigh channel is considered. In these papers, correlation is not taken into account. Since the practical channels

are correlated, our paper considers spatial correlation with Rician fading channels between the mobile users and base stations of each cell and Rayleigh fading channels for inter-cell transmission.

The linear processing, mainly, receive combining in the uplink and transmit precoding in the downlink use the channel state information obtained through various channel estimation techniques.

In this paper, an uplink multi-cell multi-antenna communication system is proposed and evaluated for average sum spectral efficiency (SE). The channel statistics are obtained using two channel estimation techniques, namely, minimum mean square error (MMSE) and least square (LS). The performance of different receive combining schemes, viz., regularized zero-forcing (RZF) combining, multi-cell minimum mean squared error (MMMSE) combining and maximal ratio (MR) combining, used for signal detection are also compared for computational complexity. Further, the impact of varying channel statistics with spatial correlation on the system performance is evaluated for different pilot reuse factors.

The novel contributions of the paper are

- A multi-cell multi-user scenario is considered for communication in the uplink with a multi-antenna base station. The proposed system is evaluated for average sum spectral efficiency (SE) and complexity using channel estimation techniques, namely MMSE and LS.
- The channel estimates are used for different receive combiners, viz., RZF, MMMSE and MR whose impact on the system performance is analyzed further.
- The impact of spatial correlation is also highlighted considering both correlated and uncorrelated fading channels between mobile user and BS.

The notations used in the paper are tabulated in Table 1.

2 System Model

Consider a multi-antenna multi-cell communication scenario in which the mobile users communicate with each other through a number of base stations (BSs). A given geographical area consists of C cells with K number of mobile users in each cell. There is one BS with N number of antennas at the centre of each cell. Figure 1 shows an illustration of system model with three cells. The mode of operation is assumed to be time division duplex (TDD) in which each coherence block consists of total of s_t samples. In each coherence block, there are s_p samples used as pilot signals for uplink channel estimation, s_u samples used as data signals for uplink transmission and s_d samples for downlink data transmission. For downlink channel estimation, TDD uses channel reciprocity.

The propagation channel denoted by h_{ck}^b represents the channel between k th mobile node in one cell c to the BS in another cell b . Rician fading model is assumed in this paper where \bar{h}_{ck}^b represents the LOS component and a covariance matrix R_{ck}^b

Table 1 Summary of notations

Notation	Description
C	Number of cells
K	Number of mobile users in each cell
N	Number of antennas
s_t	Total number of samples
s_p	Samples used as pilot signals for uplink channel estimation
s_u	Samples for uplink data transmission
s_d	Samples for downlink data transmission
h_{ck}^b	Propagation channel between a mobile node k in cell c to the BS in cell b
\tilde{h}_{ck}^b	Spatially correlated Rician fading LOS component
R_{ck}^b	Spatial correlation covariance matrix
β_{ck}^b	Large-scale fading
α	Path loss exponent
d_{ck}^b	Distance between the mobile user and the BS antenna
F_{ck}^b	Shadow fading
σ^2	Variance of shadow fading
γ	Channel gain at the reference distance of 1 km
x_{ck}	Uplink signal from user k in cell c
p_{ck}	Uplink signal Power
n_b	Additive receiver noise
ϑ_{bk}	Pilot sequence transmitted by user terminal k
\mathcal{V}_{bk}	Pilot set
y_{bci}^p	Processed received pilot signal
\tilde{h}_{ci}^b	Estimation error
C_{ci}^b	Estimation error covariance matrix
\hat{h}_{ci}^b	Estimated channel
r_{bk}^{MMSE}	MMSE receive combining
r_{bk}^{MR}	MR receive combining
r_{bk}^{RZF}	RZF receive combining

represents the spatial correlation of the NLOS component. The following relation gives the average channel gains from the antenna at BS in cell b to the mobile user k in cell c .

$$\beta_{ck}^b = \frac{1}{N} \text{tr}(R_{ck}^b) \quad (1)$$

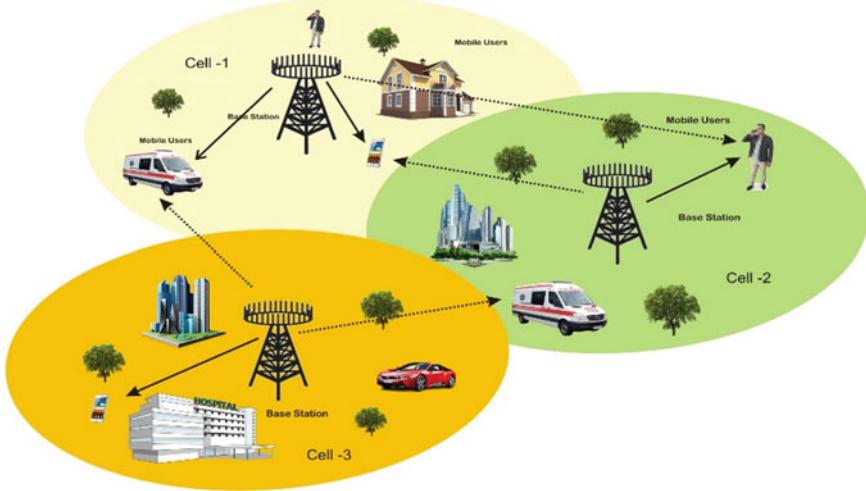


Fig. 1 A multi-cell multi-antenna communication scenario

The parameter β_{ck}^b defines the large-scale fading which includes path loss, shadowing and other propagation effects

$$\beta_{ck}^b = \gamma - 10\alpha \log_{10}\left(\frac{d_{ck}^b}{1\text{km}}\right) + F_{ck}^b \quad (2)$$

where α is the path loss exponent, d_{ck}^b is the distance between the mobile user and the BS antenna and F_{ck}^b defines the shadow fading, $F_{ck}^b \sim N(0, \sigma^2)$, where σ^2 is the variance of shadow fading. γ denotes the channel gain at the reference distance of 1 km.

2.1 Uplink System Model

The received signal at BS in cell b is given by

$$y_b = \sum_{c=1}^C \sum_{k=1}^{K_c} h_{ck}^b x_{ck} + n_b \quad (3)$$

$$= \sum_{k=1}^{K_b} h_{bk}^b x_{bk} + \sum_{c=1}^C \sum_{i=1}^{K_c} h_{ci}^b x_{ci} + n_b \quad (4)$$

$c \neq b$

where x_{ck} is the uplink signal from user k in cell c which has power p_{ck} , such that $p_{ck} = E\{x_{ck}^2\}$ and n_b is the additive receiver noise with zero mean and unit variance.

The BS in cell b needs the channel estimates of its user terminals and also the channel estimates of interfering users from other cells for receive processing and interference suppression respectively. The uplink pilot signals are used for the estimation of channels for which s_p samples are reserved in each coherence block. The pilot sequence $\vartheta_{bk}\mathbb{C}^{s_p}$ is transmitted by user terminal k in cell b . Let us define a pilot set $\mathcal{V}_{bk} = \{(c, i) : \vartheta_{ci} = \vartheta_{bk}, c = 1, 2 \dots C, i = 1, 2 \dots K_c\}$ which gives all users in the system using the same pilot sequence as that used by k th user in cell b . $\sqrt{p_{bk}}$ is the transmit power by which the pilot sequence is scaled before transmission. The received pilot signals at BS b is given by

$$y_b^p = \sum_{k=1}^{K_b} \sqrt{p_{bk}} h_{bk}^b \vartheta_{bk}^T + \sum_{\substack{c=1 \\ c \neq b}}^C \sum_{i=1}^{K_c} \sqrt{p_{ci}} h_{ci}^b \vartheta_{ci}^T + n_b^p \quad (5)$$

he BS in cell b multiples y_b^p with the user's pilot sequence ϑ_{ci}^* to obtain the processed received pilot signal y_{bci}^p which is used for estimating the channel h_{ci}^b .

$$y_{bci}^p = y_b^p \vartheta_{ci}^* = \sqrt{p_{ci}} s_p h_{ci}^b + (c', i') \in \vartheta_{ci} / (l, i) \sqrt{p_{c'i'} s_p} h_{c'i'}^b + n_b^p \vartheta_{ci}^* \quad (6)$$

2.2 MMSE Channel Estimation

The MMSE channel estimate from BS b to user i in cell c is given by

$$\hat{h}_{ci}^b = \bar{h}_{ci}^b + \sqrt{p_{ci}} R_{ci}^b \varphi_{ci}^b (y_{bci}^p - \bar{y}_{bci}^p) \quad (7)$$

where $\bar{y}_{bci}^p = \sum_{(c', i') \in \vartheta_{ci} / (l, i)} \sqrt{P_{c'i'}} s_p \bar{h}_{c'i'}^b$

$$\varphi_{ci}^b = s_p \text{Cov}\{y_{bci}^p\}^{-1} = \left(\sum_{(c', i') \in \vartheta / l, i} \sqrt{p_{c'i'}} s_p + R_{c'i'}^b + \sigma^2 I_N \right)^{-1} \quad (8)$$

$\tilde{h}_{ci}^b = h_{ci}^b - \hat{h}_{ci}^b$, is the estimation error whose covariance matrix is given by

$$\mathbf{C}_{ci}^b = R_{ci}^b - p_{ci} s_p R_{ci}^b \varphi_{ci}^b R_{ci}^b \quad (9)$$

The mean square error (MSE) is given by $E\{\|h_{ci}^b - \hat{h}_{ci}^{b2}\|\} = \text{tr}(\mathbf{C}_{ci}^b)$.

2.3 LS Channel Estimation

The LS channel estimates is given by

$$\hat{h}_{ci}^b = \frac{1}{\sqrt{p_{ci}\tau_p}}y_{bci}^p \quad (10)$$

2.4 Uplink Spectral Efficiency and Receive Combining

It is a linear processing technique in which the BS in cell b uses a receive combining vector $r_{ck} \in \mathbb{C}^N$ in order to separate the desired signal from the interfering signals. During data transmission, BS b correlates the received signal y_b with the combining vector to obtain

$$r_{bk}^H y_b = r_{bk}^H h_{bk}^b s_{bk} + \sum_{\substack{i=1 \\ i \neq k}}^{K_b} r_{bk}^H h_{bi}^b s_{bi} + \sum_{c=1}^C \sum_{i=1}^{K_c} r_{bk}^H h_{ci}^b s_{ci} + r_{bk}^H n_b \quad (11)$$

The ergodic channel capacity of user k in cell b , SE_{bk}^{UL} is given by:

$$SE_{bk}^{UL} = \frac{s_u}{s_t} \log_2 (1 + Y_{bk}^{UL}) \quad (12)$$

$$Y_{bk}^{UL} = \frac{p_{bk} \left| E \left\{ r_{bk}^H \hat{h}_{bk}^b \right\} \right|^2}{\sum_{c=1}^C \sum_{\substack{i=1 \\ (c, i) \neq (b, k)}}^{K_c} p_{ci} E \left\{ \left| r_{bk}^H \hat{h}_{ci}^b \right|^2 \right\} - p_{bk} \left| E \left\{ r_{bk}^H \hat{h}_{bk}^b \right\} \right|^2 + \sigma^2 E \{ r_{bk} \}} \quad (13)$$

r_{bk} depends on the channel estimates and different receive combining vectors are summarized below:

$$r_{bk}^{MMSE} = p_{bk} \left(\sum_{c=1}^C \sum_{i=1}^{K_c} p_{ci} \left((\hat{h}_{ci}^b)^H (\hat{h}_{ci}^b) + C_{ci}^b \right) + \sigma^2 I_N \right)^{-1} \hat{h}_{bk}^b, \text{ MMSE combining} \quad (14)$$

$$r_{bk}^{RZF} = \hat{h}_{bk}^b, \text{ MR combining} \quad (15)$$

$$r_{bk}^{RZF} = \hat{h}_{bk}^b \left(\left(\hat{h}_{bk}^b \right)^H + \sigma^2 p_{bk}^{-1} \right)^{-1}, \text{ RZF combining} \quad (16)$$

3 Results and Discussions

Here, the simulations results are presented which are obtained using MATLAB with 10^4 number of iterations are averaged to evaluate each simulation point. In our multi-cell setup, 9 cells are considered with each cell covering an area of 250×250 m. Each cell is having a multi-antenna BS at the centre with number of antennas in the range of 10–100. In each cell, 10 number of users are uniformly distributed in each cell. The parameters values listed in Table 2 are considered. The system performance is evaluated in terms of average sum spectral efficiency (SE) for different channel estimation algorithms and for different receive combiners. The computational complexity of the system with different receive combiners is also evaluated.

The average sum SE of the system depends on number of data samples in each coherence block, the channel statistics and the receive combining vector. The channel estimates are obtained from MMSE channel estimation technique and LS channel estimation technique. Figure 2 gives the average sum SE of the proposed system for uncorrelated Rayleigh and Rician fading channels for different number of BS antennas. It is observed that maximum SE is obtained for MMSE channel estimate with Rician fading channels. For LS channel estimate, the average SE coincide in case of both Rayleigh and Rician fading channels.

The impact of spatial correlation on the system average SE is depicted in Fig. 3. The channel fading models namely Rayleigh and Rician are considered for evaluation. For spatially correlated Rician channels, maximum SE of 14 bits/s/Hz/cell is achieved for 100 number of antennas at the BS. This value is 11.8 bits/s/Hz/cell for uncorrelated Rician channels. The improvement in average sum SE with spatial correlation is 18%. The same trend is observed for Rayleigh fading channels with average sum SE values of 7.8 bits/s/Hz/cell and 7.02 bits/s/Hz/cell for correlated and uncorrelated channels respectively, thus an improvement of 11% with spatial correlation.

Table 2 Simulation parameters

Parameter	Value	Parameter	Value
C	9	B	20 MHz
K	10	σ	4
N	10–100	f	1,2 or 3
s_t	200	α	3.76
s_p	10	n_b	-94 dBm

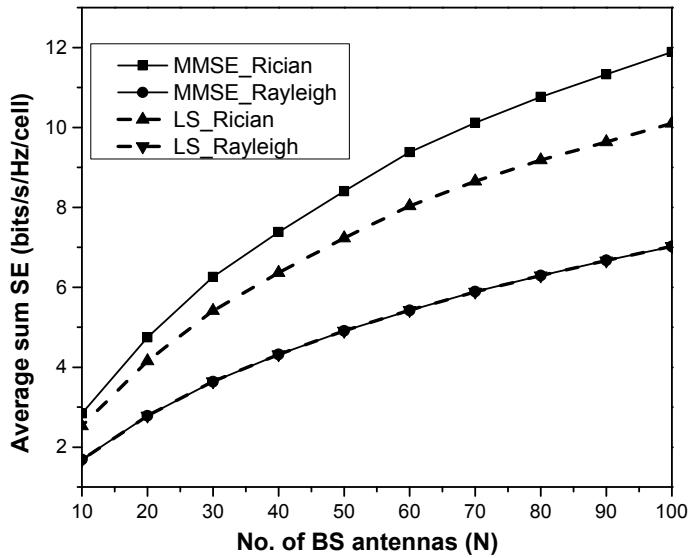


Fig. 2 Variation of sum SE with number of BS antennas for $K = 10$, $C = 9$, $f = 1$ for spatially uncorrelated Rician and Rayleigh fading channels

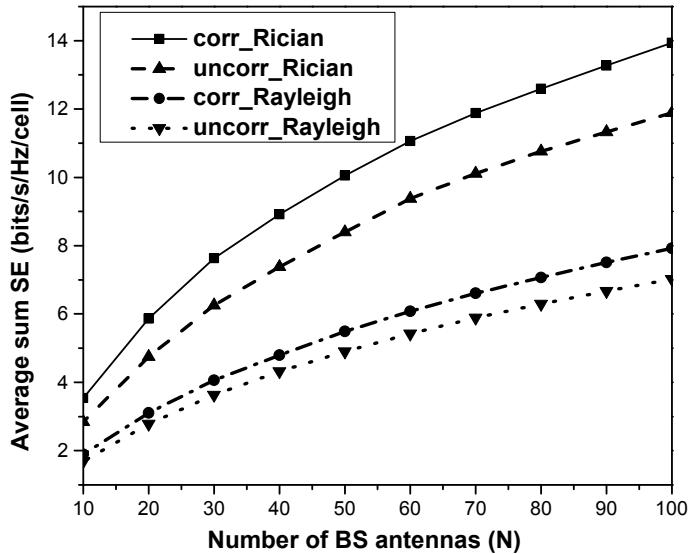


Fig. 3 Effect of spatial correlation on the average sum SE of the system

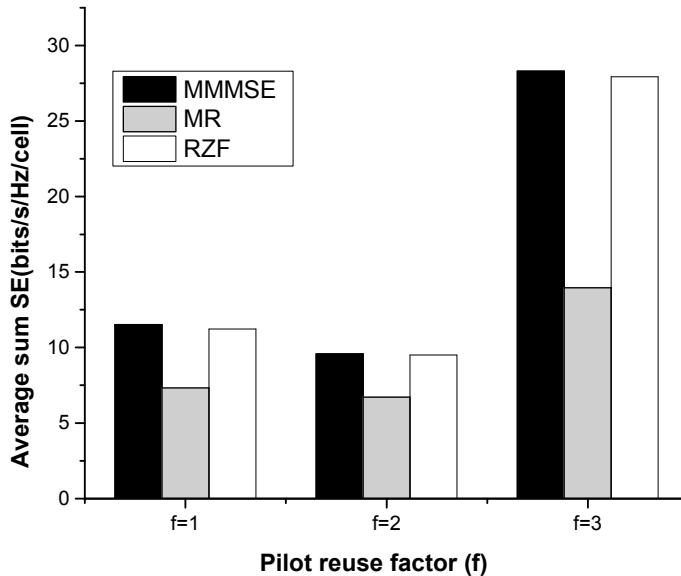


Fig. 4 Average sum SE with different combiners for varying f ($f = 1, 2$ and 3)

The system average sum SE for different combiners and for different pilot reuse factors is shown in Fig. 4. MMSE channel estimates are used with $N = 100$ and $K = 10$. In each coherence block, s_p samples are reserved for pilot sequence which is distributed among the users and are reused across the cells. $s_p = fK$ where f is the pilot reuse factor and K is the number of users. MMSE and RZF receive combiners outperform the MR receive combiner for different pilot reuse factors.

The computational complexity of the three different receive combiners is shown in Fig. 5 for varying number of BS antennas with $K = 10$. MMSE receive combiner has more complexity compared to RZF and MR receive combiners. MR combining provides the lowest computational complexity since it involves no matrix inversion.

4 Conclusion

In this paper, the multi-cell multi antenna communication system is evaluated for two different channel estimation techniques, namely MMSE and LS. The different practical channel scenarios are considered for performance evaluation. It is observed that maximum average sum SE is obtained with MMSE channel estimation for Rician fading channel and the lowest with LS estimation. The impact of spatial correlation on the system performance is also analysed. The paper shows that for 100 number of BS antennas, the average sum SE improves by 11% with spatial correlation for Rayleigh fading channels and by 18% for Rician fading channels. The use of different

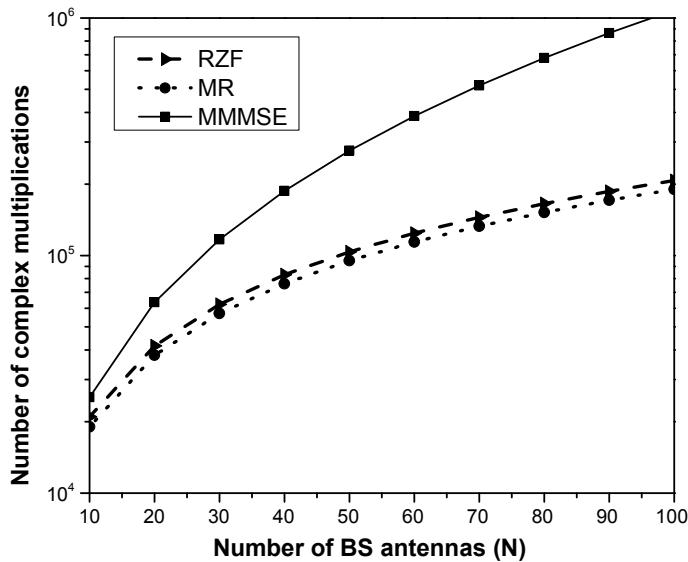


Fig. 5 Complexity of different receive combiners for $K = 10$ and varying N

receive combiners suggest that MMMSE and RZF receive combiners outperform MR receive combiner in terms of average sum SE for different values of pilot reuse factors. However, MR combining provides the lowest computational complexity since it involves no matrix inversion.

References

1. Erik, G.L., Edfors, O., Tufvesson, F.: Massive MIMO for next generation wireless systems. *IEEE Commun. Mag.* **52**(2), 186–195 (2014)
2. Junyoung, N., Caire, G., Debbah, M.: Capacity scaling of massive MIMO in strong spatial correlation regimes. *IEEE Trans. Inf. Theory* **66**(5), 3040–3064 (2019)
3. Abbas, Al.W., Rizzo, Al.H.: Efficient evaluation of massive MIMO channel capacity. *IEEE Syst. J.* **14**(1), 614–620 (2019)
4. Jiayi, Z., Chen, S., Lin, Y.: Cell-free massive MIMO: A new next-generation paradigm. *IEEE Access* **7**, 99878–99888 (2019)
5. Emil, B., Hoydis, J., Sanguinetti, L.: Massive MIMO networks: Spectral, energy, and hardware efficiency. *Found. Trends Signal Process.* **11**(3–4), 154–655 (2017)
6. Chuili, K., Zhong, C., Matthaiou, M., Zhang, Z.: Performance of downlink massive MIMO in ricean fading channels with ZF precoder. In: *IEEE International Conference on Communications (ICC)*, pp. 1776–1782. IEEE, London (2015)
7. Apoorva, C., Patel, A., Jagannatham, A.K.: Distributed detection in massive MIMO wireless sensor networks under perfect and imperfect CSI. *IEEE Trans. Signal Process.* **67**(15), 4055–4068 (2019)
8. Rajat, T., Karaman, S., Modiano, E.: Improving age of information in wireless networks with perfect channel state information. *IEEE/ACM Trans. Networking* **28**(4), 1765–1778 (2020)

9. Trinh, V.C., Mollén, C., Björnson, E.: Large-scale-fading decoding in cellular massive MIMO systems with spatially correlated channels. *IEEE Trans. Commun.* **67**(4), 2746–2762
10. Yu, H., Jin, S., Wen, C.K.: Channel estimation for extremely large-scale massive MIMO systems. *IEEE Wirel. Commun. Lett.* **9**(5), 633–637 (2020)
11. Dian, F., Gao, F., Liu, Y., Deng, Y., Wang, G., Zhong, Z.: Angle domain channel estimation in hybrid millimeter wave massive MIMO systems. *IEEE Trans. Wireless Commun.* **17**(12), 8165–8179 (2018)
12. Amin, K., Minn, H.: On channel estimation for massive MIMO with pilot contamination. *IEEE Commun. Lett.* **19**(9), 1660–1663 (2015)
13. Peng, X., Wang, J., Qi, F.: Analysis and design of channel estimation in multicell multiuser MIMO OFDM systems. *IEEE Trans. Veh. Technol.* **64**(2), 610–620 (2014)
14. Yanrong, Z., Zhao, W., Wang, G., Ai, B., Putra, H.H., Juliyanto, B.: AoA-based channel estimation for massive MIMO OFDM communication system on high speed rails. *China Commun.* **17**(3), 90–100 (2020)
15. Mingming, F., Jin, H.: An improved channel estimation algorithm based on DFT in OFDM systems. In: IEEE International Conference on Computer Information and Big data Applications, pp. 321–325. IEEE (2020)
16. Zhao, Y., Zou, W.: A novel NE-DFT channel estimation scheme for milli-meter wave massive mimo vehicular communication. *IEEE Access* **8**, 74965–74976 (2020)
17. Zhaocheng, W., Zhao, P., Qian, C., Chen, S.: Location-aware channel estimation enhanced TDD based massive MIMO. *IEEE Access* **7**(4), 7828–7840 (2016)
18. Xiao, W., Peng, W., Chen, D., Jiang, T.: Joint channel parameter estimation in multi-cell massive MIMO system. *IEEE Trans. Commun.* **67**(5), 3251–3264 (2018)
19. Cheng, Q., Fu, X., Sidiropoulos, N.D.: Algebraic channel estimation algorithms for FDD massive MIMO systems. *IEEE J. Selected Topics Signal Process.* **13**(5), 961–973 (2019)
20. Evangelos, V., Alexandropoulos, G.C., Thompson, J.: Wideband MIMO channel estimation for hybrid beamforming millimeter wave systems via random spatial sampling. *IEEE J. Select. Topics Signal Process.* **13**(5), 1136–1150 (2019)
21. Zhen, G., Hu, C., Dai, L., Wang, Z.: Channel Estimation for millimeter-wave massive MIMO with hybrid precoding over frequency-selective fading channels. *IEEE Commun. Lett.* **20**(6), 1259–1262 (2016)
22. Marzetta, T. L., Larsson, E. G., Yang, H.: Fundamentals of Massive MIMO. Cambridge University Press (2016)
23. Hong, Y., Marzetta, T.L.: Massive MIMO with max-min power control in line-of-sight propagation environment. *IEEE Trans. Commun.* **65**(11), 4685–4693 (2017)
24. David, T., Viswanath, P.: Fundamentals of Wireless Communications. Cambridge University Press (2005)
25. Qi, Z., Jin, S., Wong, K.-K., Zhu, H., Matthaiou, M.: Power scaling of uplink massive MIMO systems with arbitrary-rank channel means. *IEEE J. Select. Topics Signal Process.* **8**(5), 966–981 (2014)
26. Chuil, K., Zhong, C., Matthaiou, M., Zhang, Z.: Performance of downlink massive MIMO in Ricean fading channels with ZF precoder. In: IEEE International Conference on Communications (ICC), pp. 1776–1782. IEEE (2015)
27. Yeqing, H., Hong, Y., Evans, J.: Angle-of-arrival-dependent interference modeling in Rician massive MIMO. *IEEE Trans. Veh. Technol.* **66**(7), 6171–6183 (2017)
28. Lou, Z., Yang, T., Geraci, G., Yuan, J.: Downlink multiuser massive MIMO in Rician channels under pilot contamination. In: IEEE International Conference on Communications (ICC), pp. 1–6. IEEE (2016)
29. Liang, W., Zhang, Z., Dang, J., Wang, J., Liu, H., Wu, F.: Channel estimation for multicell multiuser massive MIMO uplink over Rician fading channels. *IEEE Trans. Veh. Technol.* **66**(10), 8872–8882 (2017)
30. Luca, S., Kammoun, A., Debbah, M.: Asymptotic analysis of multicell massive MIMO over Rician fading channels. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3539–3543. IEEE (2017)

Correction to: Anomalies in Punjabi Language WordNet: An IndoWordNet Evaluation



Monica Bianchini, Vincenzo Piuri, Sanjoy Das, and Rabindra Nath Shaw

Correction to:

Chapter “Anomalies in Punjabi Language WordNet: An IndoWordNet Evaluation” in: M. Bianchini et al. (eds.), *Advanced Computing and Intelligent Technologies, Lecture Notes in Networks and Systems 218*, https://doi.org/10.1007/978-981-16-2164-2_44

The original version of the book was inadvertently published with incorrect Gurmukhi (Panjabi) words in chapter “Anomalies in Punjabi Language WordNet: An IndoWordNet Evaluation”. The correction chapter and the book have been updated with the change.

The updated version of this chapter can be found at
https://doi.org/10.1007/978-981-16-2164-2_44

Author Index

A

- Agarwal, Ritu, 73
Akella, Sundeep V. V. S., 453
Arathi, V. P., 121
Aulakh, Karanpartap Singh, 37

B

- Bajaj, Varnica, 607
Banerjee, Avishek, 269
Barhanpurkar, Kanishk, 319
Barman, Hillol, 83
Basak, Debangho, 577
Baskaran, Santhi, 389
Bedi, Pradeep, 279, 293, 307, 319
Bhardwaj, Diwakar, 99
Bhati, Nishi, 245
Bhirud, S. G., 537, 569
Bhoyar, Aniket, 523

C

- Carvajal, Juan P., 401
Chakraborty, Aritra, 257
Chatterjee, Santanu, 257
Chowdhury, Titash, 207
Chugh, Priya, 577

D

- Dalvi, Ashwini, 111, 537
Das, Indrani, 329
Das, Nataraj, 25
Das, Upamanyu, 577
Das, Victor, 269

Dawdi, Takwa, 429

Deb, Suman, 25
Deepak, Gerard, 15
De, Sudip Kumar, 269

E

Egumadiri, Vijay, 131
Emimal, M., 1

F

Fajardo, Arturo, 401
Fathima Shemim, K. S., 59
Fernando, W. M. J. H., 145

G

Gamage, M. P. A. W., 145
Gandhi, Kanika, 37
Garg, Anupam, 183
Garg, Chakshu, 593
Garg, Sarita Bansal, 417
Gayathri, N., 461
Ghosh, Ankush, 207, 219, 257, 269, 279, 293, 307, 319, 593, 607
Giri, Debasis, 269
Goyal, S. B., 279, 293, 307, 319

H

Haque, Merajul, 523
Harikrishna, T. H., 121

I

- Idhis, Sally M., 429
Inbamalar, T. M., 1

J

- Jain, Apoorva, 537
Jain, Nipun, 195
Jayakumar, Amrutha, 355
Jino Hans, W., 1

K

- Karthika, C., 481
Katiyar, Sparsh, 73
Kaur, Amrita, 183
Kaur, Preeti, 195
Kazi, Faruk, 537
Khareta, Ritika, 245
Kharoud, Gurpreet Singh, 37
Kishor, Netalkar Rohan, 83
Kokane, Rohan, 523
Krishnan, Rahul, 159
Kumar, Nimmala Chaitanya Sai, 131
Kumar, Rakesh, 99

L

- Lohi, Loshima, 339

M

- Mahiban Lindsay, N., 1
Majumder, Koushik, 257, 269
Mal, Arpan Kumar, 607
Mallikharjuna Rao, K., 505
Megalingam, Rajesh Kannan, 131, 493
Modanval, Rajiv K., 461
Moradiya, Smit, 537
Motiani, Mohit, 195

N

- Nailwal, Sharad, 73
Nair, Prashant R., 159, 453
Namdeo, Aman, 593
Nandhakishore, C S, 15
Narayanan, A. G. Hari, 121, 481
Nasir, Qassim, 429
Neogi, Debosmit, 25
Nirmala Devi, M., 165
Nivelkar, Mukta, 569

O

- Omran, Yara, 429

P

- Paez, Carlos, 401
Pai, Maya L., 339, 355
Palimkar, Prajyot, 219, 607
Panigrahi, Sweta, 83
Parashar, Anshu, 183
Patel, Abhishek, 111
Patel, Saurabh, 73
Pathak, Debanjan, 83
Patil, Sachin, 523
Prithvi, Darla Vineeth, 131
Puram, Hari Sudarshan Rahul, 493

R

- Rajawat, Anand Singh, 279, 293, 307, 319
Rajendra Prasath, S. S., 453
Raju, U. S. N., 83
Rama Satish, Aravapalli, 505
Rana, Ankita, 633
Reddy, Chennareddy Pavanth Kumar, 493
Rohit Surya, A. T., 453

S

- Saluja, Nitin, 633
Samsani, Venkata Chanakya, 245
Samudrala, Naveen, 493
Sankar, Vaishnavi, 165
Santhanavijayan, A., 15
Shah, Priyanka, 515
Sharma, Sandeep K., 461
Sharma, Seemu, 37
Sharma, Shalu, 245
Shaw, Rabindra Nath, 207, 219, 257, 269, 279, 293, 307, 319, 593, 607
Siddavatam, Irfan, 111, 537
Singhal, Abhishek, 593
Singh, Arshdeep, 577
Singh, Priyanka, 593
Sinha, Trisha, 207
Sonekar, Shrikant V., 523
Sreekumar, K., 373
Subrahmanyam, V. V., 417
Suklabaidya, Sudip, 329
Suresh, Gayathri, 121

T

- Talib, Manar Abu, 429

Tammana, Hemanth Sai Surya Kumar, [493](#)
Taneja, Ashu, [633](#)
Tantravahi, Santosh, [493](#)
Tayal, Prakhar, [461](#)
Tavar, Niraj, [515](#)
Thakkar, Viraj, [111](#)
Thind, Jobanpreet Singh, [37](#)
Thokala, Nagasai, [493](#)
Titre, Manoj, [523](#)

V

Vashisth, Tushar, [245](#)

Vedpathak, Aditya, [111](#)
Vijayalakshmi, P. P., [481](#)
Vineetha Sankar, P., [373](#)

W

Witkowski Dr., Ulf, [59](#)

Y
Yadav, Jyoti, [577](#)
Yasaswini, V., [389](#)
Yaswanthram, P., [453](#)