

Lecture Notes on Data Engineering
and Communications Technologies 106

Pankaj Verma
Chhagan Charan
Xavier Fernando
Subramaniam Ganesan *Editors*



Advances in Data Computing, Communication and Security

Proceedings of I3CS2021

Lecture Notes on Data Engineering and Communications Technologies

Volume 106

Series Editor

Fatos Xhafa, Technical University of Catalonia, Barcelona, Spain

The aim of the book series is to present cutting edge engineering approaches to data technologies and communications. It will publish latest advances on the engineering task of building and deploying distributed, scalable and reliable data infrastructures and communication systems.

The series will have a prominent applied focus on data technologies and communications with aim to promote the bridging from fundamental research on data science and networking to data engineering and communications that lead to industry products, business knowledge and standardisation.

Indexed by SCOPUS, INSPEC, EI Compendex.

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

More information about this series at <https://link.springer.com/bookseries/15362>

Pankaj Verma · Chhagan Charan ·
Xavier Fernando · Subramaniam Ganesan
Editors

Advances in Data Computing, Communication and Security

Proceedings of I3CS2021

 Springer

Editors

Pankaj Verma
Department of Electronics
and Communication Engineering
National Institute of Technology
Kurukshetra
Kurukshetra, India

Chhagan Charan
Department of Electronics
and Communication Engineering
National Institute of Technology
Kurukshetra
Kurukshetra, India

Xavier Fernando
Department of Electrical and Computer
Engineering
Ryerson University
Toronto, ON, Canada

Subramaniam Ganesan
Department of Electrical and Computer
Engineering
Oakland University
Rochester, MI, USA

ISSN 2367-4512

ISSN 2367-4520 (electronic)

Lecture Notes on Data Engineering and Communications Technologies

ISBN 978-981-16-8402-9

ISBN 978-981-16-8403-6 (eBook)

<https://doi.org/10.1007/978-981-16-8403-6>

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

List of Advisory and Board Members

Chief Patron

Padma Shri Dr. Satish Kumar, Director, National Institute of Technology, Kurukshetra, India

Patrons

Dr. N. P. Singh, Head of Electronics and Communication Engineering Department, National Institute of Technology, Kurukshetra, India

Dr. Mayank Dave, Head of Computer Engineering Department, National Institute of Technology, Kurukshetra, India

Programme Chairs

Dr. Chhagan Charan, National Institute of Technology, Kurukshetra, India

Dr. Sébastien Roy, University of Sherbrooke, Quebec, Canada

Publicity Chairs

Dr. Rajendra Prasad, Technion-Institute of Technology, Israel

Dr. Shweta Meena, National Institute of Technology, Kurukshetra, India

Publication Chairs

Dr. Santosh Kumar, National Institute of Technology, Kurukshetra, India
Dr. Vijayakumar Vandarajan, University of New South Wales, Sydney, Australia

Conference Secretaries

Dr. Pankaj Verma, National Institute of Technology, Kurukshetra, India
Dr. Kriti Bhushan, National Institute of Technology, Kurukshetra, India

Advisory Board Members

Dr. Sameer Sonkusale, Medford, Massachusetts, USA
Dr. R. Venkatesan, Memorial University of Newfoundland, Canada
Dr. Theodoros Tsiftsis, Jinan University, China
Dr. Vijay Arora, Wilkes University, USA
Dr. George Giakos, Director, Quantum Cognitive Systems and Bioinspired Engineering Laboratory, New York, USA
Dr. Miroslav Skoric, Secretary at Amateur Radio Union of Vojvodina (SRV), Serbia
Dr. Ghanshyam Singh, University of Johannesburg, South Africa
Dr. Masood Ur Rehman, University of Glasgow, London, UK
Dr. Ahmed Abdelgawad, Central Michigan University, USA
Dr. Srijib Mukherjee, Oak Ridge National Laboratory, North Carolina, USA
Dr. Mangilal Agarwal, IUPUI, USA
Dr. Paul Ranky, New Jersey Institute of Technology, USA
Dr. Yusuf öztürk, Antalya Science University, Turkey
Dr. Srinivas Talabattula, IISc, Bangalore, India
Dr. Rohit Srivastava, IIT Bombay, India
Dr. Anand Mohan, IIT, BHU, India
Dr. Y. N. Singh, IIT Kanpur, India
Dr. Brahmjit Singh, NIT, Kurukshetra, India
Dr. Rajoo Pandey, NIT, Kurukshetra, India
Dr. Rajiv Kapoor, DTU, Delhi, India
Dr. Shalli Rani, Chitkara University, Chandigarh, India

Technical Programme Committee Members

Dr. Ali Kashif Bashir, The Manchester Metropolitan University, UK
Dr. Wai-keung Fung, Cardiff Metropolitan University, UK
Dr. Rahim Tafazolli, University of Surrey, UK
Dr. Vishal Sharma, Queen's University Belfast, UK
Dr. Amin Malek Mohammadi, Design Engineer at California State University, USA
Dr. Imran Shafique Ansari, University of Glasgow, UK
Dr. Saraju Mohanty, University of North Texas, USA
Dr. Muhammad Bilal, Hankuk University of Foreign Studies, Korea
Dr. Kishore Naik Mude, R&D, Solace Power Inc., Canada
Dr. Saptarshi Das, Pennsylvania State University, USA
Dr. Vijayalakshmi Saravanan, University of Texas, San Antonio, USA
Dr. Ajeet Kaushik, Florida Polytechnic University, USA
Dr. Khaled Rabie, The Manchester Metropolitan University, UK
Dr. Nima Karimian, San Jose State University, USA
Dr. Sunil Aryal, Deakin University, Australia
Dr. Michele Casula, Sorbonne University, France
Dr. Deep Jariwala, University of Pennsylvania, USA
Dr. Nawab Faseeh Qureshi, Sungkyunkwan University, South Korea
Dr. Hamed Taherdoost, OBS Tech Limited, British Columbia, Canada
Dr. Khondker Shajadul Hasan, University of Houston, USA
Dr. Amit Kumar, IIT, Jodhpur, India
Dr. Mainuddin, Jamia Millia Islamia, New Delhi, India
Dr. Devendra Gurjar, NIT, Silchar, India
Dr. Umesh Ghanekar, NIT, Kurukshetra, India
Dr. Arvind Kumar, NIT, Kurukshetra, India
Dr. Naveen Chauhan, NIT Hamirpur, India
Dr. Tarun Rawat, NSIT, Delhi, India
Dr. Kalpana Chauhan, Central University of Haryana, India
Dr. Anshul Agarwal, NIT Delhi, India
Dr. Ashish Chittora, BITS Goa, India
Dr. Aneesh Kumar Sharma, Scientist, RCI, DRDO, India
Dr. Harpal Singh Panwar, Scientist, IRDE, DRDO, India
Dr. Hemant Singh Ajal, CSIR, Chandigarh, India
Dr. Santosh Kumar Meena, NCL, CSIR, Pune
Mr. Rajesh Singh, Scientist, VSSC, ISRO, India
Mr. Anuj Solanki, Samsung Electronics, India
Mr. Hariom Meena, C-DoT, India

Preface

The book titled *Advances in Data Computing, Communication and Security* is a collection of high-quality peer-reviewed contributions from the academicians, researchers, practitioners and industry professionals which were accepted in the International Conference on Advances in Data Computing, Communication and Security (I3CS2021) organized by the Department of Electronics and Communication Engineering in collaboration with the Department of Computer Engineering, National Institute of Technology, Kurukshetra, India, during 08–10 September 2021.

The fast pace of advancing technologies and growing expectations of the next generation requires that the researchers must continuously reinvent themselves through new investigations and development of the new products. Therefore, the theme of this conference was devised as “Embracing Innovations” for the next generation data computing and secure communication systems.

Data engineering is an act of collecting and managing data for analysis. Its domain is very wide, enveloping everything from analysing data to building predictive models. In today’s time, it is almost impossible to think of any industry, which has not been revolutionized by the data engineering. The next generation data engineering is moving towards complete transmutation as the recent developments in data computing are greatly impacted by machine learning (ML), artificial intelligence (AI) and Internet of Things (IoT). To keep pace with this transformation, there is a need to develop and design new protocols and methods for data technologies. Therefore, one part of this book focuses on the recent developments, open challenges and future scope in the area of data science that concerns industrial merchandise, businesses and standardization.

The NextG (next generation) communication technologies envisage offering ubiquitous wireless communication systems, seamless high-quality wireless services, ultimate throughput, self-aggregating network fabrics and highly efficient radio frequency spectrum utilization with greater energy effectiveness. This includes dynamic and adaptive technologies that provide a new standard for radio spectrum accessibility, dynamic and fast adaptive multilayer schemes, smart radio and adaptive networking. The future generation wireless communication technologies, i.e. IoT, software-defined radio, cognitive radio networks, device-to-device communication

and machine-to-machine communication, visible light network and terahertz wireless network, artificial intelligence-based wireless networking tactile, are expected to operate in very complex, uncertain, dynamic and diverse environment. Such an operating environment along with the hardware intricacies and pragmatic challenges limits the performance of the end-to-end communication. Also, the computing functionality of these systems is highly distributed and decentralized in nature. Designing and developing these systems ensuring consistent performance in complex and uncertain environment is indeed a very challenging task. Recently, the accessibility of data irrespective of the time and place is an essential characteristic for NextG communication technology. However, security and privacy are of prime concern when applied to the end-user applications. A recent report of the National Sanitation Foundation (NSF) has revealed that 78% of large organization and 63% of small businesses are cyber attacked annually, and it is expected that these figures will continue to rise in future making the cyber security distress across the world. Therefore, data security and privacy are of primary concern for future technological advances.

This book aims at providing a comprehensive and insightful understanding of the critical review, recent developments, open technical challenges and future directions in the areas of data computing, communication technologies, security and VLSI design and materials.

Salient features of the book:

- The censorious review of the literature in the area of data engineering and secure communication technologies
- Awareness of the advantages and disadvantages of the latest advances in data and communication technologies for future problem formulation for the keen researchers
- Analysis of the inspiring developments in data computing, communication technologies and security to motivate students to generate attentiveness in the next wave of the technologies
- State-of-the-art research methodologies and their utilization in data science, communication, security and VLSI design and materials
- Aims to bridge the gap from fundamental research to application-oriented research on data science and communication

Aims to enlighten the readers including academia, researchers, practitioners and industry professionals with the most advanced engineering task of building and deploying distributed, scalable and reliable data infrastructures and communication systems.

Kurukshetra, India
Kurukshetra, India
Toronto, Canada
Rochester, USA

Pankaj Verma
Chhagan Charan
Xavier Fernando
Subramaniam Ganesan

Acknowledgements

At the outset, we express our sincere gratitude to Prof. Fatos Xhafa, Series Editor of the Book Series *Lecture Notes on Data Engineering and Communications Technologies*, Springer, for accepting our proposal for the publication of the proceedings of the International Conference on Advances in Data Computing, Communication and Security (I3CS2021). We also acknowledge the help and cooperation extended by the Team Springer Nature, especially Sh. Aninda Bose, Senior Publishing Editor, for his valuable inputs, guidance and administrative support throughout this project. We are also grateful to Ms. Sharmila Mary Panner Selvam for her prompt and diligent communication guiding us for preparing the manuscripts in conformity with the standard norms.

We owe special thanks to Prof. M. P. Poonia, Vice Chairman, AICTE, for sparing his valuable time to join us on the inaugural function of the conference. We thank all eminent speakers from academia, industry and R&D organizations for delivering the keynote talks. We are immensely grateful to Dr. Satish Kumar, Director, NIT, Kurukshetra, for his constant support and motivation for organizing the I3CS2021. We are also thankful to Dr. N. P. Singh, Head, ECE Department, NIT, Kurukshetra, and Dr. Mayank Dave, Head of CS Department, NIT, Kurukshetra, for their unconditional support and cooperation extended towards successful organization of the conference. Heartfelt acknowledgement is due to the members of the advisory/technical programme committee, reviewers, session chairs for their valued contribution and of course the authors for sharing their research work on the platform of I3CS2021.

We thank one and all who have contributed directly or indirectly in the smooth conduct and grand success of the conference organized through online mode. Finally, we the Team I3CS2021 express our gratitude to the Almighty for blessing us all with good health during the challenging times of the COVID-19 pandemic situation.

Contents

Data Computing

OntoIntAIC: An Approach for Ontology Integration Using Artificially Intelligent Cloud	3
V. Adithya, Gerard Deepak, and A. Santhanavijayan	
Identifying the Most Frequently Used Words in Spam Mail Using Random Forest Classifier and Mutual Information Content	15
Mohammad A. N. Al-Azawi	
Comparative Analysis of Intelligent Learning Techniques for Diagnosis of Liver Tumor from CT Images	27
Rutuja Nemane, Anuradha Thakare, Shreya Pillai, Nupur Shiturkar, and Anjitha Nair	
A New Model for COVID-19 Detection Using Chest X-ray Images with Transfer Learning	39
Vaibhav Jaiswal and Arun Solanki	
Violence Detection in Videos Using Transfer Learning and LSTM	51
Rohan Choudhary and Arun Solanki	
Kernel Functions for Clustering of Incomplete Data: A Comparative Study	63
Sonia Goel and Meena Tushir	
Neuroimaging (Anatomical MRI)-Based Classification of Alzheimer’s Diseases and Mild Cognitive Impairment Using Convolution Neural Network	77
Yusera Farooq Khan and Baijnath Kaushik	
Design and Implementation of Stop Words Removal Method for Punjabi Language Using Finite Automata	89
Tanveer Singh Kochhar and Gulshan Goyal	

Meticulous Presaging Arrhythmia Fibrillation for Heart Disease Classification Using Oversampling Method for Multiple Classifiers Based on Machine Learning	99
Ritu Aggarwal and Prateek Thakral	
Unsupervised Modeling of Workloads as an Enabler for Supervised Ensemble-based Prediction of Resource Demands on a Cloud	109
Karthick Seshadri, C. Pavana, Korrapati Sindhu, and Chidambaran Kollengode	
A Technique to Find Out Low Frequency Rare Words in Medical Cancer Text Document Classification	121
Falguni N. Patel, Hitesh B. Shah, and Shishir Shah	
Image-Based Spammer Analysis Using Suitable Features Selection by Genetic Algorithm in OSNs	133
Somya Ranjan Sahoo, Asish Kumar Dalai, Sanjit Ningthoujam, and Saroj Kumar Panigrahy	
An Improved Firefly Algorithm Based Cluster Analysis Technique	145
Manju Sharma and Sanjay Tyagi	
Stock Market Analysis of Beauty Industry During COVID-19	157
Satya Verma, Satya Prakash Sahu, and Tirath Prasad Sahu	
An OWA-Based Feature Extraction and Ranking for Performance Evaluation of the Players in Cricket	169
Khalid Anwar, Aasim Zafar, Arshad Iqbal, and Shahab Saquib Sohail	
Cardiac Problem Risk Detection Using Fuzzy Logic	181
T. Sai Vyshnavi, Shruti Prakash, Vyomikaa Basani, and K. Uma Rao	
Deep Learning Approach for Motor-Imagery Brain States Discrimination Problem	193
Saptarshi Mazumdar and Rajdeep Chatterjee	
Automated Diagnosis of Breast Cancer: An Ensemble Approach	207
Surbhi Gupta	
Survey on Formation Verification for Ensembling Collective Adaptive System	219
Muhammad Hamizan Johari, Siti Nuraishah Agos Jawaddi, and Azlan Ismail	
An Inquisitive Prospect on the Shift Toward Online Media, Before, During, and After the COVID-19 Pandemic: A Technological Analysis	229
Anshul Gupta, Sunil Kr. Singh, Muskaan Chopra, and Shabeg Singh Gill	

A Comparative Study of Learning Methods for Diabetic Retinopathy Classification 239
 Qazi Mohammad Areeb and Mohammad Nadeem

CNN for Detection of COVID-19 Using Chest X-Ray Images 251
 Ashish Karhade, Abhishek Yogi, Amit Gupta, Pallavi Landge, and Manisha Galphade

Crop Yield Prediction Using Weather Data and NDVI Time Series Data 261
 Manisha Galphade, Nilkamal More, Abhishek Wagh, and V. B. Nikam

Automated Multiple-Choice Question Creation Using Synonymization and Factual Confirmation 273
 M. Pranav, Gerard Deepak, and A. Santhanavijayan

EASDisco: Toward a Novel Framework for Web Service Discovery Using Ontology Matching and Genetic Algorithm 283
 N. Krishnan and Gerard Deepak

An Approach Towards Human Centric Automatic Ontology Design 293
 S. Manaswini, Gerard Deepak, and A. Santhanavijayan

A Review on Dataset Acquisition Techniques in Gesture Recognition from Indian Sign Language 305
 Animesh Singh, Sunil Kr. Singh, and Ajay Mittal

OntoReqC: An Ontology Focused Integrative Approach for Classification of Software Requirements 315
 R. Dheenadhayalan and Gerard Deepak

SemUserProfiling: A Hybrid Knowledge Centric Approach for Semantically Driven User Profiling 325
 Rituraj Ojha and Gerard Deepak

Prediction of Stroke Disease Using Different Types of Gradient Boosting Classifiers 337
 Astik Kumar Pradhan, Satyajit Swain, Jitendra Kumar Rout, and Niranjana Kumar Ray

Bi-objective Task Scheduling in Cloud Data Center Using Whale Optimization Algorithm 347
 Srichandan Sobhanayak, Isaac Kennedy Alexandre Mendes, and Kavita Jaiswal

NDVI-Based Raster Band Composition for Classification of Vegetation Health 361
 Rishwari Ranjan, Ankit Sahai Saxena, and Hemlata Goyal

FAMDM: An Approach to Handle Semantic Master Data Using Fog-based Architecture 371
 Saravjeet Singh, Jaiteg Singh, and Jatin Arora

A Smart Mobile Application for Stock Market Analysis, Prediction, and Alerting Users 379
 Rutvi Boda, Saroj Kumar Panigrahy, and Somya Ranjan Sahoo

Impact of Blockchain Technology on the Development of E-Businesses 391
 Jatin Sharma and Hamed Taherdoost

Automatic Audio and Video Summarization Using TextRank Algorithm and Convolutional Neural Networks 397
 Kriti Saini and Mayank Dave

Dealing with Class Imbalance in Sentiment Analysis Using Deep Learning and SMOTE 407
 Shweta Kedas, Arun Kumar, and Puneet Kumar Jain

Classification of Machine Learning Algorithms 417
 Hamed Taherdoost

Performance Analysis of Object Classification System for Traffic Objects Using Various SVM Kernels 423
 Madhura M. Bhosale, Tanuja Satish Dhope (Shendkar), Akshay P. Velapure, and Dina Simunic

Analysis and Prediction of Liver Disease for the Patients in India Using Various Machine Learning Algorithms 433
 U. Sinthuja, Vaishali Hatti, and S. Thavamani

Vision-Based Human-Following Robot 443
 Ajay Thakran, Akshay Agarwal, Pulkit Mahajan, and Santosh Kumar

MRI Cardiac Images Segmentation and Anomaly Detection Using U-Net Convolutional Neural Networks 451
 Kriti Srikanth, Sapna Sadhwani, and Siddhaling Urolagin

Communication

GSM-Based Smart Marine Tracking System for Location Monitoring and Emergency Protection 467
 T. Kesavan and K. Lakshmi

Defected Ground UWB Antenna for Microwave Imaging-Based Breast Cancer Detection 477
 Anupma Gupta, Paras Chawla, Bhawna Goyal, and Aayush Dogra

Impact of Beam Formation on 5G Mobile Communication Network in mmWave Frequency Band 487
 Nallapalem Neeraj Srinivas, Yaraswini Vellisetty, and P. C. Jain

Multi-transform 2D DCT Architecture Supporting AVC and HEVC Video Codec 497
 K. Phani Raghavendra Sai and I. Mamatha

Performance Analysis of SDN-Inspired Swarm Intelligence-Based Routing Optimization Algorithm in Vehicular Network 509
 K. Abinaya, P. Praveen Kumar, T. Ananth kumar, R. Rajmohan, and M. Pavithra

Blending of Window Functions in Sonar Array Beamforming 521
 S. Vijayan Pillai, T. Santhanakrishnan, and R. Rajesh

Photonic Crystal Fiber Based Refractive Index Sensor for Cholesterol Sensing in Far Infrared Region 533
 Amit Kumar, Pankaj Verma, and Poonam Jindal

An Efficient, Low-Cost, and Reliable Monostatic Microwave Imaging (MWI) Approach for the Detection of Breast Tumor Using an Ultra-Wideband Dielectric Resonator Antenna 543
 Gagandeep Kaur and Amanpreet Kaur

Effect of Link Reliability and Interference on Two-Terminal Reliability of Mobile Ad Hoc Network 555
 Ch. Venkateswara Rao and N. Padmavathy

Security

User Information Privacy Awareness Using Machine Learning-Based Tool 569
 Aaditya Deshpande, Ashish Chavan, and Prashant Dhotre

Target Node Protection from Rumours in Online Social Networks 581
 Saranga Bora and Shilpa Rao

Risk Detection of Android Applications Using Static Permissions 591
 Meghna Dhalaria and Ekta Gandotra

Cryptography-Based Efficient Secured Routing Algorithm for Vehicular Ad Hoc Networks 601
 Deepak Dembla, Parul Tyagi, Yogesh Chaba, Mridul Chaba, and Sarvjeet Kaur Chatrath

An Efficient Feature Fusion Technique for Text-Independent Speaker Identification and Verification 613
 Savina Bansal, R. K. Bansal, and Yashender Sharma

Identifying Forged Digital Image Using Adaptive Over Segmentation and Feature Point 623
Nemani Nithyusha, Rahul Kumar Chaurasiya, and Om Prakash Meena

A Granular Access-Based Blockchain System to Prevent Fraudulent Activities in Medical Health Records 635
Megha Jain, Dhiraj Pandey, and Krishna Kewal Sharma

Suspicious Activity Detection in Surveillance Applications Using Slow-Fast Convolutional Neural Network 647
Mitushi Agarwal, Priyanka Parashar, Aradhya Mathur, Khushi Utkarsh, and Adwitiya Sinha

Licence Plate Recognition System for Intelligence Transportation Using BR-CNN 659
Anmol Pattanaik and Rakesh Chandra Balabantaray

VLSI Design and Materials

First Principle Study of Mechanical and Thermoelectric Properties of In-Doped Mg₂Si 671
Abdullah bin Chik and Lam Zi Xin

Design and Analysis of Area and Energy-Efficient Quantum-Dot Half Adder Using Intercellular Interaction Technique 679
Neeraj Tripathi, Mohammad Mudakir Fazili, Abhishek Singh, Shivam, and Suksham Pangotra

Observation of Proposed Triple Barrier δ -Doped Resonant Tunneling Diode 687
Man Mohan Singh, Ajay Kumar, and Ratneshwar Kr. Ratnesh

A Node-RED-Based MIPS-32 Processor Simulator 695
Ethan Anderson, S. M. Abrar Jahin, Niloy Talukder, Yul Chu, and John J. Lee

Simulating Modern CPU Vulnerabilities on a 5-stage MIPS Pipeline Using Node-RED 707
Samuel Miles, Corey McDonough, Emmanuel Obichukwu Michael, Valli Sanghami Shankar Kumar, and John J. Lee

Author Index 717

About the Editors

Pankaj Verma received AMIETE (Associate Member of Institute of Electronics and Telecommunication Engineering) degree from Institute of Electronics and Telecommunication Engineers, New Delhi in 2009, M.Tech. in Microwave and Optical Communication from Delhi Technological University (formerly Delhi College of Engineering) 2011 and Ph.D. from National Institute of Technology, Kurukshetra in 2017. He has been awarded Gold Medal in AMIETE (2009). Currently, he is working as Assistant Professor in Electronics and Communication Engineering Department, National Institute of Technology, Kurukshetra, India. He has published several research papers in International/National journals and conferences. He is also having one-year industrial experience in Intellectual Property (IP) domain and earlier served Indian Air Force for three years. He has served as Conference Chair in “International Conference on Cutting-Edge Technologies in Computing and Communication Engineering (IC4E2020). He has also given many expert talks in various institutions across the country and has served as session chairs in many conferences. His research interests are in wireless communications, cognitive radio systems, optical communication, signal processing, visible light communication, security, machine learning, artificial intelligence, and photonics crystal fiber sensors.

Chhagan Charan is working as Assistant Professor in the Department of Electronics and Communication Engineering, National Institute of Technology Kurukshetra, India. From 2011 to 2013, he was employed as System Engineer in Tata Consultancy Services (TCS). He received his B.Tech. (Bachelor of Technology) degree in Electronics and Communication Engineering from the University of Rajasthan, Jaipur in 2009 and M.Tech. degree (Master of Technology) in Microwave and Optical Communication Engineering from Delhi Technological University (DTU), Delhi in 2011. He completed his Ph.D. degree in Electronics and Communication Engineering at National Institute of Technology, Kurukshetra in 2018. He has experience of about 09 years in teaching and industry. He has authored many research papers in International and National journals and conferences. His research interests include cognitive radio networks, NOMA, digital communication, microwave engineering,

wireless sensor networks, wireless communications, MIMO systems, next generation of communication systems, data computing, and IoT.

Xavier Fernando is Professor at the Department of Electrical and Computer Engineering, Ryerson University, Toronto, Canada. He has co-authored over 200 research articles, two books (one translated to Mandarin) and holds few patents and non-disclosure agreements. He is Director of Ryerson Communications Lab that has received total research funding of \$3,185,000.00 since 2008 from industry and government (including joint grants). He was IEEE Communications Society Distinguished Lecturer and delivered close over 50 invited talks and keynote presentations all over the world. He was Member in the IEEE Communications Society (COMSOC) Education Board Working Group on Wireless Communications. He was Chair IEEE Canada Humanitarian Initiatives Committee 2017–18. He was also Chair of the IEEE Toronto Section and IEEE Canada Central Area. He was General Chair for IEEE International Humanitarian Technology Conference (IHTC) 2017 and General Chair of IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), 2014. He has been in the organizing/steering/technical program committees of numerous conferences and journals. He was Member of Board of Governors of Ryerson University during 2011–12. He is Program Evaluator for ABET (USA). He was Visiting Scholar at the Institute of Advanced Telecommunications (IAT), UK in 2008, and MAPNET Fellow visiting Aston University, UK in 2014. Ryerson University nominated him for the Top 25 Canadian Immigrants award in 2012 in which was a finalist. His research interests are in signal processing for optical/wireless communication systems. He mainly focuses on Physical and MAC layer issues. He has special interest in underground communications systems, cognitive radio systems, visible light communications, and wireless positioning systems.

Subramaniam Ganesan is Professor of Electrical and Computer Engineering, Oakland University, Rochester, MI 48309, USA. He has over 25 years of teaching and research experience in Digital systems. He served as Chair of the CSE department from 1991 to 1998. He is with Electrical and Computer Engineering department since 2008. He received his masters and Ph.D. from Indian Institute of Science (IISc) Bangalore, India. He worked at National Aeronautical Laboratory (NAL) India, Ruhr University Germany, Concordia University Canada, and Western Michigan University before joining Oakland University. He is Member of Editorial board of *Journal of Computer Science (JCS)*, *International Journal of Embedded System and Computer Engineering (IJESC)*, *Journal of Concurrent Engineering (CE SAGE publications)*, *International Journal of Information Technology (IJIT)*, *International Journal of Agile Manufacturing (IJAM)*, *Karpagam Journal of Computer Science (KJCS)*. He is also acting as Editor-in Chief, *International Journal of Embedded system and Computer Engineering* and *International Journal of Sensors and Applications*, South Korea. He has received “Life Time Achievement award” for illustrious career in teaching, research, and administration and contribution to the fields of engineering, conferred by ISAM, USA, in December 2012. He is Distinguished Speaker of IEEE

Computer society for 6 years and has visited many countries to give speeches. His research interests include parallel computer architecture, real-time multiprocessor systems for image processing, embedded systems security, DSP processor-based systems and applications, automotive embedded systems and condition-based maintenance, sensor network, mobile protocol and applications, divisible load theory, and applications.

Data Computing

OntoIntAIC: An Approach for Ontology Integration Using Artificially Intelligent Cloud



V. Adithya, Gerard Deepak, and A. Santhanavijayan

Abstract Ontologies are pedagogical models used in knowledge engineering for knowledge representation, management, and integration. Ontologies play a vital role in representation and reasoning of domain knowledge and provisioning background knowledge for artificially intelligent application. This paper proposes a unique strategic framework to integrate ontologies using machine learning with the help of the artificially intelligent cloud which provides enormous computational power and reduces the processing time. First, the existing domain ontologies and contents of the World Wide Web and summarized e-book contents are fed into the cloud, where the preprocessing is done and the data is sampled using mappers and mergers and fed for classification using random forest. The yielded content is used to create a thesaurus where the information gain is computed and optimized using honey bee optimization algorithm. The performance is compared with the baseline approaches and models and was found to be very much superior. The proposed OntoIntAIC architecture achieved an accuracy of 97.33%.

Keywords AI cloud · Honey bee optimization · Mappers · Ontology integration · Random forest · Thesaurus

1 Introduction

Rapid evolution and development in cloud infrastructure have allowed service providers to offer a wide variety of cloud services to users with varying characteristics at a variety of prices, seeking an acceptable service from a growing number of people. Cloud platforms that meet consumer expectations, such as efficiency, cost, and security, have become a major challenge. Artificial intelligence which is also a

V. Adithya

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, India

G. Deepak (✉) · A. Santhanavijayan

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

recent technology has become an important domain in the information technology industry. Cloud computing and artificial intelligence combine in multiple ways, and according to data scientists, artificial intelligence could just be the technology to revolutionize cloud computing solutions.

Artificial intelligence as a service enhances current cloud infrastructure technologies and provides new avenues for growth. It is a well-known fact that text data is always larger in volume when compared to other types of data. To solve problems related to text data natural language processing and knowledge engineering, a field of artificial intelligence is used. Ontologies are one of an effective set of tools used for knowledge engineering which helps in adding real-world knowledge to the data and making the solution more practically acceptable. They reflect a powerful collection of tools for adding semantic meaning to the data which serves as the foundation here for common terminology allowing data to be interpreted in such a way that different data can be connected, even though they are in different formats using different languages, which enables the effective reuse of information. For mixed knowledge problems, integrating these ontologies is required for multiple domain knowledge problems [1].

Motivation: Text form of information has gained significance in this modern age [2], which has led to the formation of a wide variety of ontologies. There are a wide variety of ontologies and ontological resources available. The functionality of ontology depends explicitly on how information is expressed and what sort of rational ability is needed. For every domain, there is a domain ontology but what is the case if the data is not domain specific, so an ontology has to be created that must generalize very well for a problem and integrating these ontologies with an artificially intelligent (AI) cloud can provide an advanced amount of knowledge when compared to the domain ontologies by combining real-world knowledge and forming a thesaurus of ontologies which automatically generalizes well to the data such that it can be used to solve knowledge engineering problems with high accuracy and make the solution more relevant by adding all the real-world knowledge to it. The ontology model aids semantic Web technologies in classifying and integrating information [3].

Contribution: This paper proposes a strategic OntoIntAIC framework for integrating Internet of Things (IoT) ontologies to an artificial intelligence cloud which can prove to be effective to generalize for IoT-related problems. First, the World Wide Web contents, SSN ontologies, IoT ontologies, and sensor ontologies and summarized contents of e-book on IoT, circuit components, and VLSI are passed to the cloud, and the thesaurus comprising of these domain ontologies is also passed to the cloud. When they are mapped accordingly using mappers and are preprocessed using the artificial intelligent classifier, cloud classifies the contents using a random forest classifier and is used for computing information gain, and this is optimized using an optimization genetic algorithm called honey bee optimization and then establishment of the relationship between axioms, axiomatization is done, and then ontology is integrated.

Organization: The remaining part of the paper is as follows: Sect. 2 addresses the relevant research that has been previously done related to this topic. Section 3

briefly explains the architecture proposed. Section 4 consists of the implementation. Section 5 explains the performance and comparison. Section 5 concludes the paper.

2 Related Works

It is very difficult to find a single ontology composed of all related information and expertise for a specific domain that generalizes well for data so integration of ontologies that generalizes well can be done and several people have done research on this topic to find an effective technique to integration ontologies and also using cloud computing to solve knowledge engineering and data science problems. Makwana and Ganatra [1] have explained why knowledge representation is very important, also have said that ontology integration is necessary to solve mixed domain problems, and have presented a technique for clustering of the ontology using global similarity score of a pair of ontologies using the K-means clustering algorithm. Feng and Fan [2] have put forth a strategic approach for deleting invalid records by filling in vacancies in a manual way, and these are used to preprocess semantic ontologies and have done feature extraction using X2 statistics, high-dimensional data are mapped to low-dimensional data, and finally, classification has been performed using CNN. Asfand-E-Yar and Ali [3] have proposed a technique that uses query execution model, and Jena API has been used to retrieve related documents semantically. Their findings are said to demonstrate the feasibility and scalability of the technique. Ramisetty et al. [4] have proposed an ontology service that has been integrated into a manufacturing e-commerce app for product recommendation and have used mapping and merging principles to translate the collaborative requirement of the manufacturing app into appropriate resource specifications. Jiménez-Ruiz et al. [5] presented a technique that can be used to map ontologies using mappers, and they have also given a content map that evaluates the feasibility of their approach. Caldarola and Rinaldi [6] have said that the ability to efficiently and effectively reuse knowledge is a very important thing in knowledge management systems and proposed the approach for ontology reuse by integrating ontologies and can be applied to the food domain. Zhang et al. [7] have used a guided semantic data integration framework that can support data analysis of cancer survival. In [8–20], the usage and classification of various ontologies and relevant semantic approaches have been discussed.

3 Proposed System Architecture

The architecture of the proposed OntoIntAIC is depicted in Fig. 1. AI cloud is used for a reason to simplify or to simplify and parallelize the classification tasks in the environment of highly linked large-scale datasets. So, when these tasks are parallelized, there is better throughput, less computational time, less complexity, more productivity, and less computationally expensive. The proposed OntoIntAIC system

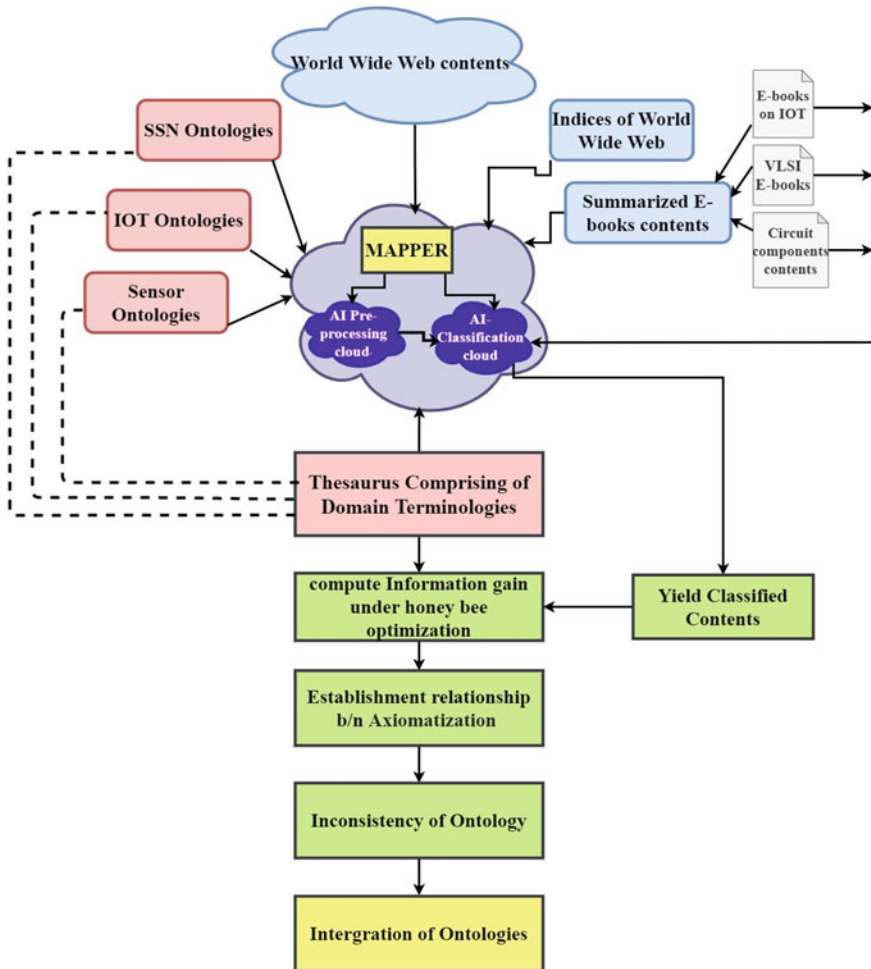


Fig. 1 Proposed architecture diagram of the OntoIntAIC

architecture can be divided into three phases, namely data preparation phase where the data is sampled and fed into the cloud, classification phase where the data is classified using a random forest [21] machine learning algorithm, and the ontology integration phase where the yielded contents information gain is checked using an optimization algorithm and ontology is created and is integrated into integration phase.

Data Preparation Phase: Firstly, for the input to the AI cloud, three distinct ontologies, namely SSN ontologies, IoT ontologies, and sensor ontologies, are chosen in this experiment, apart from these ontologies, based on these ontologies, the indices of the World Wide Web which is based on IoT and similar domains and also the summarized contents of the e-book and the e-book metadata and the table of contents of the

e-book are also taken and sent into a mapper. The mapper is run by an agent which computes the term similarity between the keywords from the summarized e-book contents and relevant indexes with that of the ontological terms resulting in fragmenting the ontologies as smaller subsets and passing it as an input to the AI cloud. This thesaurus containing domain ontologies is used as a topic for classification. Under these topics from the thesaurus, all the inputs from the ontologies as well as all these inputs are mapped using the mapper. The schematic representation of the mappers used is being depicted in Fig. 2.

Since it is highly difficult to handle highly linked data and also going to the high complexity in the size of the data as it is extracted from the World Wide Web and also from the e-book contents, the mapper samples the ontologies along with the summarized e-book indices, indices of the World Wide Web, and the contents of the World Wide Web to make it feasibly fit into this AI architecture. Since data is from several heterogeneous sources, the AI classification cloud requires preprocessed sampled data.

Classification Phase: Since data is from several heterogeneous sources, the AI classification cloud requires preprocessed sampled data. For this AI preprocessing cloud is utilized which does several preprocessing of data like stop word removal, inconsistency removal, redundancy removal, convert to lower case, etc., with the help of NLP algorithms. While the AI cloud is the classification unit, a single mapper feeds in the samples of data into each of the classification sections and also the thesaurus comprising of the domain ontologies is also fed into the individual classification units, and this classification unit classifies based on classification algorithm where 80% is used for training and 20% is used for testing. This classification cloud uses random forest regression to classify the contents.

Random Forest Classification: A supervised machine learning algorithm is used for both regression and classification. Random forest algorithm builds decision trees

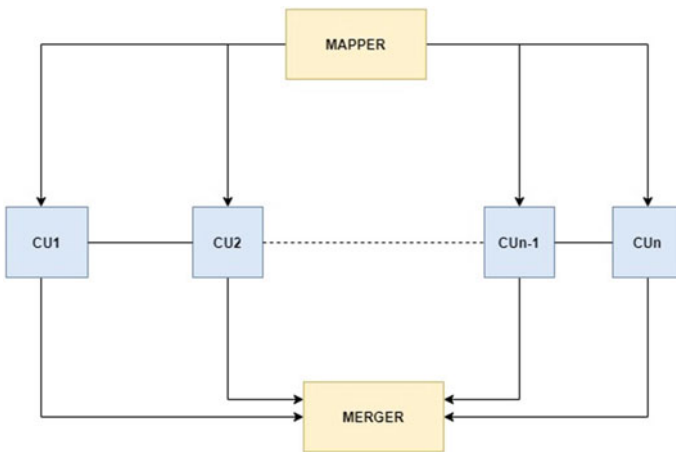


Fig. 2 Schematic representation of the mapper used

on data samples and then gets the estimate from each of them and eventually chooses the best solution by voting. It is an ensemble approach that is stronger than a single decision tree since it eliminates overfitting by integrating the result. This classification algorithm uses a score called as Gini index which is used to find how important a feature is. The higher the Gini index, more important the feature is. The formulae of the Gini index are as shown in Eq. (1).

$$\text{Gini Index} = 1 - \sum_{i=1}^c p_i^2 \quad (1)$$

p_i = relative frequency of classes, c = Number of classes in the dataset.

The classification cloud is nothing but a segment of AI classification units arranged sequentially, and to parallelize this, a mesh architecture is followed.

Ontology Integration Phase: Further the classified contents are yielded, and under each classified content, information gain is computed and is optimized using the honey bee optimization [22] algorithm.

Information Gain: Information gain is the reduction of entropy or surprise by the transformation of a dataset which is also used in training decision trees. It is determined by comparing the dataset entropy before and after transformation. The formulae for calculating information gain are shown in Eq. (2).

$$\text{Gain}(H, A) = H(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} H(S_v) \quad (2)$$

where v is the possible values of A , S is the set of examples $\{x\}$, and S_v is the subset where $X_A = v$ and H is the entropy function.

Honey Bee Optimization: It is a nature-inspired computational algorithm that mimics and uses swarm intelligence of bees to optimize the target value of a problem by iterating continuously. Here the total number of available solutions is compared to the total number of flowers available where n —number of bee's population can be used to visit the flower and calculate its food availability, and here, we call it as a fitness function.

Finally, the most appropriate or the most reliable or relevant entities are yielded. This entity will be linked and will be converted into the taxonomy, and a hierarchical relationship will be established based on the information gain value. Ultimately, the ontologies inconsistency is checked using a pellet reasoner, and finally, the ontologies created will be integrated with the existing ontology, and a bigger ontology is generated.

4 Implementation and Performance Evaluation

The proposed OntoIntAIC system architecture was successfully designed, evaluated, and implemented using a Windows 10 operating system equipped with Intel Core I5 8th generation processor with an 8 GB RAM. An automatic data classifier that classifies data uploaded to the cloud storage was successfully set up using Google cloud services with the help of cloud functions, cloud storage, and cloud data loss prevention services. To design the AI preprocessing, cloud Python script was used that uses various NLP libraries to preprocess the input data, namely NLTK, Re, Num2words, and Sklearn, which were used. Pellet reasoner. Web Protégé is a cloud-based ontology creator which has been used to create and visualize ontologies. Mappers were implemented using Smart Python multi-Agent Development Environment (SPADE), the merger communicates to the merger using agent communication language (ACL), and merger stores a log of the mapped objects. Ontology datasets that were used for this experimentation purposes are IoT-Lite, IoT stream, and IoT-O ontologies along with SSN and sensor ontologies.

The performance metrics were drawn and compared with baseline approaches, and they are tabulated as shown in Table 1.

$$\text{Precision \%} = \frac{\text{True number of Positives}}{\text{True number of Positives} + \text{False number of Positives}} \tag{3}$$

$$\text{Recall \%} = \frac{\text{True number of Positives}}{\text{True number of Positives} + \text{False number of Negatives}} \tag{4}$$

$$\text{Accuracy \%} = \frac{\text{Precision} + \text{Recall}}{2} \tag{5}$$

Table 1 Performance comparison of the OntoIntAIC with baseline models

Metrics	Makwana and Ganatra [1]	Feng and Fan [2]	OntoIntAIC (Proposed)	Eliminating AI cloud from OntoIntAIC with a traditional classifier	Eliminating honey bee optimization from OntoIntAIC
Precision %	82.41	90.31	96.84	92.36	88.89
Recall %	86.81	94.38	98.71	94.72	90.35
Accuracy %	85.89	92.02	97.33	93.89	87.39
F-Measure %	84.55	92.30	97.77	93.53	89.61
Percentage of new and yet relevant concepts discovered	74.69	81.68	88.68	87.69	83.12

$$\text{F-Measure \%} = \frac{2(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (6)$$

$$\text{False Negative Rate} = 1 - \text{Recall} \quad (7)$$

The performance metrics are computed using Eqs. (3–7).

It can be seen that the precision, recall, accuracy, F-measure, and percentage of new and yet relevant concepts discovered increase of the proposed OntoIntAIC framework to the baseline approaches are 14.43%, 11.90%, 11.44%, 13.22%, and 13.99%, respectively, to the Makwana and Ganarat [1] approach that uses global similarity score and K-means clustering for ontology integration, and when compared to Feng and Fan’s [2] paper, it is 6.53%, 4.43%, 5.53%, 5.47%, and 7%, respectively, which uses CNN for ontology integration, and by removing AI cloud and performing the classification on the local machine, the increase was observed to be 4.48%, 3.99%, 3.44%, 4.24%, 0.99%, and while eliminating the honey bee optimization, the increase was observed to be 7.79%, 8.36%, 9.94%, 8.16%, and 1.44%.

Figure 3 shows the false discovery rate (FDR) comparison of the baseline models which is shown above, it is seen that the FDR of the proposed OntoIntAIC framework is 0.01 which in turn tells that our model is very superior when compared to the baseline models, and it can also be found that the optimization technique has reduced the FDR rate drastically which is mainly responsible for high performance of the proposed architecture. It is also seen that the proposed OntoIntAIC framework integrating AI cloud with the use of mapper has also contributed to a reduction in FDR due to high computational power offered by the cloud.

It can be seen from Fig. 4 that the processing time of the proposed OntoIntAIC framework is 5.33 ms also very less when compared to the baseline approaches and the baseline model despite using the optimization algorithm, and this is only because of the clouds high-performance computing which reduces time consumption rapidly.

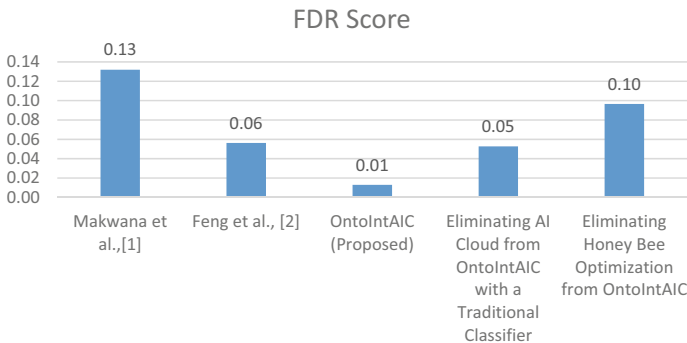


Fig. 3 FDR comparison

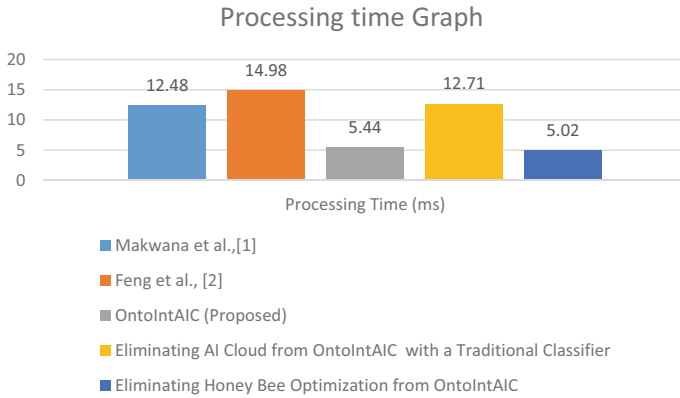


Fig. 4 Processing time comparison

5 Conclusions

There are a large number of domain ontologies existing in the world but to solve a heterogeneous domain-based problem ontology that generalizes well or ontologies that resonates to that specific domain are very much required so the only way to find an ontology like that is to create an ontology by using all the existing ontologies and contents of the World Wide Web and e-book. This paper has proposed an OntoIntAIC framework to that using an AI cloud, and it is seen that the optimization technique and usage of the AI cloud have been very much efficient in archiving a high superiority in terms of performance when compared to the baseline approaches. The processing time was observed to be 5.33 ms which is responsible because the imbibing of the AI cloud and optimization algorithm reduced the FDR to 0.01. It can be concluded that the utilization of AI cloud for knowledge engineering tasks will provide a very high-performance increase, and also as a future work, this can be presented as a functions API and can be hosted on the Internet, where the input of the ontology and topic name and e-Book can be given and the integrated ontology will be given as an output.

References

1. A. Makwana, A. Ganatra, A better approach to ontology integration using clustering through global similarity measure. *J. Comput. Sci.* **14**, 854–867 (2018). <https://doi.org/10.3844/jcssp.2018.854.867>
2. Y. Feng, L. Fan, Ontology semantic integration based on convolutional neural network. *Neural Comput. Appl.* **31**, 8253–8266 (2019). <https://doi.org/10.1007/s00521-019-04043-w>
3. M. Asfand-E-Yar, R. Ali, Semantic integration of heterogeneous databases of same domain using ontology. *IEEE Access* **8**, 77903–77919 (2020). <https://doi.org/10.1109/ACCESS.2020.2988685>

4. S. Ramisetty, P. Calyam, J. Cecil, A.R. Akula, R.B. Antequera, R. Leto, Ontology integration for advanced manufacturing collaboration in cloud platforms, in *Proceedings of the 2015 IFIP/IEEE International Symposium on Integrated Network Management, IM 2015* (2015), pp. 504–510. <https://doi.org/10.1109/INM.2015.7140329>
5. E. Jiménez-Ruiz, B.C. Grau, I. Horrocks, R. Berlanga, Ontology integration using mappings: towards getting the right logical consequences, in *Proceedings of the 6th European Semantic Web Conference*, vol. 5554 of LNCS (2009), pp. 173–187. https://doi.org/10.1007/978-3-642-02121-3_16
6. E.G. Caldarola, A.M. Rinaldi, An approach to ontology integration for ontology reuse, in *2016 IEEE 17th International Conference on Information Reuse and Integration (IRI), Pittsburgh, PA* (2016), pp. 384–393. <https://doi.org/10.1109/IRI.2016.58>
7. H. Zhang, Y. Guo, Q. Li, T.J. George, E. Shenkman, F. Modave, J. Bian, An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival. *BMC Med. Inf. Decis. Making* **18** (2018). <https://doi.org/10.1186/s12911-018-0636-4>
8. V. Adithya, G. Deepak, OntoReq: an ontology focused collective knowledge approach for requirement traceability modelling, in *European, Asian, Middle Eastern, North African Conference on Management & Information Systems* (Springer, Cham, March, 2021), pp. 358–370
9. V. Adithya, G. Deepak, A. Santhanavijayan, HCODF: hybrid cognitive ontology driven framework for socially relevant news validation, in *International Conference on Digital Technologies and Applications* (Springer, Cham, January, 2021), pp. 731–739
10. G.L. Giri, G. Deepak, S.H. Manjula, K.R. Venugopal, OntoYield: A semantic approach for context-based ontology recommendation based on structure preservation, in *Proceedings of International Conference on Computational Intelligence and Data Engineering: ICCIDE 2017*, vol. 9 (Springer, December, 2017), p. 265
11. G. Deepak, N. Kumar, A. Santhanavijayan, A semantic approach for entity linking by diverse knowledge integration incorporating role-based chunking. *Procedia Comput. Sci.* **167**, 737–746 (2020)
12. G. Deepak, S. Rooban, A. Santhanavijayan, A knowledge centric hybridized approach for crime classification incorporating deep bi-LSTM neural network. *Multimedia Tools Appl.*, 1–25 (2021)
13. K. Vishal, G. Deepak, A. Santhanavijayan, An approach for retrieval of text documents by hybridizing structural topic modeling and pointwise mutual information, in *Innovations in Electrical and Electronic Engineering* (Springer, Singapore, 2021), pp. 969–977
14. G. Deepak, V. Teja, A. Santhanavijayan, A novel firefly driven scheme for resume parsing and matching based on entity linking paradigm. *J. Discr. Math. Sci. Cryptogr.* **23**(1), 157–165 (2020)
15. G. Deepak, N. Kumar, G.V.S.Y. Bharadwaj, A. Santhanavijayan, OntoQuest: an ontological strategy for automatic question generation for e-assessment using static and dynamic knowledge, in *2019 Fifteenth International Conference on Information Processing (ICINPRO)* (IEEE, December, 2019), pp. 1–6
16. G. Deepak, A. Santhanavijayan, OntoBestFit: a best-fit occurrence estimation strategy for RDF driven faceted semantic search. *Comput. Commun.* **160**, 284–298 (2020)
17. M. Arulmozhivarman, G. Deepak, OWLW: ontology focused user centric architecture for web service recommendation based on LSTM and whale optimization, in *European, Asian, Middle Eastern, North African Conference on Management & Information Systems* (Springer, Cham, March, 2021), pp. 334–344
18. G. Deepak, J.S. Priyadarshini, Personalized and enhanced hybridized semantic algorithm for web image retrieval incorporating ontology classification, strategic query expansion, and content-based analysis. *Comput. Electr. Eng.* **72**, 14–25 (2018)
19. Z. Gulzar, A.A. Leema, G. Deepak, Pcrs: personalized course recommender system based on hybrid approach. *Procedia Comput. Sci.* **125**, 518–524 (2018)
20. G. Deepak, J.S. Priyadarshini, A hybrid semantic algorithm for web image retrieval incorporating ontology classification and user-driven query expansion, in *Advances in Big Data and Cloud Computing* (Springer, Singapore, 2018), pp. 41–49

21. J. Ali, R. Khan, N. Ahmad, I. Maqsood, Random forests and decision trees. *Int. J. Comput. Sci. Issues (IJCSI)* **9**(5), 272 (2012)
22. B. Yuce, M.S. Packianather, E. Mastrocinque, D.T. Pham, A. Lambiase, Honey bees inspired optimization method: the bees algorithm. *Insects* **4**(4), 646–662 (2013)

Identifying the Most Frequently Used Words in Spam Mail Using Random Forest Classifier and Mutual Information Content



Mohammad A. N. Al-Azawi

Abstract Nowadays, email is an important medium of communication used by almost everyone whether for official or personal purposes, and this has encouraged some users to exploit this medium to send spam emails either for marketing purposes or for potentially harmful purposes. The massive increase in the number of spam messages led to the need to find ways to identify and filter these emails, which encouraged many researchers to produce work in this field. In this paper, we present a method for identifying and detecting spam email messages based on their contents. The approach uses the mutual information contents method to define the relationship between the text the email contains and its class to select the most frequently used text in spam emails. The random forest classifier was used to classify emails into legitimate and spam due to its performance and the advantage of overcoming the overfitting issue associated with regular decision tree classifiers. The proposed algorithm was applied to a dataset containing 3000 features and 5150 instances, and the results obtained were carefully studied and discussed. The algorithm showed an outstanding performance, which is evident in the accuracy obtained in some cases, which reached 97%, and the optimum accuracy which reached 96.4%.

Keywords Spam email · Mutual information content · Random forest classifier

1 Introduction

Due to the availability of the Internet and the increase of people using it daily, email communication has become the most popular medium of communication for personal and official use. According to Statista, the number of daily emails has increased from 269 billion in 2017 to 306 billion in 2020 and is expected to reach 376 billion by 2025 [1]. This massive amount of email has lured some users to send emails in bulk for various purposes, some are with good intentions such as marketing, and others are with bad intentions and intended to harm the users. These emails, or what are

M. A. N. Al-Azawi (✉)

Oman College of Management and Technology, Muscat, Oman
e-mail: mohd.alazawi@omancollege.edu.om

known as spam, are annoying if not harmful and cause a lot of confusion due to the misleading information they contain. The massive increase in the volume of spam has created an urgent need for ways to filter these emails and encouraged researchers to propose different ways to do so.

The use of spam filters is one of the main approaches to identify spam emails based on the analysis of their contents and information [2]. These filters rely on knowledge engineering techniques such as user-defined rules, keywords, or senders blacklists to differentiate legitimate emails from spam/ham emails. Other approaches use AI techniques such as classification and regression to categorise emails. These methods need to extract some features from email messages and use them to train the model which is used then to identify and detect spam emails. To increase the accuracy of the classifier, the used features or words contained in the email are usually increased, and this leads to making the training process slow and sometimes unfeasible. To overcome this issue, feature selection techniques can be used to filter the features and get rid of useless, irrelevant, and redundant features or features with a small impact on the training process and focus on the relevant features only. The feature selection methods are mainly divided into three types, namely filters, wrappers, and hybrid, where most research work focuses on the first two types.

Although a lot of features have been used to identify spam emails, the content of the email itself is still one of the most important features to be considered. Some features such as the sender's address and metadata are not always reliable as it is easy for spammers to change their methods to bypass spam filters. Therefore, in this work, we present a method for classifying emails based on their contents and the relationships between the words the emails contain and the category they belong to.

In this work, we present an algorithm that uses the mutual information contents to identify the most frequently used words in spam emails and uses these words in classifying the emails as legitimate or spam. The algorithm was applied to a dataset containing 3000 features and 5150 cases, and the results obtained were carefully studied and discussed. The algorithm showed excellent performance, which is evident in the accuracy obtained in some cases, which reached 97%, and the optimum accuracy which reached 96.4%.

The remainder of this paper is structured as follows. Section 2 presents the necessary theoretical background and reviews some of the state-of-the-art work. The proposed approach is presented in Sect. 3. In Sect. 4, a thorough discussion of the obtained results is presented. Finally, the conclusions are derived in Sect. 5.

2 Background and Existing Work

2.1 Spam Email Detection

In this discussion, we categorise emails into three categories, namely legitimate, ham, and spam. The first category is the important emails that contain useful information,

while the other two are unwanted or unimportant emails. Spam email is unsolicited bulk email and using the word unsolicited means that the recipient has not requested or approved receiving this type of email. On the other hand, ham email is not much different from the previous type, but it is sent based on the user's consent or request, as in the case when accepting the terms of reference when downloading software. Ham and spam are both handled in the same way in spam filters but with different features since they are sent in bulk but with different contents. Therefore, what is applied to spam can be applied to ham but with different features that can be derived from their contents.

Most of the proposed techniques use either knowledge engineering or machine learning in designing spam filters. In knowledge engineering-based approaches, the available data is used to create a set of rules and protocols upon which the filter decides whether an email is spam or not. The main limitation of such methods is that the set of rules must be maintained and updated continuously as the spammers update their techniques and tactics constantly. In machine learning-based approaches, the algorithm uses some features extracted from the spam email corpus and uses them to train the model which is used to classify other emails. Although both classification and clustering can be used in machine learning-based approaches, the classification is used more widely as the class of each email under consideration in training is known. Mainly, two types of approaches are used in machine learning-based filters which are single standard machine learning and hybrid machine learning [3, 4].

Nowadays, most of the publications are using artificial intelligence and machine learning techniques in identifying spam emails. Techniques such as neural nets, decision trees, random forest, and others are widely used in such applications. Due to space limitations, we will not review the existing work, instead, we will list some of the recent methods, as given in Table 1, and the reader can refer to references such as [5–7] for more details.

Table 1 Some of the existing work

Research	Approach used	Accuracy (%)
Trivedi and Dey [8]	NB, Bayesian	92.9
Bassiouni et al. [9]	RF, ANN, LR, SVM, random tree, k-NN, Bayes Net, NB, RBF	95.45
Agarwal and Kumar [10]	Naïve Bayes and particle swarm optimisation	95.5
Gaurav et al. [11]	Naive Bayes and decision tree	92.9
Douzi et al. [12]	Neural network and paragraph vector-distributed memory	93–96

2.2 Feature Extraction and Selection

Feature extraction is an important process in solving recognition problems, and these features must be carefully chosen to be useful in describing the problem or the object under consideration. Feature extraction methods aim at converting raw data into a set of meaningful measures that can be processed by identification and recognition algorithms. Several methods have been suggested to extract features from the email data such as those extracted from email text and its attachments [13], or from the metadata such as sender address, subject, browser used, date, sending server, and return path. Some of the extracted features can help identify the type of email and some are not. Sending data and time are examples of features that are not quite useful in the classification process, while features extracted from the email body can be more useful. To get reliable results from any AI model, the model should be trained using adequate data both in the number of features and the number of instances used in the training.

As mentioned above, some features may not be completely helpful in categorising emails and may degrade the classification accuracy; therefore, a supplementary process must be undertaken to identify the best features and eliminate redundant and irrelevant ones from the dataset. Such processes are known as feature selection techniques where the most relevant features are selected and used in the classification process [2].

Let us define \mathcal{S} as the set of all features, i.e., $\mathcal{S} = \{f_i : 0 \leq i \leq N\}$, where f_i is the i th feature and the number of features is $(N+1)$. Furthermore, we shall define \mathcal{S}_I as the important features and \mathcal{S}_U as the unimportant features, where $\mathcal{S} = \mathcal{S}_I \cup \mathcal{S}_U$ and $\mathcal{S}_I \cap \mathcal{S}_U = \phi$. The role of the feature selector is to select the optimum subset \mathcal{S}_I from the set of all features. Two main types of feature selectors are defined which are filters and wrappers. Filter-based methods are faster than wrapper-based ones because they depend on some type of estimation of the importance of individual features. On the other hand, wrapper-based methods are more accurate as the importance of feature subsets is measured using a classification algorithm [14, 15].

In filter-based approaches, the original feature set is fed to a filter to find the best features that can be used in the classification process. To further explain this, let us define the mapping $\mathcal{F} : \mathcal{S} \rightarrow \mathcal{W}$ as the function that identifies the importance (weights) of each feature, where $\mathcal{W} = \{w_i : 0 \leq i \leq N\}$ is the set of weights corresponding to each feature. Finally, a selection criterion (\mathcal{T}) is needed to be specified either by selecting the importance values (weights) or by specifying the number of features we need to select. Figure 1 shows the schematic diagram of the feature selection process using filters. The main limitation of filter-based approaches is that

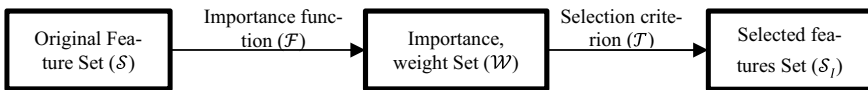


Fig. 1 Schematic diagram of feature selection using filters

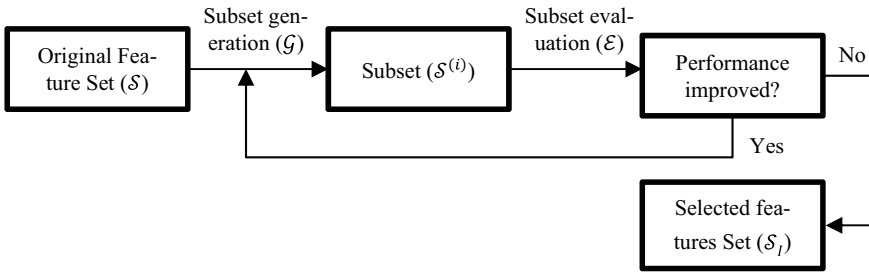


Fig. 2 Schematic diagram of feature selection using wrappers

there is a lack of relationship between the features and the classifier. In addition, they may select irrelevant or redundant features because of the limitation of the evaluation function [16].

Wrappers are more efficient in selecting the features as they decide the best features based on the accuracy obtained from the model. The model might be a classification or a regression model, and it is assumed to be predefined. The process begins with defining a random subset that includes a random number of features. Figure 2 shows the schematic diagram of the wrappers, in which a random subset of features is selected ($S^{(i)}$) using the mapping (G), then the subset is evaluated using the selected model and evaluated the results based on the accuracy obtained (E). To reduce the computation power required, the mapping (G), which is used to select the subset, can be represented by an optimiser.

The wrapper tests all possible feature combinations, where each subset is inserted into the model and the performance is measured in terms of prediction accuracy. This represents a measure of the quality of the specified subset of features.

In wrapper-based approaches, feature presence is represented by 1, and feature absence is indicated by 0. This means that if the number of features is n , the size of the search space is $O(2^n)$. It is sort of impossible to search the whole feature space when n is not small [17]. For example, if the number of features is 100, the search space is $O(1.2 \times 10^{30})$. Several optimisation approaches were used to reduce the calculations and the search space such as nature-inspired optimisation algorithms which include swarm optimisation, farmland fertility, ant lion optimiser, bat algorithm, firefly algorithm, whale algorithm, genetic algorithm, flower pollination, and grey wolf optimiser [17]. The other limitation associated with the wrappers is that the decision of selecting the best features is based on the classifier or the model used, which may not give the best result when used with other models.

2.3 Mutual Information Gain

According to Battiti, mutual information (MI) can be used in selecting the optimum feature for any classification problem [16]. It is assumed that relevant features have a

major impact on the output in any classification problem because they contain important information about it. On the other hand, some other features may be irrelevant or have minimal impact. These features are known as irrelevant features, and eliminating them may not affect the classification accuracy. Many publications utilised mutual information (MI) in feature selection such as [16–19]. Based on the definition of mutual information, it can be used to find the relationship between each feature and the output and then assign an importance value (weight) to that feature.

To measure the information contents, we can use the entropy proposed by Shannon and known as Shannon's information theory (1949). The entropy which is denoted as $H(X)$ can be calculated as given in Eq. (1).

$$H(X) = - \sum_{i=0}^n P(x_i) \log_2 P(x_i) \quad (1)$$

where $X = \{x_0, x_1, \dots, x_n\}$ is a discrete random variable and $P(x_i)$ is the probability of occurrence of the variable x_i .

To calculate the conditional entropy, we need to define another discrete random variable, be it $Y = \{y_0, y_1, \dots, y_m\}$. The conditional entropy $H(Y|X)$ of variable Y is the amount of uncertainty left in variable Y after the variable X is introduced. It can be defined as given in Eq. (2).

$$H(Y|X) = - \sum_{x_i \in X} \sum_{y_j \in Y} P(y_j, x_i) \log_2 P(y_j|x_i) \quad (2)$$

The joint entropy of X and Y , which is denoted as $H(Y, X)$, is the uncertainty that occurs simultaneously with two variables and can be calculated using Eq. (3).

$$\begin{aligned} H(Y, X) &= - \sum_{x_i \in X} \sum_{y_j \in Y} P(y_j, x_i) \log_2 P(y_j, x_i) \\ H(Y, X) &= H(X) + H(Y|X) \end{aligned} \quad (3)$$

Finally, the mutual information between variable Y and variable X , which is denoted as $I(Y; X)$, is given in Eq. (4).

$$\begin{aligned} I(Y; X) &= \sum_{x \in X} \sum_{y \in Y} P(y, x) \log_2 \left(\frac{P(y, x)}{P(x)P(y)} \right) \\ I(X; Y) &= H(X) + H(Y) - H(X, Y) \end{aligned} \quad (4)$$

The mutual information given in Eq. (4) is the amount of information that variable X contains about variable Y , or in other words, it indicates the level of shared information between the random variables. For the variables to be related to each other, the value of $I(X; Y)$ should be high [15].

3 Proposed Algorithm

In the proposed algorithm, we assume that the features used in the classification process are the words contained in the main text of the email. Therefore, the set of features (\mathcal{S}) shall contain the words extracted from the email corpus. Further, we shall define the set $\mathcal{C} = \{c_i; 0 \leq i \leq M\}$ as the set of all output corresponding to each input, where c_i is the class of the email and the number of instances in the dataset is $(M+1)$. The dataset which includes the features set and the corresponding outputs shall be denoted as \mathcal{D} .

The mutual information is calculated for every feature with the output, i.e., $w_i = I(f_i; \mathcal{C})$. This will construct the weight set \mathcal{W} which was defined earlier. The selection criterion must be carefully defined as it must fulfil the balance between accuracy and computational power required. The selection criteria either depend on the number of features or the value of the weights corresponding to each feature. In the first case where the selection process considers the number of features, the weights should be sorted in descending order, and then the features corresponding to the top \mathcal{T}_p percentile are selected to be the optimum features. On the other hand, if the value of the weights is considered as the selection criterion, a specific threshold value \mathcal{T}_T should be selected. These values should be specified manually, which is one of the main drawbacks of filter-based feature selection approaches.

Figure 3 shows the schematic diagram of the algorithm, in which the dataset (\mathcal{D}) is firstly separated into a set of features (\mathcal{S}) and the corresponding output (\mathcal{C}). The mutual information content function ($I(f_i; \mathcal{C})$) is applied to each feature (f_i) with the output set (\mathcal{C}), and the result (w_i) is stored in the weights set (\mathcal{W}). The weights set, which contains the weights corresponding to each feature, then goes through a selection process to select the best features. The selection criterion is specified either by selecting the number of features or the threshold value. Finally, the selected features are used to train the model, which is in this case a random forest classifier.

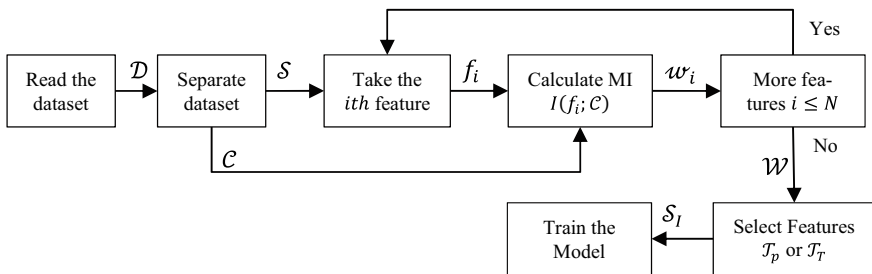


Fig. 3 Proposed algorithm schematic diagram

4 Results and Discussion

The proposed algorithm has been applied to a dataset containing 3000 features and 5150 instances which is available at [20]. The dataset contains the number of occurrences of 3000 words that were used without any natural language pre-processing or removing the common words such as, ‘the’, ‘a’, ‘an’, and ‘and’; this is to avoid prejudging the words before applying the algorithm. The data is divided into two subsets: one for training and the other for testing. 70% of the data (3620 records) were randomly selected for training and 30% (1552 records) for testing.

Table 2 shows a sample of the results obtained; in this table, we studied different percentiles and measured the precision (P), recall (R), F-scale (FS), and accuracy (Acc). The mentioned metrics were measured to both classes, not spam and spam. Furthermore, the macro average (avg) and the weighted average (wa) were calculated for all metrics. Finally, to evaluate the efficiency of the process, we presented a measure that relies on the accuracy obtained and the time needed for training the model. If we assume that the time needed for training the model in the case i is t_i and the accuracy is Acc_i , then the efficiency E_i can be calculated as given in Eq. (5).

$$E_i = (1 - \text{Norm}(t_i)) \times (\text{Norm}(Acc_i)) \quad (5)$$

where $\text{Norm}(\cdot)$ is a normalisation function used to overcome the scale difference between the time and the accuracy.

Figure 4 shows the performance evaluation of the algorithm. In this figure, (a) shows the precision (P) versus the percentile and (b) gives the recall (R) versus the percentile. In these two figures, the precision and recall were calculated for each of the two categories in addition to calculating the macro and weighted averages, (c)

Table 2 Training time, precision, recall, FS, and accuracy versus percentile

Percentile	Time ms	P avg	R avg	FM avg	P wa	R wa	FM wa	Acc	E
1	480	0.88	0.91	0.89	0.91	0.91	0.91	90	0
2	550	0.93	0.94	0.93	0.94	0.94	0.94	94	0.548957
4	680	0.93	0.93	0.93	0.94	0.94	0.94	94	0.507223
6	1090	0.93	0.94	0.93	0.95	0.94	0.94	94	0.375602
10	831	0.95	0.95	0.95	0.96	0.96	0.96	96	0.688122
20	977	0.95	0.95	0.95	0.96	0.96	0.96	96	0.617817
30	1300	0.95	0.95	0.95	0.96	0.96	0.96	96	0.462279
50	1410	0.96	0.96	0.96	0.97	0.97	0.97	97	0.477528
70	1670	0.95	0.95	0.95	0.96	0.96	0.96	96	0.284109
80	1710	0.96	0.96	0.96	0.97	0.97	0.97	96	0.264848
100	2260	0.96	0.96	0.96	0.97	0.97	0.97	96	0

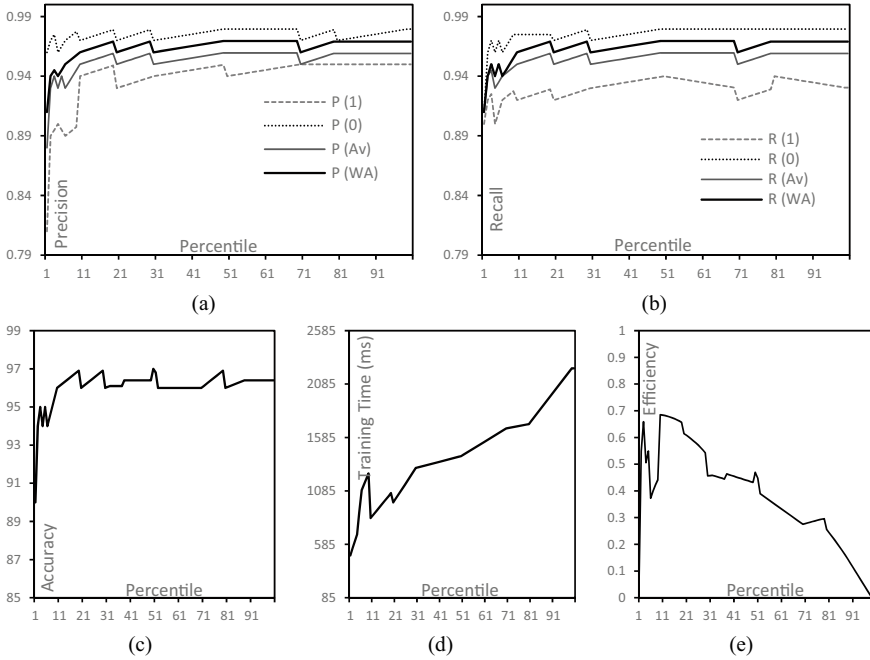


Fig. 4 Performance evaluation, **a** precision versus percentile, **b** recall versus percentile, **c** accuracy versus percentile, **d** training time versus percentile, and **e** efficiency versus percentile

shows the accuracy, which is in this case similar to the *F*-score measure, versus the percentile. Finally, (d) gives the training time versus the percentile, and (e) shows the efficiency versus the percentile.

From the curves given in Fig. 4, it is clear that the performance is strongly influenced by the number of features selected, and the largest change in the curves appears when the percentile value is between 2 and 20 while the curve tends to be nearly constant after that. It is noteworthy that in some cases, selecting more features may degrade the performance as shown in the curves in Fig. 4.

Although the highest accuracy occurred (97%) when the percentile is 50%, the number of selected features is very high (1500 features). By considering the efficiency curve in Fig. 4e, the best performance was when the percentile is 12% which is (96.4%) and the number of selected features is (360) features. Therefore, and based on that, we can find the selection criteria by finding the global maximum from the efficiency curve using any available conventional or AI-based approach.

5 Conclusions

In this research, we presented an approach to identify the most frequently used words in spam emails using the mutual information contents and random forest classifier. The mutual information contents measure was used as it can describe the relationship between each feature (word in our case) and the class the email belongs to. The mutual information contents approach is basically used to reduce and get rid of irrelevant features or features with a small impact on the classification process. In our research, we used a filter-based approach as wrapper-based approaches can work only with models that have only a few features and fail with models that have a large number of features unless a suitable optimisation algorithm is used. The main limitation of filter-based feature selection approaches is that they do not consider the relationship among different features and only consider the relationship between the feature and the output. This did not affect the performance of the algorithm as the relation between the features themselves has been considered in the classification phase.

The algorithm was applied to a dataset containing 3000 features derived from 5150 email messages. The performance of the algorithm was studied carefully, and a new method to find the optimum number of features has been presented and discussed. The algorithm showed an outstanding performance where the accuracy reaches 97% in some cases. But considering the optimum number of features, the performance was 96.4%.

References

1. J. Johnson, Number of e-mails per day worldwide 2017–2025. *Statista* (2021). <https://www.statista.com/statistics/456500/daily-number-of-e-mails-worldwide/>. Accessed 01 March 2021
2. H. Mohammadzadeh, F.S. Gharehchopogh, A novel hybrid whale optimization algorithm with flower pollination algorithm for feature selection: case study email spam detection. *Comput. Intell.* **37**(1), 176–209 (2021). <https://doi.org/10.1111/coin.12397>
3. H. Faris et al., An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks. *Inf. Fusion* **48**(August), 67–83 (2019). <https://doi.org/10.1016/j.inffus.2018.08.002>
4. M. Zhiwei, M.M. Singh, Z.F. Zaaba, Email spam detection: a method of meta-classifiers stacking. *Int. Conf. Comput. Informatics* **200**, 750–757 (2017)
5. A. Bhowmick, S.M. Hazarika, Machine learning for e-mail spam filtering: review, techniques and trends, June 2016. <http://arxiv.org/abs/1606.01042>
6. E.G. Dada, J.S. Bassi, H. Chiroma, S.M. Abdulhamid, A.O. Adetunmbi, O.E. Ajibuwa, Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon* **5**(6) (2019). <https://doi.org/10.1016/j.heliyon.2019.e01802>
7. E.Y. Desta, Spam email detection on data mining: a review. *J. Inf. Eng. Appl.* **9**(2), 1–4 (2019). <https://doi.org/10.7176/jiea/9-2-01>
8. S.K. Trivedi, S. Dey, Interplay between probabilistic classifiers and boosting algorithms for detecting complex unsolicited emails. *J. Adv. Comput. Networks* **1**(2), 132–136 (2013). <https://doi.org/10.7763/JACN.2013.V1.27>

9. M. Bassiouni, M. Ali, E.A. El-Dahshan, Ham and spam e-mails classification using machine learning techniques. *J. Appl. Secur. Res.* **13**(3), 315–331 (2018). <https://doi.org/10.1080/19361610.2018.1463136>
10. K. Agarwal, T. Kumar, Email spam detection using integrated approach of Naïve Bayes and particle swarm optimization, in *Proceedings of the 2nd International Conference on Intelligent Computing and Control Systems, ICICCS 2018*, June 2018 (2019), pp. 685–690. <https://doi.org/10.1109/ICCONS.2018.8662957>
11. D. Gaurav, S.M. Tiwari, A. Goyal, N. Gandhi, A. Abraham, Machine intelligence-based algorithms for spam filtering on document labeling. *Soft Comput.* **24**(13), 9625–9638 (2020). <https://doi.org/10.1007/s00500-019-04473-7>
12. S. Douzi, F.A. AlShahwan, M. Lemoudden, B. El Ouahidi, Hybrid email spam detection model using artificial intelligence. *Int. J. Mach. Learn. Comput.* **10**(2), 316–322 (2020). <https://doi.org/10.18178/ijmlc.2020.10.2.937>
13. U.K. Sah, N. Parmar, An approach for malicious spam detection in email with comparison of different classifiers. *Int. Res. J. Eng. Technol.* **4**(8), 2238–2242 (2017). <https://irjet.net/archives/V4/i8/IRJET-V4I8404.pdf>
14. R.N. Khushaba, A. Al-Ani, A. Alsukker, A. Al-Jumaily, A combined ant colony and differential evolution feature selection algorithm. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **5217** (LNCS, 2008), pp. 1–12. https://doi.org/10.1007/978-3-540-87527-7_1
15. J. Huang, Y. Cai, X. Xu, A hybrid genetic algorithm for feature selection wrapper based on mutual information. *Pattern Recogn. Lett.* **28**(13), 1825–1844 (2007). <https://doi.org/10.1016/j.patrec.2007.05.011>
16. A.I. Sharaf, M. Abu, I. El-Henawy, A feature selection algorithm based on mutual information using local non-uniformity correction estimator. *Int. J. Adv. Comput. Sci. Appl.* **8**(6) (2017). <https://doi.org/10.14569/ijacsa.2017.080656>
17. X. Wang, B. Guo, Y. Shen, C. Zhou, X. Duan, Input feature selection method based on feature set equivalence and mutual information gain maximization. *IEEE Access* **7**, 151525–151538 (2019). <https://doi.org/10.1109/ACCESS.2019.2948095>
18. A. El Akadi, A. El Ouardighi, D. Aboutajdine, A powerful feature selection approach based on mutual information. *Int. J. Comput. Sci. Netw. Secur.* **8**(4), 116–121 (2008). http://paper.ijcns.org/07_book/200804/20080417.pdf
19. S. Verron, T. Tiplica, A. Kobi, Fault detection and identification with a new feature selection based on mutual information. *J. Process Control* **18**(5), 479–490 (2008). <https://doi.org/10.1016/j.jprocont.2007.08.003>
20. B. Biswas, Email spam classification dataset CSV (2020). <https://www.kaggle.com/balaka18/email-spam-classification-dataset-csv>. Accessed 1 Feb 2021

Comparative Analysis of Intelligent Learning Techniques for Diagnosis of Liver Tumor from CT Images



Rutuja Nemane, Anuradha Thakare, Shreya Pillai, Nupur Shiturkar, and Anjitha Nair

Abstract Liver tumor is one of the major causes of deaths worldwide. The early-stage diagnosis of liver tumors is very important and is a time-consuming process. Hence, it is essential to have an automated method for the detection of liver tumors in order to have an accurate and efficient result of the diagnosis. In this paper, the aim is to analyze different intelligent techniques like machine learning, deep learning, and transfer learning to detect liver tumors. A number of deep learning techniques have been implemented for liver tumor detection. But since the amount of training data is less in medical imaging, the CNN models face certain challenges. To overcome these challenges, one of the new emerging techniques called transfer learning is used. This technique is currently being effectively used in the medical imaging field as it shows higher accuracy even with small dataset. Transfer learning can be combined with pretrained CNN models to get better results. This survey highlights various strategies for detecting and segmenting liver images, which will aid in providing an overall summary for future research.

Keywords Liver tumors · Machine learning · Deep learning · Hybridized fully neural network · Transfer learning · AlexNet

R. Nemane (✉) · A. Thakare · S. Pillai · N. Shiturkar · A. Nair
Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, India
e-mail: rutuja.nemane18@pccoepune.org

A. Thakare
e-mail: anuradha.thakare@pccoepune.org

S. Pillai
e-mail: shreya.pillai18@pccoepune.org

N. Shiturkar
e-mail: nupur.shiturkar18@pccoepune.org

A. Nair
e-mail: anjitha.nair18@pccoepune.org

1 Introduction

Liver is one of the most vital organs in the human body [1]. The liver prepares to process a wide range of harmful compounds in your bloodstream when blood leaves your digestive tract and goes into your liver [2]. Due to this, the cancer cells flowing through the blood are highly accessible to the liver. Liver cancer in its early stages begins in the liver but eventually starts spreading to the other body parts. This is the root cause of liver cancers. The majority of liver cancers are secondary or metastatic, which means they began elsewhere in the body. The radiologist's work is complicated because they must confirm the manual diagnosis and segmentation of a liver tumor using a 3D CT scan, which may involve several lesions. Automated techniques are found to be quite effective in the detection and diagnosis of liver tumors. Such procedures assist surgeons in detecting tumors and facilitating care, as well as physicians and radiologists in identifying the affected areas of the liver. Since these techniques ease the process of identification of tumors, they provide clinical assistance to the doctors and surgeons for enhancing the diagnosis, thus improving the survival rate of patients. In this paper, we have analyzed and compared machine learning, deep learning, and transfer learning techniques for the detection of liver tumors. Figure 1 represents some sample CT images of liver tumor [3].

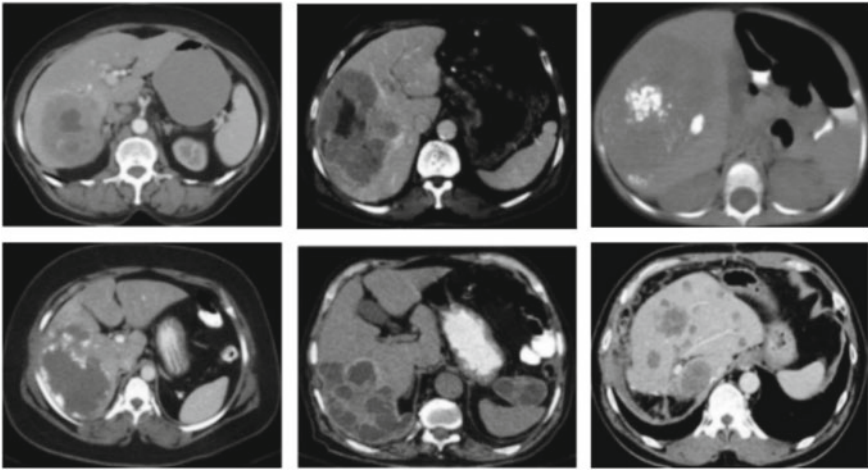


Fig. 1 Examples of CT images of liver tumors

2 Literature Review

2.1 *Analysis of Machine Learning Techniques for Liver Tumor Detection from CT Images*

A work was published on detection of liver tumor with the use of algorithms like SVM and Naïve Bayes [3]. These algorithms are classified on the basis of factors like the accuracy of classification and execution time. Naive Bayes classifier makes an assumption regarding the presence of a specific feature of a class and states that this feature is not related to any other feature. Support vector machines (SVM) are a type of supervised machine learning model used for the classification of linear and nonlinear data and regression analysis. The work stated in this paper has been implemented using Matlab.

A work on machine learning based on MRI scans was published which gives out a system that automates the diagnoses of liver tumor using MRI images [4]. This system provides the result based on normality and abnormality with more accuracy in classification. A spatial fuzzy clustering algorithm was employed for accurately detecting tumor and also to find out the areas of abnormality. Gray-level co-occurrence matrix (GLCM) for texture analysis, multilevel wavelet decomposition, and supervised probabilistic neural network are the different methods that are used during classification.

An another survey work proposes various machine learning methodologies which can be used in order to diagnose liver-related diseases [5]. Some of the algorithms highlighted in this paper are logistic regression, support vector machine, K-nearest neighbor, decision tree, random forest tree, neural network, and ensemble method. The dataset reviewed in this paper is considered in many of the existing techniques. These dataset are related to hepatitis and hepatocellular carcinoma.

A work was published on machine learning-based hybrid feature analysis for liver cancer classification using fused (MR and CT) images [6], which illustrates the power of machine learning methods applied on a dataset of 2D CT images for liver cancer classification. This study employed texture analysis to properly diagnose liver tumors using a fused hybrid dataset based on hybrid feature analysis. The classification accuracy ranged from 95.94 to 98.27%, yielding an effective outcome.

2.2 *Analysis of Deep Learning Techniques for Liver Tumor Detection from CT Images*

A work was published on the segmentation and detection of liver tumor using hybridized fully convolutional neural network based on deep learning framework [7]. In this method, first data augmentation was done during the testing phase to improve the CT data and then features were extracted using various layers of CNNs.

Then, region of interests (ROIs) were distinguished into normal and abnormal lesions. They proposed the HFCNN model for the segmentation and detection of tumors. They designed an ensemble segmentation algorithm for liver segmentation and detection of the region where the tumor is present. The algorithm showed high accuracy in terms of detecting the liver tumor.

A paper was published which showed that a convolutional neural network is one of the effective tools for medical image understanding [8]. Convolutional layers are used to extract features, edges, etc., and then the output of these layers is given into an activation function layer. Then, pooling is done for downsampling of the features. The output of pooling is flattened and then passed to fully connected networks, which compile the data to get the required output.

A work based on modified SegNet, a deep learning model for liver tumor segmentation in CT scans, was published [9]. In this, to make binary segmentation of medical images easier, a binary pixel classification layer was used in place of the classification layer. Here, the dataset utilized for training and testing was a standard 3D-IRCADb-01 dataset. In this method, the steps used for liver tumor segmentation were preprocessing, feature extraction, classification, and post-processing. The suggested technique properly detected most sections of the tumor, with a tumor classification accuracy of over 86 percent. But, upon further examination of the data, they discovered some drawbacks on which they found out that those faults could be reduced if larger dataset is used for training the model.

A work was put forward on a level-set method (LSM) with the initial region of interest and an enhanced edge indicator for liver tumor segmentation from CT images [10]. Segmentation of the liver was done in a coarse-to-fine manner using a 2D U-net and a 3D fully convolutional network. Unsupervised fuzzy C-means (FCM) was employed to improve the edge indicator in the suggested LSM. FCM calculated the likelihood of a tumor being present in a specific image. A combination of deep learning and the level-set method yields good results in medical image segmentation.

2.3 Analysis of Transfer Learning Techniques for Liver Tumor Detection from CT Images

A work was published on transfer learning which improves supervised image segmentation across imaging protocols. In this, transfer learning was used for image segmentation even when the amount of training data was less [11]. In this case, transfer learning was able to reduce the classification errors by up to 60%. They have evaluated following four transfer classifiers:

(1) weighted SVM (WSVM); (2) reweighted SVM (RSVM); (3) TrAdaBoost; (4) adaptive SVM (A-SVM). Among these classifiers, weighted SVM was the most consistent classifier.

A framework was proposed based on unsupervised deep transfer feature learning for medical image classification [12]. Deep learning approaches that are supervised

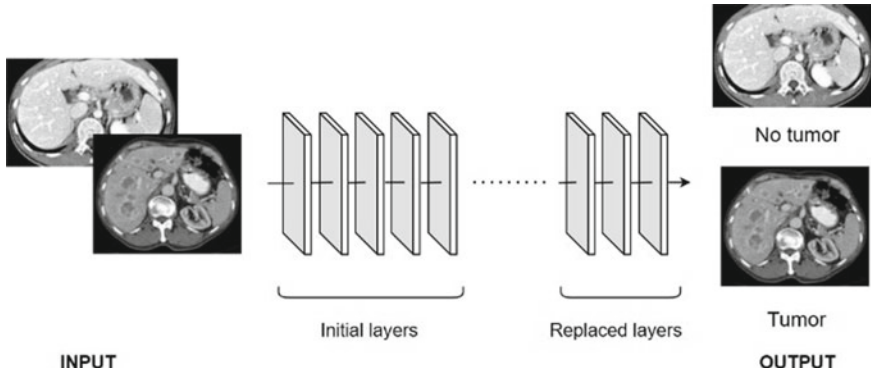


Fig. 2 Automated method for liver tumor classification using transfer learning

require a lot of labeled data. Unsupervised feature learning algorithms can learn from unlabeled data, overcoming these drawbacks. Therefore, a new unsupervised feature extractor is integrated with a convolutional auto-encoder added on top of a pretrained CNN. An accuracy of 81.33% is obtained using this method which is better than the other unsupervised feature learning algorithms.

A work was published on classification of liver cancer histopathology images based on deep learning using only global labels [13]. Early diagnosis of liver tumors was done by classifying the data into abnormal and normal images with the help of just global labels mixed with transfer learning. Here too, transfer learning proved to be better as the amount of training data was less. In transfer learning, some of the last layers are replaced as per the new model requirement and are used for the classification [14] as represented in Fig. 2.

The performance of transfer learning was compared with that of health experts in diagnosis of diseases on the basis of various diseases, different approaches of transfer learning, and many other techniques [15]. It provides a complete insight into all the solutions based on pretrained algorithms used in medical imaging for classifying diseases. They have answered some of the most important research questions which can definitely help the researchers to choose efficient algorithms. According to this comprehensive literature review, the diagnostic accuracy of the transfer learning approach is roughly the same as the diagnosis of health professionals and can be used as it reduces the amount of time and effort required for manual diagnosis.

3 Comparative Analysis of Intelligent Learning Techniques for Diagnosis of Liver Tumor from CT Images

A comparative study of the mentioned three intelligent techniques based on the factors like accuracy, efficiency, size of dataset required, etc., is represented in Fig. 3.

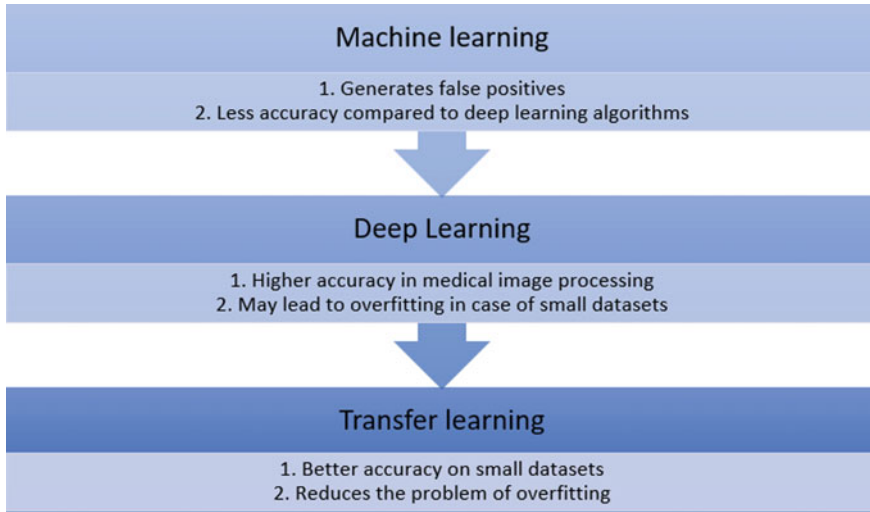


Fig. 3 Comparison between machine learning, deep learning and transfer learning used in medical imaging

It explains why transfer learning is better than other two techniques for medical image processing.

- In case of machine learning, we observed that some of the results were wrongly indicated, that is, they showed false positives. Also, ML techniques were less accurate as compared to deep learning even with the same dataset.
- Deep learning algorithms indicated higher accuracy in case of medical imaging. A lot of work has been done in this field using deep learning. However, these methods may lead to overfitting in case of small dataset.
- To overcome the disadvantages of Deep Learning, we analyzed the transfer learning methods, which are currently emerging and are found to be more efficient. They showed higher accuracy in case of small dataset, thus avoiding the problem of overfitting.

4 Intelligent Algorithms for Liver Tumor Detection

4.1 Otsu's Thresholding Method

In this model [6], image thresholding is used for binarization of the image based on pixel intensities. A grayscale image along with a threshold is passed as the input to this algorithm, and the resulting output consists of a binary image. This method processes image histograms and segments the objects by minimizing the variance on each of the classes. The main task in this method is to divide the image histograms

into two clusters having a threshold that is described as a result of reducing the weighted variance of these classes.

4.2 Weighted SVM

In this algorithm [12], same-distribution, as well as different-distribution training samples are used. The different distribution samples have a lower weight than the same distribution samples. Original SVM definition was used along with sample weighting by assigning a weight $w_i \geq 0$ to every training sample x_i . The sum of all weights, w was equal to the total number of training samples, N . This incorporation of sample weights in the SVM objective function resulted in the following objective function [12]:

$$\min_v \frac{1}{2} \|v\|^2 + C \sum_{i=1}^N w_i \xi_i \quad (1)$$

These constraints and the traditional SVM constraints are same. On performing Transfer Learning, it resulted in the following SVM objective function:

$$\min_v \frac{1}{2} \|v\|^2 + CRw \sum_{i:x_i \in T_d} \xi_i + C \sum_{i:x_i \in T_s} \xi_i \quad (2)$$

This method is known as Weighted SVM (WSVM). It gives a higher classification rate and reduces the effect of outliers. The results of this algorithm were better than the regular SVMs on all the training data.

4.3 Hybridized Fully Convolutional Neural Network

HFCNN has been a powerful tool for image classification and detection. In the paper [7], they designed an ensemble segmentation algorithm for better liver segmentation.

4.3.1 Fully Convolutional Neural Network

The neural network was divided into a two-layer structure of the hidden and visible unit and its energy was expressed as [7]:

$$E(u, g) = - \sum_{j \in \text{visible}} b_j u_j - \sum_{i \in \text{hidden}} a_i g_i - \sum_{j,i} u_j g_i s_{ji} \quad (3)$$

With the help of energy function, they determined the possibilities of hidden layers, then the hidden vectors were summed up, and the likelihood of the network was assigned to a visible vector u , given as:

$$q(u) = \frac{1}{X} \sum_g e^{-E(u,g)} \quad (4)$$

The derivative of a training vector's log-likelihood in terms of weight was shown as:

$$\frac{\partial \log q(u)}{\partial (s_{ji})} = \langle u_j g_i \rangle_{\text{data}} - \langle u_j g_i \rangle_{\text{model}} \quad (5)$$

The following equation was used to measure the hidden unit's binary states:

$$q(g_i = 1|u) = \rho \left(a_i + \sum_j u_j s_{ji} \right) \quad (6)$$

The state of the visible unit was obtained by:

$$q(u_i = 1|g) = \rho \left(b_i + \sum_j g_j s_{ji} \right) \quad (7)$$

The distribution of the k hidden layer g^l and observed layer y of the FCNN model with k layers was shown by:

$$\mathcal{Q}(y, g^1, \dots, g^k) = \left(\prod_{l=1}^{k-2} \mathcal{Q}(g^l | g^{l+1}) \right) \mathcal{Q}(g^{k-1} \cdot g^k) \quad (8)$$

The proposed HFCNN model [7] produced a high precision ratio and also improved the volume error.

5 Evaluation of Related Research

1. Of all the applied classifiers, the MLP classifier had the best performance. The classification accuracy ranged from 95.94 to 98.27%, which is a very promising outcome. Although this method gives an exceptional result, it requires a large dataset for implementation.
2. In hybridized fully neural networks, features were extracted using various layers and the algorithm produced an accurate liver volume measurements of 97.22%

and the segmentation method gave a very good accuracy. Another method used for liver tumor segmentation was SegNet. The advantage of SegNet over traditional auto-encoder architecture is that it is a simple but very powerful adjustment that saves the feature map's max-pooling indices rather than the entire feature map.

- Using pretrained CNN models with Transfer Learning shows a ceiling level accuracy. As discussed above, the Weighted SVM algorithm can be used to get proper segmentation results. With Transfer Learning, the amount of representative training data needed reduces with reduction in classification errors.

5.1 Implementation of Existing Methods

After studying different algorithms of machine learning, deep learning, and transfer learning, we implemented two methods on existing algorithms. First, using CNN and second, using DenseNet (Transfer Learning). Due to the small dataset, CNN resulted into overfitting, showed less accuracy and gave few false positives. In case of pretrained DenseNet model with transfer learning, accuracy is better than CNN that is 72%. Figure 4 represents the comparison of training accuracy and validation accuracy when the model is trained on Liver Tumor images.

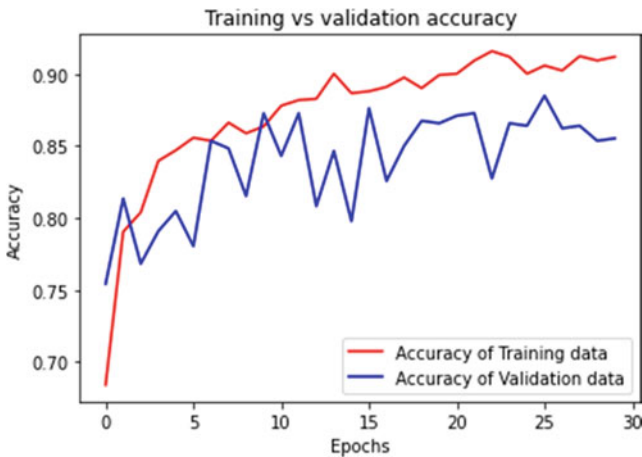


Fig. 4 Training versus validation accuracy

6 Conclusion

In this survey paper for liver tumor detection using intelligent learning methods, we found that previously a lot of methods have been used in the detection of liver cancer. But, due to the limitations of these techniques, some emerging techniques like transfer learning have been developed and are found to be more beneficial in the medical imaging field. We have analyzed various algorithms, methods, and their limitations over each other. Through this survey, we observed that a pretrained DenseNet model with transfer learning showed 72% accuracy. This is an emerging learning technology, which can give better results for small dataset that is “less data, more information.” Further, we will be working on improving the accuracy of the classification of liver tumors based on the survey done on these existing techniques. In order to do so, we need to explore more dataset, so that we can further classify the tumor into different categories.

References

1. P.R. Pruthvi, B. Manjuprasad, B.M. Parashiva Murthy, Liver cancer analysis using machine learning techniques—a review. *Int. J. Eng. Res. Technol. (IJERT) NCICCNDA—2017*. **5**(22) (2017)
2. E. Svoboda, How your liver works. WebMD (2021). <https://www.webmd.com/hepatitis/liver-function>
3. S. Vijayarani, S. Dhayanand, Liver disease prediction using SVM and Naive Bayes algorithms. *Int. J. Sci. Eng. Technol. Res. (IJSETR)* **4**, 816–820 (2015)
4. S. Devi, A. Sruthi, S. Jothi, MRI liver tumor classification using machine learning approach and structure analysis. *Res. J. Pharm. Technol.* **11**, 434 (2018). <https://doi.org/10.5958/0974-360X.2018.00080.X>
5. V. Ramalingamdran, A. Pandian, R. Ragaven, Machine learning techniques on liver disease—a survey. *Int. J. Eng. Technol.* **7**(4.19), 485–495 (2018). <https://doi.org/10.14419/ijet.v7i4.19.23207>
6. S. Naeem, A. Ali, S. Qadri, W. Khan Mashwani, N. Tairan, H. Shah, M. Fayaz, F. Jamal, C. Chesneau, S. Anam, Machine-learning based hybrid-feature analysis for liver cancer classification using fused (MR and CT) images. *Appl. Sci.* **10**(9), 3134 (2020). <https://doi.org/10.3390/app10093134>
7. X. Dong, Y. Zhou, L. Wang, J. Peng, Y. Lou, Y. Fan, Liver cancer detection using hybridized fully convolutional neural network based on deep learning framework. *IEEE Access* **8**, 76056–76068 (2020). <https://doi.org/10.1109/ACCESS.2020.2988647>
8. D.R. Sarvamangala, R.V. Kulkarni, Convolutional neural networks in medical image understanding: a survey. *Evol. Intel.* (2021). <https://doi.org/10.1007/s12065-020-00540-3>
9. S. Almutairi, G. Kareem, M. Aouf, B. Almutairi, M.A. Salem, Liver tumor segmentation in CT scans using modified SegNet. *Sensors* **20**(5), 1516 (2020). <https://doi.org/10.3390/s20051516>
10. Y. Zhang et al., Deep learning initialized and gradient enhanced level-set based segmentation for liver tumor from CT images. *IEEE Access* **8**, 76056–76068 (2020). <https://doi.org/10.1109/ACCESS.2020.2988647>
11. A. van Opbroek, M.A. Ikram et al., Transfer learning improves supervised image segmentation across imaging protocols. *IEEE Trans. Med. Imaging.* (2015). <https://doi.org/10.1109/TMI.2014.2366792>

12. E. Ahn, A. Kumar, D. Feng, M. Fulham, J. Kim, Unsupervised deep transfer feature learning for medical image classification, in *2019 IEEE 16th International Symposium on Biomedical Imaging* (2019), pp. 1915–1918. <https://doi.org/10.1109/ISBI.2019.8759275>
13. C. Sun, A. Xu et al., Deep learning-based classification of liver cancer histopathology images using only global labels. *IEEE J. Biomed. Health Inform.* **24**(6), 1643–1651 (2020). <https://doi.org/10.1109/JBHI.2019.2949837>
14. T. Kaur, T. Gandhi, Deep convolutional neural networks with transfer learning for automated brain image classification. *Mach. Vis. Appl.* **31** (2020). <https://doi.org/10.1007/s00138-020-01069-2>
15. H. Malik, M.S. Farooq, A. Khelifi, A. Abid, J. Nasir Qureshi, M. Hussain, A comparison of transfer learning performance versus health experts in disease diagnosis from medical imaging. *IEEE Access* **8**, 139367–139386 (2020). <https://doi.org/10.1109/ACCESS.2020.3004766>

A New Model for COVID-19 Detection Using Chest X-ray Images with Transfer Learning



Vaibhav Jaiswal and Arun Solanki

Abstract The novel coronavirus 2019 (COVID-19) first appeared in China, and it spreads rapidly around the world which is considered pandemic by WHO. It has left the world devastated as daily lives and public health of people around the world have been affected. It is better to detect the positive cases of COVID-19 as soon as possible, so the further spread of virus can be reduced, and the affected patients can be treated quickly. Earlier studies found that with the use of radiology imaging techniques, positive case of COVID-19 can be detected as these images contain information of the epidemic. Utilization of cutting-edge AI methods combined with radiological imaging can be useful for the precise identification of this epidemic and can likewise be assistive to doctors. Training of convolutional neural network (CNN) requires a lot of computation power and time. The COVID-19 dataset is limited, and to train a model from scratch on such dataset requires lot of time for training and heavy computation power. Such existing models trained from scratch do not perform well on dataset with multi-class classification. The study proposes an approach that utilizes transfer learning to make the CNN more efficient with less training and computation cost. In transfer learning, a model trained on very large dataset with high computation power is taken as the starting point. This pre-trained model is then trained on smaller dataset which requires less computation power and time. The proposed model is used to classify COVID-19 versus no finding versus pneumonia on X-ray images. The proposed model is evaluated on labeled and unlabeled test sets which resulted in an accuracy of 97% and 92%, respectively. The study shows that transfer learning approach performs better as compared with models trained from scratch.

Keywords Transfer learning · Computer vision · Convolutional neural network · ResNet · COVID-19 diagnosis · Chest X-ray

V. Jaiswal (✉) · A. Solanki
Gautam Buddha University, Greater Noida, India

A. Solanki
e-mail: asolanki@gbu.ac.in

1 Introduction

The novel coronavirus 2019 (COVID-19) initially appeared in China [1], and it spreads rapidly around the world and is considered pandemic by World Health Organization (WHO) [2]. COVID-19 has left the planet devastated as daily lives and public health of individuals round the world had been affected. The initial certified case of the COVID-19 in India appeared in Kerala [3]. It is better to detect the positive cases of COVID-19 as soon as possible, so the further spread of COVID-19 can be reduced, and the affected patients can be treated quickly. The transmission of this virus is mainly by the droplets which are generated and spread through air when COVID-19 infected person sneezes, coughs, and the droplets fall quickly [4]. A person can be infected by breathing or inhaling in the virus if they get in close proximity of another person infected with COVID-19. Also, a person can be affected by touching contaminated surface and then the sensitive areas of their body like eyes, nose, and mouth [5]. The severe cases of COVID-19 cause pneumonia. Some examples of cutting-edge methods to diagnose COVID-19 are RT-PCR test, rapid testing, and deep learning medical imaging.

1.1 Screening Through X-ray

Recent studies found that using radiology imaging [6] techniques can be used to detect the positive case of COVID-19 as such images contain salient information of the epidemic. The study found that radiological X-ray images provide a good solution for prediction of virus. It gives an accurate prediction to help in diagnostics of COVID-19. As the severe cases of this virus are called COVID-19 induced pneumonia [7], so the major difference in the finding of X-ray radiological images of COVID-19 and pneumonia is: In case of COVID-19 (COVID-19-induced pneumonia), the presence of bi-lateral ground glass opacity (GGO) air consolidation is observed. In case of pneumonia, the presence of central distributed and unilateral ground glass opacity (GGO) air consolidation is observed (Fig. 1).

2 Literature Review

The COVID-19 [9] is considered as pandemic by World Health Organization (WHO). It is affecting the world, and the developing country like India is critically affected by this virus. For detection, RT-PCR [10] is an effective method, but to monitor a COVID-19 patient, constant monitoring is required [11]. Research findings [12] show that chest radiology imaging can be good monitoring method. Yasin and Gouda [13] concluded from their study that X-ray radiology or medical imaging provides good monitoring of COVID-19-induced Pneumonia. Their radiological findings in lung

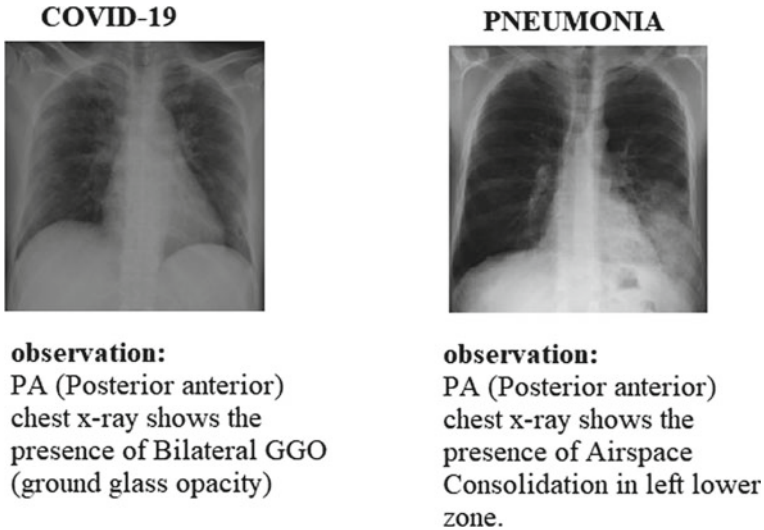


Fig. 1 Observations in COVID-19 and pneumonia X-ray images [8]

conclude that consolidation, ground glass opacity (GGO), and nodular shadowing primarily affect the lower region of lungs. Kanne et al. [14] compared the finding of both COVID-19 and pneumonia through radiology images, and they found that pneumothorax, consolidation without GGO, etc., are rare in COVID-19, but they are often found in pneumonia. Most of them have utilized the deep learning methods for classification task. Ozturk et al. [8] proposed a DCNN model containing 17 convolutional layers, and based on DarkNet-19 [15] model, their model achieved 87.02% accuracy. X. Jiang et al. [16] proposed model with location attention on chest CT dataset, and this model achieved 86.7% accuracy. Wang et al. [17] proposed CNN model, namely COVID-Net which achieved 92.4% accuracy on X-ray dataset containing 53 COVID-19, 5526 pneumonia, 8066 normal images. I. Apostolopoulos et al. [18] proposed transfer learning methodology by using pre-trained VGG-19 as starting point. The image size was rescaled to (200×266) with batch size of 64, and the use of dropout layer and Adam [19] optimizer made the model to obtain 93.48% accuracy. Michael J. Horry et al. [20] proposed transfer learning methodology for multimodal imaging datasets. The model was trained on X-ray dataset which obtained an accuracy of 86%. All the discussed models were trained and tested on labeled dataset, and most of the models did not perform well because of limited COVID-19 images. Solution to this problem is the use of transfer learning on labeled and unlabeled datasets.

Table 1 Detailed description of Dataset 1

Classes	No. of images	Dataset source
COVID-19	125	Cohen et al. [21]
No finding	500	Wang et al. [22]
Pneumonia	500	Wang et al. [22]
Total	1125	

Table 2 Detailed description of Dataset 2

Classes	No. of images	Dataset source
COVID-19	124	Cohen et al. [21]
No finding	500	Daniel S. Kermany et al. [23]
Pneumonia	500	Wang et al. [22]
Total	1124	

3 Data Preprocessing

For the multi-class classification task for COVID-19, dataset is required for accurate detection of disease. For the accurate detection of pneumonia and COVID-19, this study has considered postero-anterior (PA) view of X-ray images for all the classes. For classification, two datasets from different sources are obtained. The contrast and brightness values for COVID-19 chest X-ray images from dataset are not manipulated, and images are in original form. Images are in jpeg, jpg, and png formats with a standard resolution of 256×256 pixels.

Dataset 1: This dataset is obtained from two sources—for COVID-19 class, 125 images were selected from Cohen et al. [21]; for normal (no finding) and pneumonia classes, 500 images of each are selected from Wang et al. [22].

Dataset 2: This dataset is obtained from three different sources—for COVID-19 class, 124 images were selected from Cohen et al. [21]. For normal (no finding) class, 500 images were obtained from Daniel S. Kermany et al. [23]. For pneumonia classes, 500 images were selected randomly from Wang et al. [22] (Tables 1 and 2).

4 Methods and Model Development

The objective of proposed study is to utilize the pre-defined models and train them with the help of transfer learning to provide satisfactory performance on limited set of data.

4.1 Transfer Learning by Freezing Starting Layers of Pre-trained Model

In this type of training, a pre-trained model (trained on larger dataset like the ImageNet dataset consisting of thousands on categories and millions of images) is taken as starting point. The model is first trained by freezing starting layers up to level 3 of convolution base. Because of this, the pre-trained weights in starting convolutional layers do not get updated as these starting layers of convolution generally contain common feature detectors like edges. A custom classifier is added for the proposed task based on the required number of different categories to predict. The later convolutional layers and added custom classifier are trained on dataset (smaller dataset with limited number of images) provided by user. Whole model is then further trained by unfreezing the starting convolutional layers by training the whole model as one.

4.2 Selection of Model

Every year, researchers build models to train and compete on ImageNet dataset [24] competition which leads to advancements in computer vision and image classification field. This proposed work has considered most popular models from ImageNet competition [25], namely (AlexNet [26], MobileNetV2 [27], VGG [28], ResNet [29]). These models are available on the PyTorch library with the ImageNet weights, which is suitable for our task (Fig. 2).

For the selection of suitable models, this study trained the base architecture of all models for 100 epochs with selected learning rate of $3e-3$ after tuning of models. Image size of 256×256 is considered with batch size of 32. For weight updates, Adam optimizer is used, and cross entropy loss as loss function is used. The comparison analysis is provided in Table 3.

Based on the initial results, models are compared, and after comparison, ResNet50 model is selected as this model shows better convergence compared to other models.

4.3 Training and Testing of Proposed Model

Two different datasets are used for training of proposed model. Training of each model is done in two parts: firstly, by freezing model up to level 3 and training remaining layers with proposed classifier module. Secondly, by unfreezing whole model, then training of whole model. Images are resized to 256×256 pixel with batch size = 32. Both datasets are divided into 80/20 (%) split for training and testing of model. For Dataset 1, proposed model was trained for five epochs by freezing starting layers up to level 3 with learning rate = $3e-3$ and then for another ten epochs by

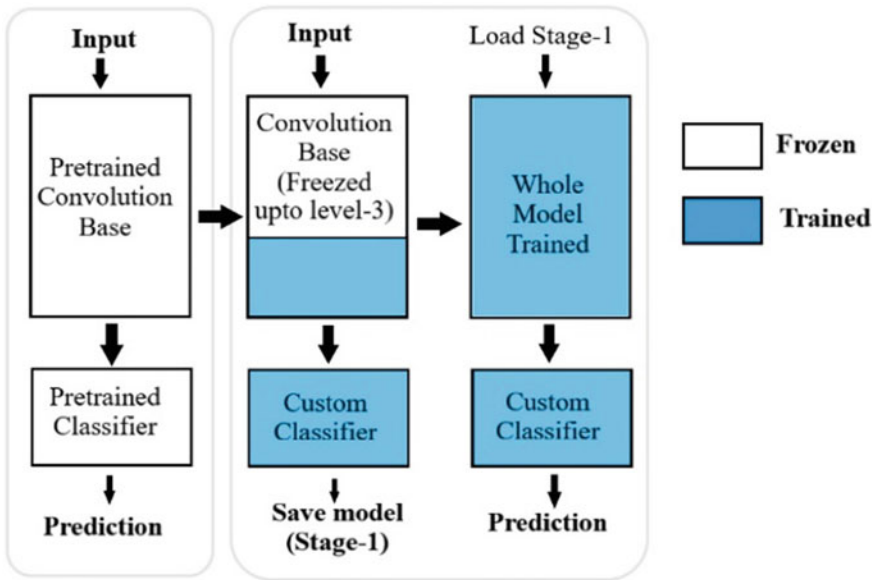


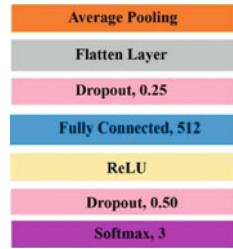
Fig. 2 Basic top-down approach of transfer learning by freezing starting layers

Table 3 Comparison analysis of models for selection

Model	Precision	Recall	F1-score	Accuracy
AlexNet	0.75	0.69	0.71	0.70
MobileNetV2	0.87	0.85	0.85	0.84
VGG16	0.87	0.85	0.86	0.84
VGG19	0.86	0.85	0.85	0.83
ResNet50	0.87	0.86	0.86	0.85

unfreezing whole model with discriminative learning rate in between (1e-5, 1e-4). For Dataset 2, proposed model was trained for five epochs by freezing starting layers up to level 3 with learning rate = 3e-3 and then for another ten epochs by unfreezing whole model with discriminative learning rate in between (2e-6, 1e-5). For weight updates, Adam optimizer is used, and cross entropy loss as loss function is used. Also, early stopping callback [30] function is used. For validation, 20 (%) split set is used for COVID-19, pneumonia, normal X-ray labeled and unlabeled classification. Proposed study has taken external test set with unlabeled data for validation of proposed model. Dataset contains 26 images of COVID-19, 100 images of normal (no finding), and 100 images of pneumonia. Obtained X-ray images are from Cohen et al. [21] for COVID-19, for normal (no finding) class from Daniel et al. [23], and for pneumonia class from Wang et al. [22].

Fig. 3 Architecture of proposed classifier (head)



4.4 Proposed Classifier (Head) Architecture

Architecture starts with average pooling layer followed by flattened layer and two dropout [31] layers (to prevent architecture from overfitting) considered with dropout rate of 0.25 and 0.50, respectively. Lastly, softmax layer is connected (Fig. 3).

5 Result

A comparison analysis of generated results is provided in Sects. 5.1 and 5.2.

5.1 Comparison of Base Model with Fine-Tuned Proposed Model

A comparison of plotted graphs is provided. In Fig. 4a, that validation loss on comparison with train loss first increases significantly in early stages of training, but in later stages of training, it decreases substantially. The loss values show that this model does not converge very well. This is due to the limited number of images in dataset

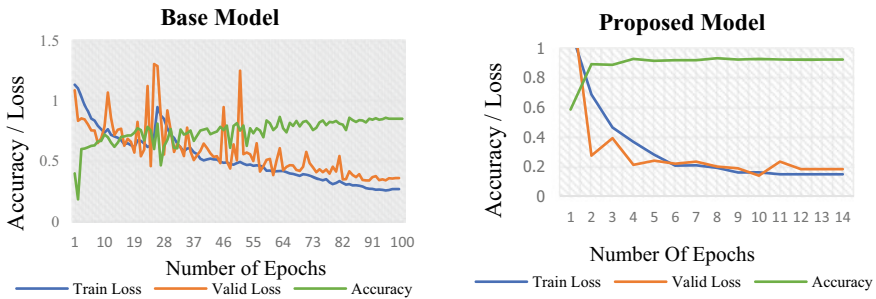


Fig. 4 Graph showing train loss, validation loss, and accuracy metrics of **a** ResNet-50 base model (on left) and **b** proposed model (on right)

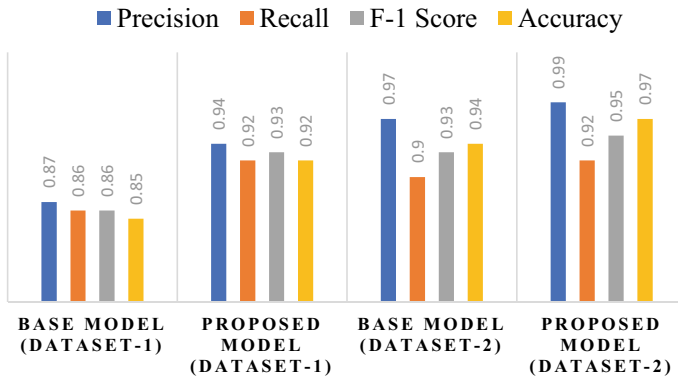


Fig. 5 Comparison of precision, recall, F1-score, and accuracy for base ResNet-50 model with proposed model on Dataset 1 and Dataset 2

which limits the performance. In Fig. 4b, we can observe that validation loss and train loss are decreasing with more training. Training of model was stopped after 15 epochs because of early stopping callback function as the model stopped improving after 15 epochs.

In Fig. 5, the comparison of ResNet-50 and proposed that are trained on Dataset 1 and Dataset 2 is done. The accuracy of base ResNet-50 model is 85% on Dataset 1 and on Dataset-2 is 94%, respectively. The accuracy of proposed model on Dataset 1 is 92% with F1-score of 93% and on Dataset 2 is 97% with F1-score of 95%, respectively. On comparison of base architecture with the proposed method, this study found that proposed model performs better with respect to its version which was trained from scratch.

5.2 Comparison with Existing Work

This proposed work performed better in comparison with existing work with sensitivity and specificity of 91% and 98% and accuracy of 97% (on labeled dataset) and 92% (on unlabeled test set), respectively (Table 4).

6 Conclusion

Due to the advancement of deep learning techniques in analysis of medical images, the convolutional neural networks (CNNs) have performed very well in disease classification. Also, the features learned by pre-trained CNN models on large-scale datasets like ImageNet are much useful for the task of image classification. This research work proposed a new model for COVID-19 detection using chest X-ray images with

Table 4 Comparison analysis of models

Existing work	Type of images	Number of images	Accuracy (in percentage)
Ioannis et al. [18]	Chest X-ray	224 COVID-19 700 Pneumonia 504 Healthy	93.48
Wang et al. [17]	Chest X-ray	53 COVID-19 5526 Pneumonia 8066 Healthy	92.4
Jiang et al. [16]	Chest CT	219 COVID-19 224 Pneumonia 175 Healthy	86.7
Ozturk et al. [8]	Chest X-ray	125 COVID-19 500 Pneumonia 500 No Findings	87.02
Michael et al. [20]	Chest X-ray	139 COVID-19 190 Pneumonia 400 No Findings	86
Proposed work	Chest X-ray	124 COVID-19 500 Pneumonia 500 No Findings 26 COVID-19 100 Pneumonia 100 No Findings	97.76 (labeled dataset) 92.30 (unlabeled dataset)

the help of transfer learning of the model to provide better diagnostics for multi-class classification (COVID-19 vs. pneumonia vs. normal case). Developed system is able to perform multi-class tasks with an accuracy of 97% on labeled dataset and 92% on unlabeled dataset, respectively. This study also validates that for the smaller dataset with limited number of images, models trained from scratch provide limited performance. The proposed model gives better performance on smaller dataset as the pre-trained weights are used for starting layers and then updated by transfer learning of model. Proposed study found that due to the limited number of COVID-19 chest

X-ray images, almost all the models confuse severe cases of COVID-19 with pneumonia. The solution to this problem is the use of chest CT images which will provide better diagnostics for COVID-19-induced pneumonia and viral pneumonia.

References

1. Y. Guo, Q.D. Cao, Z. Hong, The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak—an update on the status. *Military Med. Res.* **7**(1), 1–10 (2020)
2. C. Sohrabi, Z. Alsafi, N. O’Neill, M. Khan, K. Ahmed, A. Al-Jabir, World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). *Int. J. Surgery.* **76**, 71–76 (2020)
3. M. Andrews, B. Areekal, K.R. Rajesh, J. Krishnan, R. Suryakala, B. Krishnan, C. Muraly, P.V. Santhosh, First confirmed case of COVID-19 infection in India: a case report. *The Indian J. Med. Res.* (2020)
4. G. Kim, M. Kim, S.H. Ra, J. Lee, S. Bae, J. Jung, S.H. Kim, Clinical characteristics of asymptomatic and symptomatic patients with mild COVID-19. *Clin. Microbiol. Infect.* **26**(7), 948 (2020)
5. Y. Liu, L. Yan, L. Wan, T. Xiang, M. Peiris, W. Zhang, Viral dynamics in mild and severe cases of COVID-19. *Lancet. Infect. Dis* **20**(6), 656–657 (2020)
6. H. Shi, X. Han, N. Jiang, Y. Cao, O. Alwalid, C. Zheng, Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. *The Lancet Infect. Dis.* **20**(4), 425–434 (2020)
7. Y. Fang, H. Zhang, J. Xie, M. Lin, L. Ying, Sensitivity of chest CT for COVID-19: comparison to RT-PCR. *Radiology* **296**(2), E115–E117 (2020)
8. T. Ozturk, M. Talo, E. Yildirim, U. Baloglu, O. Yildirim, R. Acharya, Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Comput. Biol. Med.* ELSEWARE (2020). <https://doi.org/10.1016/j.compbimed.2020.103792>
9. T. Singhal, A review of coronavirus disease-2019 (COVID-19). *Indian J. Pediatrics.* **87**, 281–286 (2020)
10. Z.Y. Zu, M.D. Jiang, P.P. Xu, W. Chen, Q.Q. Ni, G.M. Lu, L.J. Zhang, Coronavirus disease 2019 (COVID-19): a perspective from China. *Radiology*, E15–E25 (2020)
11. A. Solanki, T. Singh, COVID-19 epidemic analysis and prediction using machine learning algorithms, in *Emerging Technologies for Battling Covid-19* (Nature Publishing Group, 2021), p. 57
12. E.Y. Lee, M.Y. Ng, P.L. Khong, COVID-19 pneumonia: what has CT taught us?. *The Lancet Infect. Dis.* **20**(4), 384–385 (2020)
13. R. Yasin, W. Gouda, Chest X-ray findings monitoring COVID-19 disease course and severity. *Egypt J. Radiol. Nucl. Med.* **51**, 193 (2020)
14. J. Kanne, B. Little, J. Chung, B.M. Elicker, L. Ketai, Essentials for radiologists on COVID-19: an update—radiology scientific expert panel. *Radiology-PubMed* (2020). <https://doi.org/10.1148/radiol.2020200527>
15. J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in *IEEE Conference on Computer Vision and Pattern Recognition 2017* (2017)
16. X. Xu, X. Jiang, P. Du, Y. Chen, J. Su, G. Lang, Deep learning system to screen coronavirus disease 2019 pneumonia. *arXiv* (2020)
17. L. Wang, Z.Q. Lin, A. Wong, Covid-net: a tailored deep convolutional neural network design for detection of Covid-19 cases from chest x-ray images. *Sci. Rep.* **10**(1), 1–2 (2020)
18. I. Apostolopoulos, T. Mpesiana, Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. *Phys. Eng. Sci. Med.* **43**(2), 635–640 (2020)

19. D.P. Kingma, B. Jimmy, Adam: a method for stochastic optimization. arXiv preprint, arXiv 1412.6980 (2014)
20. M.J. Horry, COVID-19 detection through transfer learning using multimodal imaging data. IEEE Access **8** (2020). <https://doi.org/10.1109/ACCESS.2020.3016780>
21. J.P. Cohen, P. Morrison, L. Dao, K. Roth, T.Q. Duong, M. Ghassemi, COVID-19 image data collection: prospective predictions are the future (2020)
22. X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017)
23. D.S. Kermany, M. Goldbaum, C. Wenzia, Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* (2018)
24. J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in *IEEE Conference on Computer Vision* (2009), p. 14
25. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 211–252 (2015)
26. A. Krizhevsky, I. Sutskever, G. Hinton, ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2012)
27. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.C. Chen, Mobilenetv2: inverted residuals and linear bottlenecks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018)
28. K. Simonyan, Z. Andrew, Very deep convolutional networks for large-scale image recognition. arXiv preprint, no. 1409.1556 (2014)
29. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition., in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016)
30. R. Apoorva, S. Arun, Sequence imputation using machine learning with early stopping mechanism, in *International Conference on Computational Performance Evaluation (ComPE)* (IEEE, 2020)
31. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014)

Violence Detection in Videos Using Transfer Learning and LSTM



Rohan Choudhary and Arun Solanki

Abstract Violence recognition in videos with high accuracy can really help law enforcement agencies to effectively cope up with violent activities and save lives. Researchers are working actively on recognizing violence in video footage. This study proposes a deep learning-based model for detecting the presence of violence in video footage. This model uses several variants of pre-trained EfficientNet convolutional neural network (CNN) architecture, which was introduced by GoogleAI, for extracting spatial features from frames of videos. The frame-level features are then fed into a long short-term memory (LSTM) layer to extract spatiotemporal features. Several combinations of fine-tuning layers and dropout layers are used to increase the accuracy. The proposed model can produce an accuracy of 98% on the benchmark hockey dataset. The comparison of the results produced by this model with state of the art indicates a promising future for this model.

Keywords Violence detection · EfficientNet · Deep learning · CNN · LSTM · Fine-tuning · Dropout

1 Introduction

Violence is the use of physical power or force to damage, threaten, or even kill someone or somebody intentionally, as defined by the Oxford dictionary. Violence is an issue which is faced by each country on this planet. People tend to monitor violent activities using surveillance cameras in public places in order to stop or act against the ones indulged in this kind of an act. There is usually a human monitoring about 10–20 streams and is alone responsible for the surveillance of those areas. Humans tend to make mistakes, and this study aims to minimize the probability of human error in surveillance by presenting a violence detection model using computer vision techniques. The improvement in different profound deep learning techniques, on account of the accessibility of large datasets and computational abilities, has brought

R. Choudhary (✉) · A. Solanki
Gautam Buddha University, Greater Noida, India

about a historic change in the section of computer vision. Thus, we tend to move toward deep learning techniques which allow a computer to learn from data itself and turn out to help humans to increase safety and take appropriate decisions accordingly.

The proposed effort attempts to develop a model that uses deep learning and modern training architectures to yield extremely high accuracies that are comparable to or better than the state of the art [1]. For training purposes, the suggested work employs frames from violent and non-violent video. The features are first extracted using EfficientNet [2], and then spatiotemporal features are extracted using LSTM [3]. Following that, numerous tuning layer combinations (dropout and dense layers) were tested to see which produced the best results.

The paper has been organized in the following form. Section 1 explains the problem and meaning of violence and how people use surveillance to cope up with it. It also briefly explains the model introduced in the proposed work. Section 2 discusses some of the work done in this domain using different techniques and methodologies. Section 3 describes the proposed work by explaining the architecture, steps, concepts, and functioning of the proposed model. Section 4 discusses the results produced by the model on a benchmark dataset and its comparison with previous works. Section 5 describes the possible work that can be done after this work, and the conclusion is Sect. 6.

2 Literature Review

For violence recognition from video, previous researchers have used several approaches which include manual feature extraction, deep learning, and traditional computer vision methodologies. Studies show that deep learning approaches to this problem show better accuracy than the ones which use handcrafted methods for violence recognition.

Use of audio content to detect violence using Bayesian networks is done in Gianakopoulos et al. [4]. Although in our proposed model, we have used just the visual data since in most cases surveillance cameras usually come without the audio content. Sudhakaran and Lanz [1] proposed a deep neural network-based solution that uses a CNN network and ConvLSTM to capture frame-level features from the video. The adjacent frame differences were used to encode the frame-level changes in the video. They evaluated their model on three benchmark datasets: hockey fight dataset, movie dataset, violent flow dataset and achieved an accuracy of 97.1%, 100%, and 94.57%, respectively. Zhou et al. [5] used three kinds of input, i.e., RGB images for spatial networks, optical flow between consecutive frames, and acceleration images for temporal networks. They constructed a FightNet which is then trained with all three kinds of input modalities. The output label was decided by fusing results from different inputs. Their method achieved an accuracy of 97% on the hockey dataset and 100% on the movie dataset [5].

A computationally fast method that used a hybrid “handcrafted/learned” feature framework was proposed in Serrano et al. [6]. Firstly, from each input video sequence,

a representative image is extracted. For classification purposes, a 2D CNN Network was used which achieved an accuracy of 94.6% on the hockey dataset and 99% on the movie dataset. For violence detection, Dai et al. [7] trained a convolutional neural network (CNN) model which extracted statics frames and motion optical flows using a two-stream CNN framework. On top of it, LSTM models are applied which can capture the longer-term temporal dynamics. Along with the deep learning features, audio features and several conventional motions were extracted as complementary information. They achieved a mean average precision of 0.296 after fusing all the features. Nievas et al. [8] used action descriptors like STIP and MoSIFT for fight detection which detected the presence of violence in a video with an accuracy of 90%. It also introduced hockey fights dataset which later became the benchmark for violence detection. Even in our research, we have used the hockey fights dataset [8] for benchmarking performance of our model.

Gong et al. [9] proposed a model based on the MoSIFT algorithm which extracted a low-level description of the videos. The most representative features were then selected using a kernel density estimation (KDE)-based feature selection method. To further process the selected MoSIFTs, a sparse coding method was applied. Tran et al. [10] proposed an approach that used deep 3D convolutional networks (3D ConvNets) for spatiotemporal feature learning. It was found that the Conv3D was able to outperform the 2DConvnets approach. On top of that, the Conv3D approach was conceptually simple, straightforward, and easy to train. Abdali and Al-Tuma [11] proposed a real-time violence detector that extracts spatial features by CNN and LSTM was applied as a temporal relation learning method. Their approach focuses on multiple factors like generality, accuracy, and fast response time. Their deep learning-based model achieved an accuracy of 98% with a speed of 131 frames/sec [11].

3 Proposed Work

The proposed work presents a model using deep learning and by utilizing modern architectures of training methodologies, which can produce very high accuracies capable of being collated to the state of the art or even better [1]. The proposed work uses frames from violent/non-violent video for training purpose. Firstly, the features are extracted using EfficientNet [2] followed by extracting spatiotemporal features using LSTM [3], after which several combinations of tuning layers (dropout and dense layers) were checked for generating the best output.

3.1 Datasets

The aim of this research is to build a deep learning-based model using transfer learning and previous works [1] to revamp the state-of-the-art results. For this

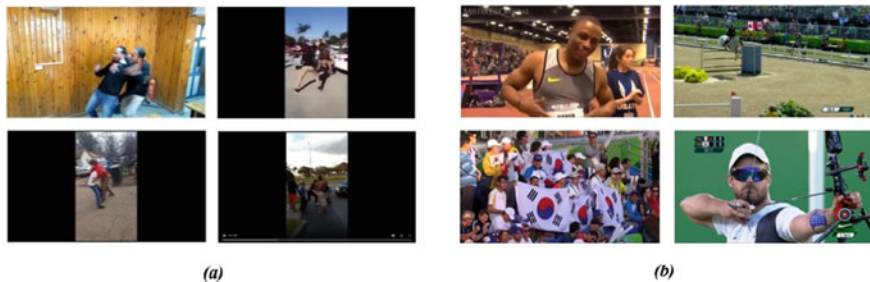


Fig. 1 RLVS training dataset samples. **a** Violence category **b** non-violence category

purpose, choosing a dataset is very crucial. The proposed model is using real-life violence situations (RLVS) dataset [12]. The model’s performance is tested on two state-of-the-art violence datasets, namely hockey fights dataset [8] and violent flow dataset [13]. The hockey dataset is a collection of 1000 videos from the hockey games of the National Hockey League which are split up into violent and non-violent categories [8]. The violent flow dataset comprises of 246 videos of violence captured in football matches in crowds [13]. The RLVS dataset is chosen for training the proposed model because the other datasets were quite similar in terms of backgrounds, environments, and resolution. The RLVS datasets provide a good variation in the dataset regarding environments (e.g., street, prison, football, swimming, weightlifting, eating, etc.) [12]. Apart from that the resolution of the videos in this dataset range from 480 to 720p, Figure 1 shows samples from both categories of the RLVS dataset.

3.2 Preprocessing

The method used in the proposed method consists of the following preprocessing steps:

- Extracting a set of frames belonging to the videos and sending them to several variants of pre-trained network called EfficientNet [2].
- Obtaining the output of one of its final layers and saving it on disk.
- From these outputs, training another network architecture with a type of special neurons called LSTM [3], which is responsible for capturing the temporal features. These neurons have memory and can analyze the temporal information of the video, if at any time they detect violence, it is classified as a violent video.
- The proposed work is extracting 20 frames from each video from both the categories.
- Each frame is resized according to the input requirements of the different pre-trained EfficientNet models used. The sizes of several variants are shown in Table 1. In other words, the variants B0, B1, B2, and B3 use 224×224 , 240×240 ,

Table 1 Resolutions of several EfficientNet variants

Base model	Resolution
EfficientNetB0	224 × 224
EfficientNetB1	240 × 240
EfficientNetB2	260 × 260
EfficientNetB3	300 × 300

260 × 260, and 300 × 300 resolution as input size, respectively. Each object in the proposed model is with the shape (res × res × 3) which corresponds to ($H \times W \times \text{RGB color channels}$).

3.3 Training and Classification

Figure 2 shows the model used in this study. It describes the proposed model’s architecture using a block diagram for each stage of the model. The model training pipeline is discussed in the following section:

- The proposed model processes 20 video frames in batches with the EfficientNet pre-trained model.
- Preceding the last classification layer of the EfficientNet model, the transfer values are saved as a.h5 file. The reason for using a.h5 file is that it requires a lot of time to process a picture with the EfficientNet architecture. If each image is processed more than once, then we can save a lot of time by caching the transfer values.
- At the point where every videos have been handled by the EfficientNet model and the subsequent transfer values are saved to a.h5 file, then those transfer values are reused as the input to a LSTM [3] neural network.
- The proposed model is then trained on the second neural network utilizing the classes from the real-life violence situations dataset (violence and non-violence),

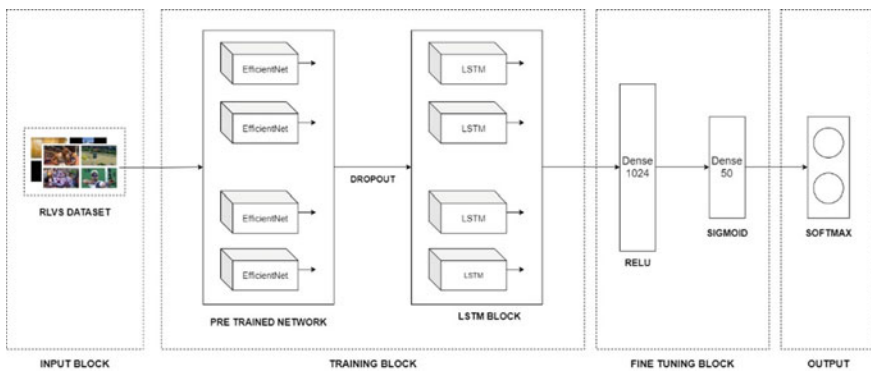


Fig. 2 Block diagram representation of the proposed model

Table 2 EfficientNet parameters for B0, B1, B2, and B3

Base model	Output size	Trainable params	Non-trainable params
EfficientNetB0	1280	5,288,548	42,023
EfficientNetB1	1280	7,794,184	62,055
EfficientNetB2	1408	9,109,994	67,575
EfficientNetB3	1536	12,233,232	87,303

so the network figures out how to classify the frames based on the transfer values from the EfficientNet model.

- The parameters for the variants of EfficientNet used are summarized in Table 2.
- The training dataset is shuffled into two sets: 80% for training and 20% for test.
- After saving the video transfer values to disk, the transfer values are loaded into memory in order to train the LSTM net. When defining the LSTM architecture, we must consider the dimensions of the transfer values.
- From the EfficientNet network, an output a vector of 1280 or 1408 or 1536 transfer values is obtained, depending upon the model of EfficientNet used.
- From each video 20 frames are processed so we will have, for example 20×1280 values per video.
- The classification must be done considering the 20 frames of the video. If any of them detects violence, the video will be classified as violent. The first input dimension of LSTM neurons is the temporal dimension, in our case it is 20.
- The second is the size of the features vector (transfer values).
- A dropout layer is added just before the LSTM layer to combat overfitting by randomly subsampling the previous layer. After an extensive fine-tuning, phase, the value of 0.4 was selected for dropout as it performed the best.
- After the LSTM layer, three dense layers are added each followed by a rectified linear unit (ReLU), Sigmoid and Softmax activation function, respectively.
- The first dense layer after the LSTM layer comprises of 1024 neurons, the second dense layer comprises of 50 neurons, and the last classification layer comprises of 2 neurons for violent and non-violent classes.
- The model is trained for 200 epochs with a batch size of 500. The values of several parameters of the model are summed up in Table 3.

3.4 System Specifications

The proposed model is written in the python programming language, version 3.6.9. It majorly uses the Keras library with TensorFlow as backend and the other libraries used are numpy, matplotlib, os, OpenCV, h5py, sys, random, EfficientNet, etc. Google Colab is used to run the code which has Intel(R) Xeon(R) CPU @ 2.20 GHz, it has 2 CPU cores, 12 GB RAM, 25 Gb disk space, and Tesla T4 GPU with CUDA version 11.2.

Table 3 Proposed model parameters

Parameter	Value
LSTM units	512
Dropout	0.4
Learning rate	0.001
Epochs	200
Batch size	500
Verbose	2
Optimizer	Adam
Loss	Binary cross-entropy

4 Results and Analysis

This section describes the model performance on several variants of the EfficientNet model and shows the model performance in form of graphs using the validation and training accuracy. It also compares the results of this proposed work with previous works.

4.1 Model Performance

The model is trained using four different variants of EfficientNet, B0, B1, B2, and B3. The model is first trained on the RLVS dataset [12], and then the same model is saved and tested on hockey [8] and violent flows dataset [13]. Usually, the validation accuracy after training the model on the benchmark datasets is considered but the proposed model is trained completely on RLVS dataset and then the same model was tested on the benchmark datasets. The model performs good on benchmark datasets. Table 4 describes the results produced by each model on the validation sets along with the performance on the benchmark datasets. Both the variants, EfficientNetB2 and

Table 4 Accuracy and loss achieved by proposed model

	B0	B1	B2	B3
Validation accuracy	0.9106	0.9376	0.9153	0.9141
Validation loss	0.4624	0.3313	0.4230	0.3992
Test accuracy	0.8850	0.9100	0.9175	0.8925
Test loss	0.5435	0.4895	0.4007	0.4770
Hockey dataset accuracy	0.9720	0.9860	0.9800	0.9800
Hockey dataset loss	0.1356	0.0705	0.1112	0.0975
Violent Flows accuracy	0.8943	0.8780	0.8862	0.9024
Violent Flows loss	0.5348	0.6100	0.5567	0.3513

EfficientNetB3, can reach 98% accuracy on the hockey dataset while EfficientNetB3 achieved the highest accuracy for violent flows dataset, i.e., 90.24%. It shows that the RLVS dataset is able to generalize well to several scenarios.

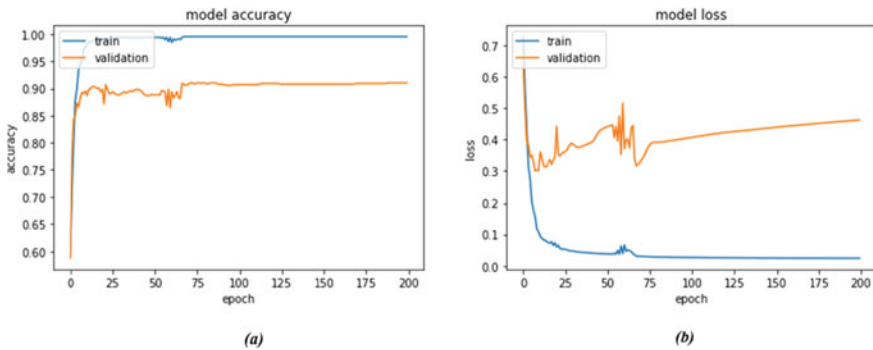
4.2 Graphs

Graphs 1, 2, 3, and 4 show the accuracy and loss graphs attained while training all the variants. The orange-colored line in the graphs depicts the validation accuracy or loss while the blue-colored lines show the training accuracy or loss.

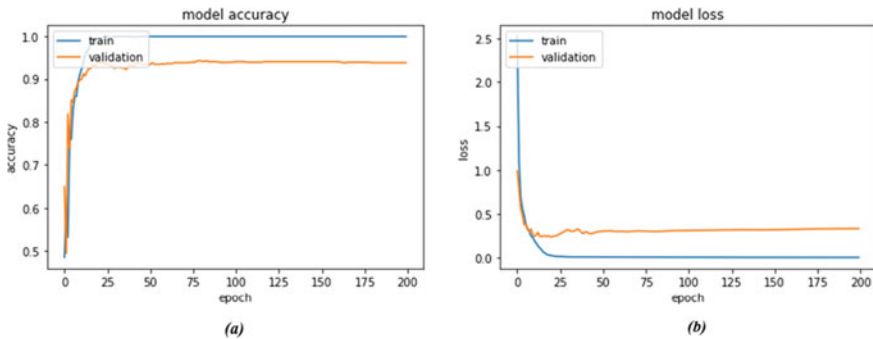
After training the model with EfficientNetB0, the proposed model achieved 91.06% on validation dataset, as we can see in Graph 1.

Graph 2 indicates the performance of EfficientNetB1 model, which was able to achieve 93.76% on validation dataset

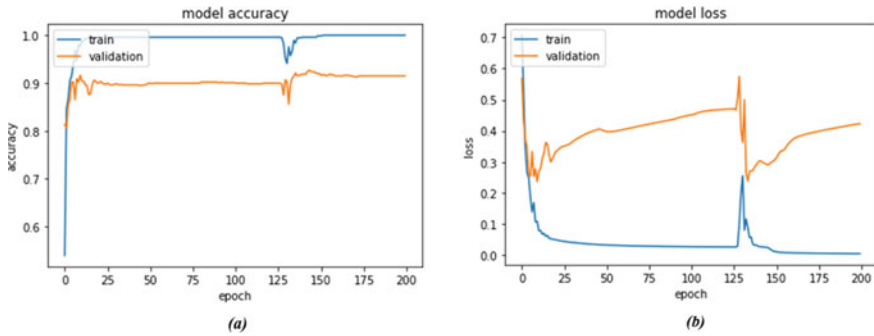
Graph 3 shows the performance of EfficientNetB2 model. It was able to achieve 91.53% on validation dataset



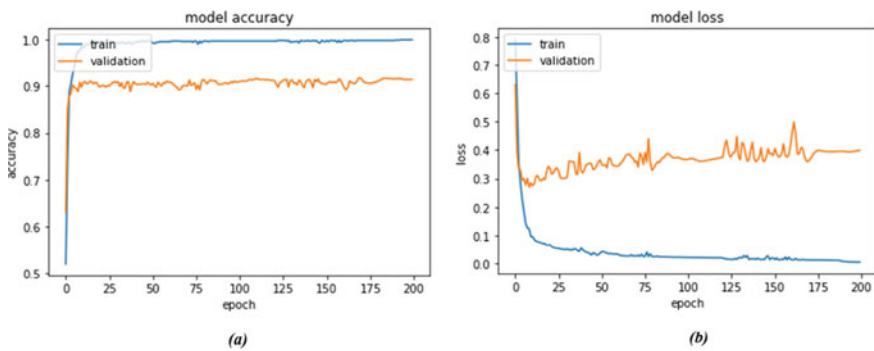
Graph 1 EfficientNetB0 model accuracy and loss graph. **a** Model accuracy and **b** model loss



Graph 2 EfficientNetB1 model accuracy and loss graph. **a** Model accuracy and **b** model loss



Graph 3 EfficientNetB2 model accuracy and loss graph. **a** Model accuracy and **b** model loss



Graph 4 EfficientNetB3 model accuracy and loss graph. **a** Model accuracy and **b** model loss

Graph 4 shows the performance of EfficientNetB3 model on the validation dataset. It was able to achieve 91.41%

The above graphs show that the model has a good learning rate. By examining the space present between the training and validation accuracy, we can conclude that the model is overfitting a bit and the model EfficientNetB1 is the least overfitting model. Several methods like data augmentation, early stopping, regularization, increasing the dataset, trying different splits for training and validation sets can be used to tackle the overfitting issues.

4.3 Comparison with Previous Work

This section compares the results of this study with previous studies done on the same problem statement. To test the generalizing ability of the proposed model and dataset, the model was only trained on the RLVS dataset and then tested on the benchmark datasets. The results generated are for the entire datasets and not a portion of it. The

Table 5 Accuracy contrast with state-of-the-art methodologies

Method	Hockey dataset (%)	Violent flows dataset (%)
MoSIFT + HIK [8]	90.0	–
ViF [13]	82.9 ± 0.14	81.3 ± 0.21
MoSIFT + KDE + Sparse Coding [9]	94.3 ± 1.68	89.05 ± 3.26
Three streams + LSTM [14]	93.9	–
ViF + OViF [15]	87.5 ± 1.7	88 ± 2.45
Sudhakaran and Lanz [1]	97.1 ± 0.55	94.57 ± 2.34
Soliman et al. [12]	95.1	90.01
Proposed work	98.00	90.00

proposed model has achieved very good outcomes on the benchmark dataset, which shows that the RLVS dataset is able to generalize well for several scenarios.

Table 5 compares the result of the proposed model with previous state-of-the-art methodologies. The proposed model can outperform all of them on hockey dataset, while it remains on third position with a margin of 0.01% on the violent flows dataset.

5 Future Work

The proposed work still has room for improvement. Working on the overfitting issue which this proposed work persists is recommended. As we saw that the RLVS dataset was able to generalize well to other datasets, but still there is a lot of scope for improving the violence detection dataset by introducing other categories rather than just violent or non-violent categories, including more scenarios with different environments, etc. Apart from that, there is always an accuracy to speed tradeoff present in studies like these. The proposed work shows the accuracy after training the model using CNN and LSTM. Work can be done to utilize this research work on surveillance footages and classify video frames as violent or non-violent in real time.

6 Conclusion

In the domain of problems related to violence detection, the proposed study presents a deep learning paradigm which can detect violence in videos. The proposed model comprises of two parts, the feature extraction part and special feature detection part. Feature extraction is done using several EfficientNet CNN architectures, and then the saved features are further passed into a LSTM network for spcaiotemporal feature

extraction. After this, the model is connected to three dense layers out of which the last one has two units for violence and non-violence classification. The RLVS dataset is used for training purpose and is later tested on two widely accepted datasets for violence detection.

This paper contributes in two ways: firstly it proposes an end to end model for violence detection using EfficientNet and LSTM. EfficientNets have not been used previously by researchers for violence detection with CNNs and LSTMs. By using this architecture, the proposed study is able to generate state-of-the-art performance on benchmark datasets. Secondly, it checks the generalization ability of the RLVS dataset. The model is never trained on the benchmark datasets, and the same model used for training (on RLVS dataset) is used on benchmark datasets. Also, the model is tested on several variants of EfficientNet and a comparative study is done among those variants. The paper also presents a comparative study on work done before this research. It shows that transfer learning, CNN, and LSTM together can provide very good accuracy for this problem statement, when we are limited by data and computational resources.

References

1. S. Sudhakaran, O. Lanz, Learning to detect violent videos using convolutional long short-term memory, in *14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce* (2017), pp. 1–6
2. M. Tan, Q.V. Le, EfficientNet: rethinking model scaling for convolutional neural networks (2019). [arXiv:1905.11946](https://arxiv.org/abs/1905.11946)
3. S. Hochreiter, J. Schmidhuber, Long short-term memory. *Neural Comput.* **9**, 8, 1735–1780 (November 15, 1997). <https://doi.org/10.1162/neco.1997.9.8.1735>
4. T. Giannakopoulos, A. Pikrakis, S. Theodoridis, A multi-class audio classification method with respect to violent content in movies using Bayesian networks, in *2007 IEEE 9th Workshop on Multimedia Signal Processing* (2007), pp. 90–93. <https://doi.org/10.1109/MMSP.2007.4412825>
5. P. Zhou, Q. Ding, H. Luo, X. Hou, Violent interaction detection in video based on deep learning. *J. Phys: Conf. Ser.* **844**, 012044 (2017). <https://doi.org/10.1088/1742-6596/844/1/012044>
6. I. Serrano, O. Deniz, J.L. Espinosa-Aranda, G. Bueno, Fight recognition in video using Hough forests and 2D convolutional neural network. *IEEE Trans. Image Process.* **27**(10), 4787–4797 (2018). <https://doi.org/10.1109/TIP.2018.2845742>
7. Q. Dai, et al., Fudan-Huawei at MediaEval 2015: detecting violent scenes and affective impact in movies with deep learning, in *MediaEval*, ed. by M.A. Larson, et al. CEUR-WS.org (2015)
8. E.B. Nievas, O.D. Suarez, G.B. Garc, R. Sukthankar, Violence detection in video using computer vision techniques, in *International Conference on Computer Analysis of Images and Patterns*, pp. 332–339, Aug 2011
9. L. Xu, C. Gong, J. Yang, Q. Wu, L. Yao, Violent video detection based on MoSIFT feature and sparse coding, in *ICASSP* (2014). <http://hdl.handle.net/10453/121610>
10. D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks, in *2015 IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 4489–4497. <https://doi.org/10.1109/ICCV.2015.510>
11. A.-M.R. Abdali, R.F. Al-Tuma, Robust real-time violence detection in video using CNN and LSTM, in *2019 2nd Scientific Conference of Computer Sciences (SCCS)* (2019), pp. 104–108. <https://doi.org/10.1109/SCCS.2019.8852616>

12. M.M. Soliman, M.H. Kamal, M.A. El-Massih Nashed, Y.M. Mostafa, B.S. Chawky, D. Khattab, Violence recognition from videos using deep learning techniques, in *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)* (2019). <https://doi.org/10.1109/icicis46948.2019.9014714>
13. T. Hassner, Y. Itcher, O. Kliper-Gross, Violent flows: real-time detection of violent crowd behavior, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (2012), pp. 1–6. <https://doi.org/10.1109/CVPRW.2012.6239348>
14. Z. Dong, J. Qin, Y. Wang, Multi-stream deep networks for person to person violence detection in videos, in *Chinese Conference on Pattern Recognition* (2016)
15. Y. Gao, H. Liu, X. Sun, C. Wang, Y. Liu, Violence detection using oriented violent flows. *Image Vis. Comput.* **48**, 37–41 (2016)

Kernel Functions for Clustering of Incomplete Data: A Comparative Study



Sonia Goel and Meena Tushir

Abstract Clustering of incomplete dataset incorporating missing features is the most prevailing problem in the literature. Several imputations, as well as non-imputation techniques, are utilized to handle this problem. The kernel-based clustering strategies are discovered to be preferred in accuracy over the conventional ones. This paper aims to compare the performance of kernel-based clustering with different kernel functions when applied to incomplete data. These kernel functions have a significant role in kernel-based clustering. The selection of kernel function is neither simple nor insignificant. Gaussian kernel function is considered to be more valuable in different kinds of kernel-based clustering techniques explored in the literature. In this work, we have discussed Gaussian, non-Gaussian, and conditionally positive definite kernel functions for kernel-based clustering for handling missing data in clustering. This investigation is carried over on one artificial dataset and two different real datasets and presents an extensive study of kernel-based fuzzy c-means clustering using different kernel functions for clustering of incomplete data. Numerical analysis shows that other powerful kernel functions exist which work well in specific clustering applications.

Keywords Kernel-based clustering · Kernel functions · Gaussian · Non-Gaussian and conditionally positive definite kernel functions

1 Introduction

In the literature, a lot of emphasis has been laid on the clustering of datasets with missing explicit features. Such datasets are described as incomplete datasets in the literature. Incomplete data problems are regularly examined through imputation,

S. Goel
USICT, GGSIPU, New Delhi, India

S. Goel (✉) · M. Tushir
Department of Electrical and Electronics Engineering, Maharaja Surajmal Institute of
Technology, GGSIPU, New Delhi, India
e-mail: soniagoel@msit.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022
P. Verma et al. (eds.), *Advances in Data Computing, Communication and Security*,
Lecture Notes on Data Engineering and Communications Technologies 106,
https://doi.org/10.1007/978-981-16-8403-6_6

where some particular ways are utilized to replace the missing explicit features [1, 2]. These incomplete datasets are not suitable for clustering because conventional clustering techniques require all features of data. Two approaches are explored in the literature to make incomplete data ideal for clustering. Some researchers highlighted the need for preprocessing of data before clustering [3, 4]. Different statistical methods are utilized to complete the data before clustering, and others customized the conventional clustering algorithms to cluster the incomplete data [5–7]. Data preprocessing is accomplished using various statistical methods. Deletion/removal technique, imputation methods, model-based techniques, and machine learning algorithms are four dominant techniques to analyze missing data are. Deletion/removal technique is the simplest strategy, which deletes the missing data from the data. But this strategy is applicable only when the percentage of incomplete data is very small. Next technique in the row is imputation which replaces the missing features with imputed value. Imputation techniques fill the incomplete data through various statistical methods, i.e., mean imputation, where mean value of all the available features is utilized to replace missing feature. Mean imputation sets up a bias in the data. Median imputation replaces the missing feature with the existing features only. Nearest neighbor method of data imputation replaces the missing attributes with the close assessment in the neighborhood [8]. This approach is precise but takes more time to trace the nearest point. Interpolation is another imputation approach used to estimate the missing value [9]. Three interpolation techniques, namely linear interpolation, quadratic interpolation, and cubic interpolation, are investigated [9]. Linear interpolation imputes the missing values of incomplete data, which attempts to fit a straight line between the two samples and locate the missing element considering the straight-line condition. Each one of the three techniques is discovered to be similarly reasonable for the assessment of missing features [9]. A data-driven approach in which the missing feature of the dataset is alternated by the regression of the available features is presented [10]. Expectation maximization (EM) algorithm, a model-based procedure, is an iterative technique utilized for assessing the missing features randomly to get the optimized outcome by reiterating the procedure until convergence [11]. Further, traditional fuzzy c-means clustering is best suited for spherical clusters only. Over the most recent couple of years, kernel methods have been embedded with fuzzy clustering. They are described as kernel-based fuzzy clustering that clusters the data with different shapes of clusters. The kernel-based classification in the input space keeps the natural grouping of data points and simplifies the related structure of the data [12]. Girolami initially presented the kernel k-means clustering technique for unsupervised classification [13]. Several researchers have shown the predominance of kernel clustering technique over other clustering techniques. The performance of the kernel-based clustering depends extensively on the choice of the kernel functions and estimating their specific constant. The choice of kernel function exclusively depends on the data, though for the particular case of data partitioning, a kernel with universal approximation qualities; for example, RBF is generally fitting. KFCM clustering, which utilizes Gaussian kernel function [14, 15], has been explored for incomplete data. In the proposed work, the linear interpolation technique is used for the imputation of the missing attributes of data, which

is further utilized for exploring some kernel functions for kernel-based clustering of incomplete data.

In this paper, we examine the outcomes of several kernel functions on the clustering of incomplete data. The paper is structured as follows. Section 2 describes the general overview of fuzzy c-means clustering and kernel version of FCM clustering along with different kernel functions. Section 3 elaborates the proposed methodology. Section 4 presents some experiments conducted with several kernel functions on different datasets with different missing proportions, and finally, Sect. 5 has conclusion.

2 Kernel Version of FCM Algorithm

2.1 Review of Fuzzy c-Means (FCM) Clustering

This section reviews FCM algorithm. The conventional fuzzy c-means technique is utilized to partition a set of objects $Z = \{z_1, z_2, \dots, z_s\}$, into fuzzy clusters $p = \{p_1, p_2, \dots, p_c\}$ by minimizing a distance-based objective function:

$$L = \sum_{s=1}^N \sum_{i=1}^c u_{is}^\lambda \|z_s - p_i\|^2 \quad (1)$$

where $s = [1, 2, \dots, N]'$ and $i = 1, 2, \dots, c$, λ is number of cluster centers. λ corresponds to the fuzziness, commonly set to 2, $\|z_s - p_i\|^2$ signifies the distance d_{is} from z_s to i th clustering center. The fuzzy partition vector u_{is} represents the membership degree of sample data z_s to the i th cluster. $u_{is} \in [0, 1]$ and satisfies the condition $\sum_{i=1}^c u_{is} = 1$.

Initial cluster centers are calculated as follows:

$$p_s^0 = \frac{\sum_{s=1}^N (u_{is})^m z_s}{\sum_{s=1}^N (u_{is})^m} \quad 1 \leq i \leq c \quad (2)$$

Therefore, the membership u_{is} is updated as follows:

$$u_{is} = \frac{d_{is}^{\frac{1}{(m-1)}}}{\sum_{i=1}^c \left(d_{Ni}^{\frac{1}{(m-1)}} \right)}, \quad 1 \leq i \leq c \quad (3)$$

$$d_{is} = \|z_s - p_i\|^2 \quad (4)$$

Finally, cluster centers are updated as:

$$p_s = \frac{\sum_{s=1}^N (u_{is})^m z_s}{\sum_{s=1}^N (u_{is})^m} \quad 1 \leq i \leq c \quad (5)$$

2.2 Review of Kernel Fuzzy c-Means (KFCM)

KFCM utilizes the kernel version of fuzzy c-means where kernel induced distance measure [13] is used and is represented as:

$$d(z, p) = \|\vartheta(z) - \vartheta(p)\| = \sqrt{K(z, z) - 2K(z, p) + K(p, p)} \quad (6)$$

Here ϑ is a nonlinear parameter that maps z_k from the input space Z to a high-dimensional new space. The kernel function $K(z, p)$ is given by the inner product in a high-dimensional new space for (z, p) in input space Z .

$$K(z, p) = \vartheta(z)^T \vartheta(p) \quad (7)$$

Numerous kernel functions are investigated in the literature [16, 17].

(i) Gaussian kernel function

$$K(z, p) = \exp\left(-\frac{\|z_s - p_i\|^2}{2\sigma^2}\right), K(z, z) = 1 \quad \text{for all } z \quad (8)$$

The performance of kernel function mainly depends on the value of σ .

The objective function of FCM is altered using Gaussian kernel function (G-KFCM) which is as follows:

$$L_{G-KFCM} = 2 \sum_{s=1}^N \sum_{i=1}^c u_{is}^\lambda (1 - K(z_s, p_i)) \quad (9)$$

(ii) Hyper-Tangent kernel function

It is also known as the sigmoid kernel

$$K(z, p) = 1 - \tanh\left(-\frac{\|z_s - p_i\|^2}{2\sigma^2}\right), K(z, z) = 1 \quad \text{for all } z \quad (10)$$

This kernel function is relatively more accepted for support vector machines because of its basis from neural network theory.

The objective function of FCM is altered using hyper-tangent kernel function (H-KFCM) which is as follows:

$$L_{H-KFCM} = 2 \sum_{s=1}^N \sum_{i=1}^c u_{is}^\lambda \left[\tanh \left(-\frac{\|z_s - p_i\|^2}{2\sigma^2} \right) \right] \quad (11)$$

(iii) Cauchy kernel function

The Cauchy kernel is a long-tailed kernel and can be utilized to give long-range influence and sensitivity over the high-dimensional space.

$$K(z, p) = \frac{1}{1 + \beta \|z_s - p_i\|^2} \quad (12)$$

The objective function of FCM is altered using Cauchy kernel function (C-KFCM) which is as follows:

$$L_{C-KFCM} = \sum_{s=1}^N \sum_{i=1}^c u_{is}^\lambda \left[\left[\frac{1}{1 + \beta \|z_s - p_i\|^2} \right] \right] \quad (13)$$

(iv) Log kernel function

$$K(z, p) = \log \left(\frac{1}{1 + \beta \|z_s - p_i\|^2} \right) \quad (14)$$

The objective function of FCM is altered using log kernel function (L-KFCM) which is as follows:

$$L_{L-KFCM} = \sum_{s=1}^N \sum_{i=1}^c u_{is}^\lambda \log(1 + \beta \|z_s - p_i\|^2) \quad (15)$$

The partition matrix u_{is} and the cluster centers p_i for different kernel-based clustering are updated as

$$u_{is} = \frac{\left(1/(1 - K(z_s, p_i)) \right)^{1/m - 1}}{\sum_{i=1}^c \left(1/(1 - K(z_s, p_i)) \right)^{1/m - 1}} \quad (16)$$

$$p_i = \frac{\sum_{s=1}^N u_{is}^\lambda K(z_s, p_i) z_s}{\sum_{s=1}^N u_{is}^\lambda K(z_s, p_i)} \quad (17)$$

3 Proposed Methodology

This section will elaborate on the methodology adopted for the proposed work to examine the performance of different kernel functions embedded in imputation-based KFCM clustering that can handle incomplete data. The work in this paper assumes that value may miss from the dataset completely at random (MCAR). Missing values in the proposed work are initially estimated by linear interpolation imputation method [18]. The algorithm of proposed work is summarized as:

Algorithm

1. Input: Z incomplete dataset; p -number of clusters, $m = 2$ is a fuzzification parameter, $\epsilon > 0$ is termination criterion.
2. Output: Cluster center p .
3. The given dataset is represented as $Z = [Z_W, Z_M]$.
where $Z_W = \{z_W \in Z | z_W \text{ is a complete datum}\}$ and $Z_M = \{z_M \in Z | z_M \text{ is a incomplete datum}\}$.
4. Initialize all missing features in Z_M with linear interpolation imputation.
5. Initialize p applying standard FCM on complete data.
6. Initialize fuzzy partition matrix randomly.
7. Choose different kernel function, i.e., Gaussian, hyper-tangent, Cauchy and log and their parameters.
Repeat.
8. $p_1 = p$.
9. $Z = [Z_W, Z_M]$.
10. Membership degrees of partition matrix for each kernel function are updated.
11. Cluster center p_1 for each kernel function is updated.
12. Update all missing features in Z_M .
Until $\|p_1 - p\| < \epsilon$ then Stop else go to (5).

4 Numerical Analysis

One artificial dataset and two real datasets, namely IRIS dataset and SEED datasets from UCI repository [19], are used to examine the performance of different kernel functions embedded in imputation-based KFCM clustering that can handle incomplete data. We have made use of three evaluation indices, i.e., misclassification rate, percentage accuracy, and mean prototype error, to evaluate the performance of different kernel functions embedded in imputation-based KFCM clustering that can handle incomplete data. The effectiveness of any clustering algorithm is commonly determined by the recognized indices, namely misclassification rate and accuracy

[20]. Misclassification rate is calculated by taking the sum of imperfectly allocated data points in each cluster. Better clustering results correspond to less value of misclassification rate. Let A is the number of imperfectly allocated data samples in each cluster.

$$\text{Misclassification rate}(M) = \sum_{s=1}^N A_s \quad (18)$$

Percentage accuracy (% accuracy) is calculated as

$$\% \text{ Accuracy} = \frac{N - \sum_{s=1}^N A_s}{N} \times 100 \quad (19)$$

To evaluate the clustering algorithm, mean prototype error is calculated between actual cluster centers and cluster center calculated by algorithm techniques.

$$\text{Error} = \|v_{\text{actual}} - v_{\text{calculated}}\|^2 \quad (20)$$

where v_{actual} and $v_{\text{calculated}}$ signify the actual cluster center and cluster center calculated by different algorithms, respectively.

Different proportions of missing data, i.e., 10, 20, 30, and 40%, are introduced to check the performance of different kernel functions. Same sample of data is generated for all algorithms. All the experiments are carried out in a MATLAB 2019a environment.

First simulation is carried out on a randomly generated artificial dataset. The complete artificial dataset shown in Fig. 1 is 2-D with two unequal-sized clusters with centers $\begin{pmatrix} 4.65 & 0.0001 \\ 16.17 & -0.0004 \end{pmatrix}$. There are 35 points in 1st cluster and 81 points in second cluster. The complete dataset is changed into incomplete datasets by arbitrarily choosing a particular level of missing values, i.e., 10, 20, 30, and 40% with the constraint that one feature value of each sample must be known. Centroids produced

Fig. 1 Artificial dataset
(Cluster centers are
presented by '□')

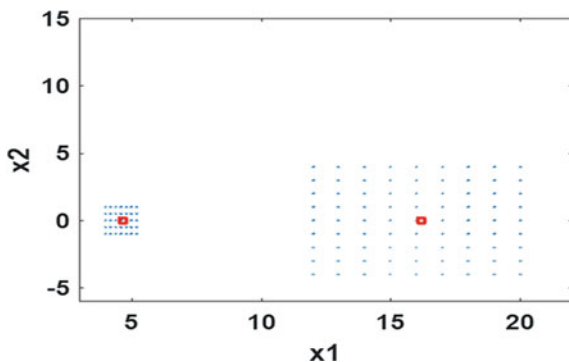


Table 1 Centroids produced by G-KFCM, H-KFCM, C-KFCM, and L-KFCM algorithms for artificial dataset

Clustering algorithm	Centroids produced by all algorithms at different missing rates							
	10%		20%		30%		40%	
G-KFCM $\sigma = 6$	4.59	-0.01	4.61	-0.01	4.61	-0.01	4.61	-0.02
	16.19	-0.05	16.12	-0.07	16.07	-0.10	15.99	-0.002
H-KFCM $\sigma = 6$	4.60	-0.01	4.62	-0.01	4.62	-0.01	4.61	-0.02
	16.12	-0.03	16.06	-0.05	16.02	-0.10	15.96	-0.03
C-KFCM $\beta = 5$	4.58	0.004	4.58	-0.05	4.55	-0.02	4.60	-0.06
	16.00	-0.01	16.00	0.003	16.00	-0.01	15.94	-0.001
L-KFCM $\beta = 5$	4.63	0.003	4.62	-0.07	4.60	-0.03	4.63	-0.08
	16.01	-0.10	16.00	0.011	16.02	-0.10	15.82	-0.04

by clustering algorithms, i.e., G-KFCM, H-KFCM, C-KFCM, and L-KFCM, with different missing rates are shown in Table 1.

Percentage error of Gaussian, hyper-tangent, Cauchy, and log kernel functions when applied on artificial dataset with KFCM clustering at different missing rate of 10–40% with the step size of 10% is calculated and is reported in Table 2. Numerical analysis shows that for the artificial dataset, the performance of hyper-tangent kernel function is comparable with Gaussian kernel function.

To validate the results, experiments are also carried out on real datasets, i.e., IRIS and SEED datasets with different missing rates for misclassification rate (M), % accuracy, and error rate. Table 3 indicates the comparisons of Gaussian, hyper-tangent, Cauchy, and log kernel functions when applied on IRIS data with KFCM clustering at different missing rate of 10–40% with the step size of 10%. At 10% missing rate, Gaussian kernel function performs better than rest of the kernel functions. At 20 and 30% missing rate, the performance of G-KFCM, H-KFCM, and L-KFCM is almost same. However as the missing rate is increased, the performance of Cauchy kernel function deteriorates where misclassification rate and error both are high as compared to other kernel functions.

Experiments are conducted on SEED dataset, and results are presented in Table 4. It is clear from the results that H-KFCM is giving the best results in terms of all indices followed by G-KFCM algorithm.

5 Conclusion

This paper has introduced the thorough analysis of the impact of kernel functions on kernel-based clustering of incomplete data. In the literature, Gaussian kernel function is viewed as the most appropriate kernel function for kernel-based clustering algorithms. We have carried out simulation on two non-Gaussian kernel functions, namely hyper-tangent and Cauchy kernel and a conditionally positive definite log

Table 2 Percentage error with G-KFCM, H-KFCM, C-KFCM, and L-KFCM algorithms for artificial data

Clustering algorithm	Percentage error at different missing rates for artificial dataset											
	10%			20%			30%			40%		
	Ist Cluster	2nd Cluster	Average error	Ist Cluster	2nd Cluster	Average error	Ist Cluster	2nd Cluster	Average error	Ist Cluster	2nd Cluster	Average error
G-KFCM $\sigma = 6$	0.16	0.03	0.09	0.14	0.05	0.10	0.14	0.09	0.12	0.15	0.12	0.13
H-KFCM $\sigma = 6$	0.15	0.05	0.10	0.13	0.06	0.11	0.14	0.10	0.13	0.14	0.15	0.14
C-KFCM $\beta = 5$	0.16	0.12	0.14	0.17	0.12	0.14	0.19	0.12	0.16	0.15	0.16	0.16
L-KFCM $\beta = 5$	0.13	0.12	0.13	0.14	0.12	0.13	0.15	0.11	0.13	0.13	0.27	0.20

Table 3 Error, misclassification rate, and percentage accuracies with G-KFCM, H-KFCM, C-KFCM, and L-KFCM algorithms for IRIS dataset

Clustering algorithm	Percentage of missing proportion for IRIS dataset											
	10%			20%			30%			40%		
	<i>M</i>	%Accuracy	Error	<i>M</i>	%Accuracy	Error	<i>M</i>	%Accuracy	Error	<i>M</i>	%Accuracy	Error
G-KFCM $\sigma = 0.5$	9	94	0.001	12	92	0.001	13	91.3	0.002	14	90.7	0.002
H-KFCM $\sigma = 0.5$	10	93.3	0.001	10	93.3	0.002	13	91.3	0.002	13	91.3	0.002
C-KFCM $\beta = 10$	11	92.7	0.002	12	92	0.002	14	90.7	0.002	19	87.3	0.004
L-KFCM $\beta = 10$	11	92.7	0.001	13	91.3	0.002	14	90.7	0.002	14	90.7	0.002

Table 4 Error, misclassification rate, and percentage accuracies with G-KFCM, H-KFCM, C-KFCM, and L-KFCM algorithms for SEED dataset

Clustering algorithm	Percentage of missing proportion for SEED dataset											
	10%			20%			30%			40%		
	<i>M</i>	%Accuracy	Error	<i>M</i>	%Accuracy	Error	<i>M</i>	%Accuracy	Error	<i>M</i>	%Accuracy	Error
G-KFCM $\sigma = 5$	17	91.9	0.043	19	90.9	0.044	21	90	0.045	22	89.5	0.048
H-KFCM $\sigma = 5$	15	92.9	0.011	16	92.4	0.014	19	90.9	0.026	22	89.5	0.031
C-KFCM $\beta = 0.25$	19	90.9	0.036	20	90.5	0.042	21	90	0.048	24	88.6	0.051
L-KFCM $\beta = 0.25$	19	90.9	0.041	20	90.5	0.041	22	89.5	0.043	24	88.6	0.045

kernel function. We embedded these different kernel functions in KFCM clustering of incomplete data. Extensive analysis on an artificial data and two real-life datasets with different missing rate is done. From this study, we can say kernel functions other than Gaussian kernel function also perform well with different datasets.

References

1. M.N. Ramli, A.S. Yahaya, N.A. Ramli, N.F.F.M. Yusof, M.M.A. Abdullah, Roles of imputation methods for filling the missing values: a review. *Adv. Environ. Biol.* **7**(12 S2), 3861–3870 (2013)
2. A.R.T. Donders, G.J. Van Der Heijden, T. Stijnen, K.G. Moons, A gentle introduction to imputation of missing values. *J. Clin. Epidemiol.* **59**(10), 1087–1091 (2006)
3. K. Kirchner, J. Zec, B. Delibašić, Facilitating data preprocessing by a generic framework: a proposal for clustering. *Artif. Intell. Rev.* **45**(3), 271–297 (2016)
4. D. Li, H. Gu, L. Zhang, A fuzzy c-means clustering algorithm based on nearest-neighbor intervals for incomplete data. *Expert Syst. Appl.* **37**(10), 6942–6947 (2010)
5. R.J. Hathaway, J.C. Bezdek, Fuzzy c-means clustering of incomplete data. *IEEE Trans. Syst. Man, Cybern. Part B (Cybern.)* **31**(5), 735–744 (2001)
6. J. Li, S. Song, Y. Zhang, Z. Zhou, Robust k-median and k-means clustering algorithms for incomplete data, in *Mathematical Problems in Engineering, 2016* (2016)
7. L. Zhang, W. Lu, X. Liu, W. Pedrycz, C. Zhong, Fuzzy c-means clustering of incomplete data based on probabilistic information granules of missing values. *Knowl.-Based Syst.* **99**, 51–70 (2016)
8. K. Kornelsen, P. Coulibaly, Comparison of interpolation, statistical, and data-driven methods for imputation of missing values in a distributed soil moisture dataset. *J. Hydrol. Eng.* **19**(1), 26–43 (2014)
9. M.N. Noor, A.S. Yahaya, N.A. Ramli, A.M.M. Al Bakri, *Filling Missing Data Using Interpolation Methods: Study on the Effect of Fitting Distribution*, vol. 594 (Trans Tech Publications Ltd., 2014), pp. 889–895
10. H. Toutenburg, T. Nittner, Linear regression models with incomplete categorical covariates. *Comput. Statistics* **17**(2), 215–232 (2002)
11. Y.G. Jung, M.S. Kang, J. Heo, Clustering performance comparison using K-means and expectation maximization algorithms. *Biotechnol. Biotechnol. Equip.* **28**(sup1), S44–S48 (2014)
12. D.Q. Zhang, S.C. Chen, Kernel-based fuzzy and possibilistic c-means clustering, in *Proceedings of the International Conference Artificial Neural Network* (vol. 122, June, 2003), pp. 122–125
13. M. Girolami, Mercer kernel-based clustering in feature space. *IEEE Trans. Neural Networks* **13**(3), 780–784 (2002)
14. D.Q. Zhang, S.C. Chen, Clustering incomplete data using kernel-based fuzzy c-means algorithm. *Neural Process. Lett.* **18**(3), 155–162 (2003)
15. T. Li, L. Zhang, W. Lu, H. Hou, X. Liu, W. Pedrycz, C. Zhong, Interval kernel fuzzy C-means clustering of incomplete data. *Neurocomputing* **237**, 316–331 (2017)
16. M. Achirul Nanda, K. Boro Seminar, D. Nandika, A. Maddu, A comparison study of kernel functions in the support vector machine and its application for termite detection. *Information* **9**(1), 5 (2018)
17. M. Tushir, S. Srivastava, Exploring different kernel functions for kernel-based clustering. *Int. J. Artif. Intell. Soft Comput.* **5**(3), 177–193 (2016)
18. S. Goel, M. Tushir, Different approaches for missing data handling in fuzzy clustering: a review. *Recent Adv. Electr. Electron. Eng. (Formerly Recent Patents on Electr. Electron. Eng.)* **13**(6), 833–846 (2020)

19. C. Blake, UCI repository of machine learning databases (1998). <http://www.ics.uci.edu/~mlearn/MLRepository>
20. Z. Huang, M.K. Ng, A fuzzy k-modes algorithm for clustering categorical data. *IEEE Trans. Fuzzy Syst.* 7(4), 446–452 (1999)

Neuroimaging (Anatomical MRI)-Based Classification of Alzheimer's Diseases and Mild Cognitive Impairment Using Convolution Neural Network



Yusera Farooq Khan  and Baijnath Kaushik

Abstract *Background:* Alzheimer's disease is one of the leading causes of dementia, including memory impairment and cognitive and behavioral conditions caused by the degeneration of neurons that renders the patient unable to function independently. However, mild cognitive impairment includes impairment in some cognitive spheres, and the patient can still function independently. Different neuroimaging modalities help the radiologist to diagnosis with Alzheimer's dementia. Deep learning in the neuroimage analysis is expanding tremendously from detection to diagnosis of Alzheimer's dementia and enhancing health treatment outcomes having Alzheimer's dementia and mild cognitive impairment. *Method and Materials:* In this study, we used deep learning approach for neuroimage analysis which is a state-of-the-art method. MRI data is taken from Alzheimer's disease neuroimaging initiative (ADNI). The convolution neural network (CNN) algorithm was applied on 975 AD MR images, 612 CN MR images and 538 MCI MR images for the classification of Alzheimer's diseases (AD) and mild cognitive impairment (MCI) from control normal (CN). Principal component analysis was used for dimensionality reduction of high resolution neuroimages (MRI). *Results and Conclusion:* Study comprises classification and accuracy evaluation based on parameters like sensitivity, specificity and AUC curve. Results showed a classification accuracy of 89%. The study demonstrates the profound deep learning capacity for difficult computer-aided diagnosis.

Keywords Neuroimaging · Neurodegenerative diseases · Deep learning · Neural network · Brain atrophy · CNN

Y. F. Khan (✉) · B. Kaushik

School of CSE, Shri Mata Vashino Devi University, Katra, Jammu and Kashmir, India

B. Kaushik

e-mail: baijnath.kaushik@smvdu.ac.in

1 Introduction

Neurodegenerative diseases are increasing as a global epidemic as they occur with age and keep growing up exponentially. Neurodegeneration is the slow and progressive death of neurons which ultimately leads to the atrophy of the brain, which manifests itself as the loss of cytoplasmic proteins, cell death and ultimately loss of brain tissue [1]. Neurons of different brain regions are involved with different functions of the human body. All these diseases have two processes, one of which is the loss of synapse over time and the other one is an accumulation of misfolded proteins [2] as shown in Fig. 1. Proteins are the building blocks of these neurons that are manufactured by the information stored in the genome of a cell. Proteins are responsible for the proper functioning of every neuron in brain. Once any protein gets misfolded, it affects neuron functionality [3]. These misfolded proteins also known as brain killers are the factors behind the neurodegeneration leading to many diseases called neurodegenerative diseases. In cells, proteins are formed in the cell cytoplasm following two-step process called translation and transcription. DNA contains the information for protein synthesis. In the transcription process [4], the codes containing the information from a cell genome that is DNA are carried out to cell cytoplasm, and there it translates the code to ribosomes to make a polypeptide chain of amino acids [5].

These polypeptide chains of amino acids then take different structures to perform the specific function. When these proteins get misfolded, they become unable to perform their specific function. Misfolded proteins then start getting accumulated leading to a slow process of neuron degeneration [6]. There are two mechanisms: One is neuronal degeneration losing synaptic connections [7], and the accumulation

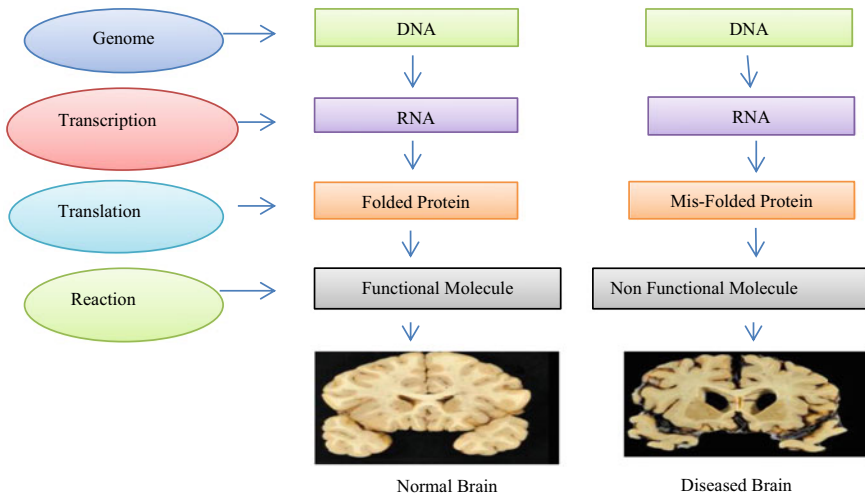


Fig. 1 Alzheimer’s patient brain atrophy caused by the misfolded proteins

of misfolded proteins is the irreversible cause of brain atrophy [8]. Neuronal degeneration leads to the death of neurons and ultimately brain atrophy which means the brain gets shrunk that is clearly depicted in the figure given above. A normal brain is differentiated from a diseased brain by the volume shrinkage. People lose 60% of synopsis before they identify the symptom of memory loss [9].

Different tools used for the identification and diagnosis of ND are clinical tests, questionnaires, biomarkers and neuroimaging [10]. Neuroimaging techniques allow visualization of the brain processes going on to better understanding the molecular activities in the brain [11]. A medical image in radiology is distinguished by low-level properties such as texture, boundaries, vertices, contours and so on. There are different medical imaging modalities to capture neuroimages [12]. Clinicians with a detailed examination of neuroimages make diagnoses for some particular form of neurodegenerative disease.

Deep learning algorithms have revolutionized computer vision using its one of the important tool called TensorFlow. Approaches earlier to deep learning needed feature engineering explicitly, whereas in deep learning feature engineering is done implicitly [13]. One of the most specific neural networks that uses perceptron is the convolution neural network (CNN) which works in the principle of neuron to analyze images the way human vision does. In this article, we will discuss deep learning algorithms applications in neuroimage analysis [14].

2 Types of Neurodegenerative Diseases (ND)

Alzheimer's disease (AD) is one of the most prevailing of all the neurodegenerative diseases and the pathogenesis involved in the amyloid-B peptides and Tau-protein. AD brain is differentiated from a normal brain in neuroimages by the cortical atrophy and loss of areas associated with memory and language found in neuroimages [15]. People with AD are seen with problems related to language and memory [16].

Parkinson's disease (PD) involves the depletion of neurotransmitter called dopamine [17] in the substantia nigra part of the brain and characterized by tremors, rigidity, Bradykinesia and postural instability. Dopamine secretion in PD is less as compared to the normal brain which leads to symptoms of PD [18]. Table 1 given below enlists different proteins associated with different neurodegenerative diseases affecting different parts of the brain.

Amyotrophic lateral sclerosis (ALS) which is known by the name of motor neuron disease affects the neurons that control the motor aspect of the movement, cardiac function and the things to do with the motor. The motor neuron disease causes muscle weakness and subsequently leads to progressive disability [4].

Huntington's disease (HD) is an autosomal overriding disease characterized and performs movements, dystonia, psychiatric problems as well as dementia. Dilation of ventricles decreases in brain size due to loss of striatal neurons in caudate putamen of basal ganglia [24].

Table 1 Different brain tissues affected in different ND

S. No	Neurodegeneration disease	Associated protein	Affected parts/tissues
1	Alzheimer's disease	Amyloid- β , Tau	Brain/hippocampus and entorhinal cortex [19]
2	Parkinson's disease	α -Synuclein	Brain/base of the brain called substantia nigra [20]
3	Huntington's disease	Huntington	Brain/caudate nuclei and striatum [17]
4	Lewy body dementia	A Synuclein	Brain/substantia nigra [21]
5	Amyotrophic lateral sclerosis	Superoxide	Brain/motor cortex and spinal cord [22]
6	Frontotemporal dementias	Tau	Brain/frontal and temporal lobes [23]

3 Motivation

The study of neurodegenerative diseases is determined to explore the clinical, cognitive, imaging and biomarker characteristics across the entire spectrum of such diseases. All the neurodegenerative diseases have a specific pattern of neuronal degeneration or brain cell loss and the specific protein associated with them [9]. Much of the research has been focused on the causes and patterns but has not yielded any successful treatment and specific therapies to these disorders to date, since neurodegeneration begins at an early age which is slow deterioration of brain functions like memory, movements, language and other functional decline resulting in brain atrophy which is irreversible [9].

The death rate of patients having neurodegenerative diseases for last 5 years worldwide is shown in the graph below according to the data available on Google trends [25]. We have also explored the comparison between the occurrence rates of dementia and other brain diseases and plotted a graph to show the comparison in Fig. 2. It is depicted in the figure that the volume of patients having dementia is more than patients having other brain diseases.

So, our study is to explore the computer-aided diagnostic approach for AD using deep learning computer vision techniques. DL algorithms are implemented for the analysis of MR imaging for early characterization of pathological changes in the brain [26] so that early detection and prediction of neurodegenerative diseases can be done for better health care [27].

4 Deep Learning Applications

DL is a subset of machine learning in artificial intelligence that has got tremendous potential in medical imaging data analysis, medical diagnostic and healthcare

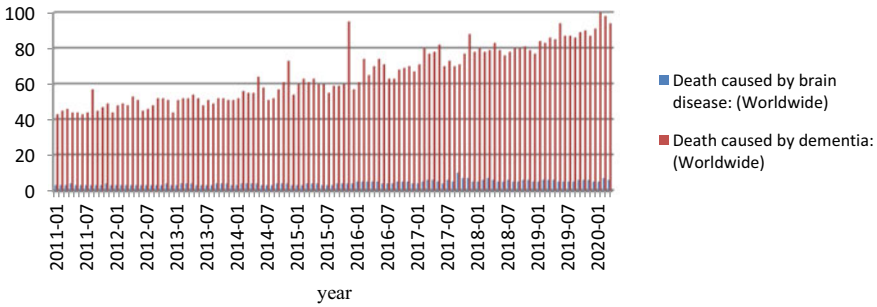
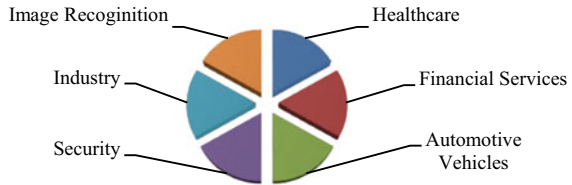


Fig. 2 Comparison between dementia patients and other brain diseases (X-axis represents number of deaths per year)

Fig. 3 Representation of various applications of deep learning in computer vision



industry. Deep learning algorithms efficiency is the reason to bring revolution in clinical practice [28]. Assortments of applications of deep learning are shown in the pie chart in Fig. 3.

The impact of ML and DL is dramatically transforming our lives across many spheres. ML and DL make computers to make decisions about our health by analyzing a large amount of data, offer clinical tools to diagnose diseases and identify best treatment options and predict better outcomes [25]. Deep learning and machine learning models play their role in the healthcare industry by assisting clinicians to make more accurate and better diagnoses which is also known as computer-aided diagnosis (CAD) [29].

Convolution neural network (CNN or ConvNet) is a category of deep neural networks (DNN) most frequently applied to computer vision applications [30]. Neural networks work similar to a neuron of brain called as perceptron [31]. Hence, an artificial neuron is a building block of deep learning architecture. Multilevel perceptron is applied in computer vision which is done by implementing CNN [32, 33]. CNN is an intricate network of interrelated processes organized in layers. Each layer in the CNN can identify elevated, more abstract features. CNN looks for these building blocks as convolution which is a mathematical operation. Images are made up of pixel that is the building blocks of the images. When CNN identifies those features, it uses something called a filter [31, 34]. In the first layer, the filters used by a CNN are small pixel squares that correspond to such features as texture, edges or contrast between two colors. Neuroimaging is a powerful technique used for the detailed examination of brains with neurodegenerative diseases [35]. Imaging findings and

clinical information are examined, and once this process is completed, the diagnosis is more thoroughly defined [36]. Impressive medical imaging breakthroughs have been accomplished, but with significant difficulties due to variable image quality and restrictions on availability. DL algorithms work fantastic in analyzing these neuroimages and extracting features [14] like low level, middle level and high level that contains important information by learning the hidden patterns, hence making accurate predictions and decision in the diagnosis process.

5 Material and Method

In this article, the CNN classification model is applied for three-class classification for Alzheimer's diseases. Here we train our CNN model to classify Alzheimer's diseases (AD), mild cognitive impairment (MCI) and normal control (NC).

Martial

Alzheimer's disease neuroimaging initiative (ADNI): MRI datasets are downloaded from the ADNI [37] based on the neuroimaging modality and three stages (CN, MCI and AD). ADNI is the largest program aimed at understanding the progression from normal aging to mild memory decline to Alzheimer's disease (Fig. 4).

It involves more than a thousand subjects that include people with normal healthy people, patients having mild cognitive impairment and patients suffering from dementia [38]. According to ADNI, there are 1196 female subjects and 1328 male subjects having different stages of cognitive impairments shown in the pie chart above [29].

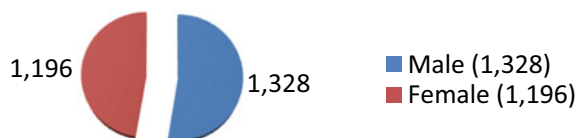
ADNI project provides clinical, neuroimaging like MRI and PET scans, biomarkers and genomic data related to the subjects involved with this project to carry on research on Alzheimer's disease.

Methodology

CNN classification model is trained on three classes of disease which are normal control (CN), Alzheimer's diseases (AD) and mild cognitive impairment (MCI). For dimensionality reduction, principal component analysis (PCA) was used in this study. Axial view of MRI dataset is taken to train our model. The proposed framework is illustrated in Fig. 5.

The input to the convolution layer of CNN is MRI axial view data as shown in Table 2 for all the three classes labeled: Alzheimer's disease (AD) as class 0, control

Fig. 4 Female and male subjects in ADNI project



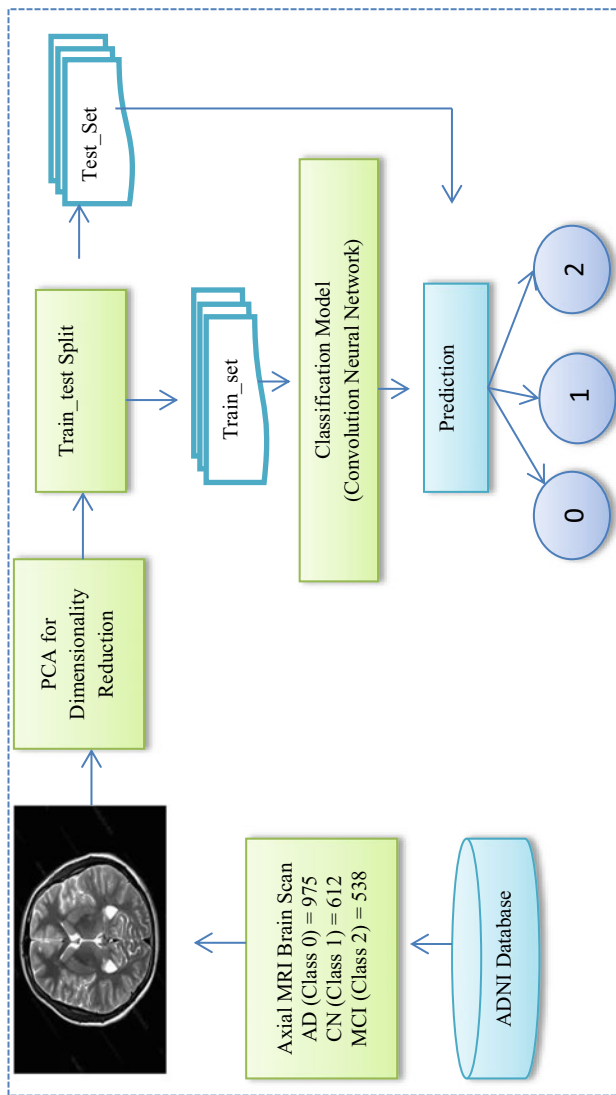


Fig. 5 Framework of methodology used in three-class (AD, CN, and MCI) classification of Alzheimer's disease

Table 2 Division of dataset into classes

	AD (class 0)	CN (class 1)	MCI (class 2)
Axial-View SLICES	975	612	538
AGE-RANGE	55–91	60–90	55–88

normal (CN) as class 1 and mild cognitive impairment (MCI) as class 2. The model is trained by splitting the dataset into a test set and train set. CNN model is trained on the train_set and latter will be tested and validated on test_set.

Conv2D(32, (3, 3)	target_size = (256, 256)
input_shape = (256, 256, 3)	batch_size = 32
Activation = relu	class_mode = 'categorical'
MaxPooling2D(pool_size = (2, 2))	rescale = 1./255
Dense(units = 128, activation = 'relu')	steps_per_epoch = 50,
Dense(units = 3, activation = 'softmax')	epochs = 20
optimizer = 'rmsprop',	validation_data = test_set,
loss = 'categorical_crossentropy',	validation_steps = 20
metrics = ['accuracy']	

6 Result and Discussion

The study contains 2125 anatomical MRI axial brain scans which are further categorized into classes AD, CN and MCI. Model has three layers including convolution layer, max pooling layer and flattening. Training hyperparameters includes filter of size (3×3) and dimension (64×64) with kernel size (3×3) . The activation function used is ReLU. Batch size was set to 32 and epochs = 20 where steps_per_epoch = 50. The framework evaluates the classification of AD and MCI using CNN. CNN has shown better accuracy as compared to the classical machine learning models.

CNN achieved training accuracy of **91%** and testing accuracy of **89%**. However, performance of the model for the AD and MCI classification can be improved by training the model with much larger datasets. The results obtained in this study are shown in Table 3. This approach can also be extended to predict various prodromal stages of AD with multiple populations of various age zones. This profound learning capability enables researchers to perform classification with improved computer-aided diagnosis.

Table 3 Results evaluation based on performance matrices

Classification model	Total MR images axial scan	Training accuracy %	Validation accuracy %
CNN	2125	91	89
Performance metrics	AD	CN	MCI
Sensitivity %	81.52	78.36	80.32
Specificity %	79.92	83.02	79.99

7 Conclusion

Deep learning is contributing to the healthcare industry by analyzing electronic healthcare records. This way, it supports clinicians to find the root cause for the undiagnosed diseases by analyzing the genomic data available in undiagnosed disease networks. Presented work shows the use of CNN in neuroimage analysis to predict Alzheimer's diseases as a result of neurodegeneration. Deep learning algorithms efficiency is the reason to bring revolution in clinical practice. Given a vast amount of neuroimaging data, with DL algorithms, it is possible to extract key features, learn patterns among the images and learn a relationship between input and output, which is highly difficult to analyze manually because of mathematical and various nonlinear equations behind it. In the future, we will work on other more datasets with sagittal and coronal views of neuroimages for predicting AD with multiple prodromal stages of cognitive decline by implementing hybrid deep learning algorithms for better performance accuracy.

References

1. A.M. Gorman, Neuronal cell death in neurodegenerative diseases : recurring themes around protein handling. **12**, 2263–2280 (2008). <https://doi.org/10.1111/j.1582-4934.2008.00402.x>
2. L. Lillemark, L. Sørensen, A. Pai, E.B. Dam, M. Nielsen, Brain region's relative proximity as marker for Alzheimer's disease based on structural MRI. *BMC Med. Imaging*. **14**, 1–12 (2014). <https://doi.org/10.1186/1471-2342-14-21>
3. S.K. Maji, A. Anoop, P.K. Singh, R.S. Jacob, CSF biomarkers for Alzheimer's disease diagnosis. *Int. J. Alzheimers. Dis.* (2010). <https://doi.org/10.4061/2010/606802>
4. M. Agrawal, A. Biswas, Molecular diagnostics of neurodegenerative disorders. *Front. Mol. Biosci.* **2**, 1–10 (2015). <https://doi.org/10.3389/fmolb.2015.00054>
5. C.M. Dobson, Protein folding and misfolding. **426** (2003)
6. J.W. Langston, The MPTP story. **7**, 11–19 (2017). <https://doi.org/10.3233/JPD-179006>
7. W. Noble, D.P. Hanger, C.C.J. Miller, S. Lovestone, The importance of tau phosphorylation for neurodegenerative diseases, *Front. Neurol.* **4**, 1–11 (2013). <https://doi.org/10.3389/fneur.2013.00083>
8. M. Dyrba, M. Grothe, T. Kirste, S.J. Teipel, Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM. *Hum. Brain Mapp.* **36**, 2118–2131 (2015). <https://doi.org/10.1002/hbm.22759>
9. K. Kwak, H.J. Yun, G. Park, J.M. Lee, Multi-modality sparse representation for Alzheimer's disease classification. *J. Alzheimer's Dis.* **65**, 807–817 (2018). <https://doi.org/10.3233/JAD-170338>
10. C. Davatzikos, P. Bhatt, L.M. Shaw, K.N. Batmanghelich, J.Q. Trojanowski, Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiol. Aging*. **32**(2322), e19-2322.e27 (2011). <https://doi.org/10.1016/j.neurobiolaging.2010.05.023>
11. P.L. Freddolino, F. Liu, M. Gruebele, K. Schulten, Ten-microsecond molecular dynamics simulation of a fast-folding WW domain. *Biophys. J.* **94**, L75–L77 (2008). <https://doi.org/10.1529/biophysj.108.131565>
12. L. Umr, HHS public access. 398–412 (2018). <https://doi.org/10.1016/j.neuroimage.2014.10.002.Machine>

13. J. Qiao, Y. Lv, C. Cao, Z. Wang, A. Li, Multivariate deep learning classification of Alzheimer's disease based on hierarchical partner matching independent component analysis. **10**, 1–12 (2018). <https://doi.org/10.3389/fnagi.2018.00417>
14. N. Tajbakhsh, J. Y. Shin, S.R. Gurudu, R.T. Hurst, C.B. Kendall, M.B. Gotway, J. Liang, Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE Trans. Med. Imaging*. **35**, 1299–1312 (2016). <https://doi.org/10.1109/TMI.2016.2535302>
15. R. Wolz, V. Julkunen, J. Koikkalainen, E. Niskanen, D.P. Zhang, D. Rueckert, H. Soininen, J. Lötjönen, Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease. *PLoS ONE* **6**, 1–9 (2011). <https://doi.org/10.1371/journal.pone.0025446>
16. S.L. Risacher, W.H. Anderson, A. Charil, P.F. Castelluccio, A.J. Saykin, A.J. Schwarz, Alzheimer disease brain atrophy subtypes are associated with cognition and rate of decline (2017)
17. J. Blesa, J.L. Lanciego, J.A. Obeso, Editorial: Parkinson's disease: cell vulnerability and disease progression. *Front. Neuroanat.* **9**, 9–11 (2015). <https://doi.org/10.3389/fnana.2015.00125>
18. S. Tenreiro, K. Eckermann, T.F. Outeiro, Protein phosphorylation in neurodegeneration: friend or foe? *Front. Mol. Neurosci.* **7**, 1–30 (2014). <https://doi.org/10.3389/fnmol.2014.00042>
19. L. Sørensen, C. Igel, N. Liv Hansen, M. Osler, M. Lauritzen, E. Rostrup, M. Nielsen, Early detection of Alzheimer's disease using MRI hippocampal texture. *Hum. Brain Mapp.* **37**, 1148–1161 (2016). <https://doi.org/10.1002/hbm.23091>
20. S.N. Aslan, B. Karahalil, Oxidative stress and Parkinson disease. *Ankara Univ. Eczac. Fak. Derg.* **43**, 94–116 (2019). <https://doi.org/10.33483/jfpau.519964>
21. S. Shimizu, D. Hirose, H. Hatanaka, N. Takenoshita, Y. Kaneko, Y. Ogawa, H. Sakurai, H. Hanyu, Role of neuroimaging as a biomarker for neurodegenerative diseases. *Front. Neurol.* **9**, 1–6 (2018). <https://doi.org/10.3389/fneur.2018.00265>
22. L.M. Sharkey, N. Safren, A.S. Pithadia, J.E. Gerson, M. Dulchavsky, S. Fischer, R. Patel, G. Lantis, N. Ashraf, J.H. Kim, A. Meliki, E.N. Minakawa, S.J. Barmada, M.I. Ivanova, H.L. Paulson, Mutant UBQLN2 promotes toxicity by modulating intrinsic self-assembly. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E10495–E10504 (2018). <https://doi.org/10.1073/pnas.1810522115>
23. N.P. Oxtoby, D.C. Alexander, Imaging plus X: multimodal models of neurodegenerative disease. *Curr. Opin. Neurol.* **30**, 371–379 (2017). <https://doi.org/10.1097/WCO.0000000000000460>
24. A. Lanzillotta, V. Porrini, A. Bellucci, M. Benarese, C. Branca, E. Parrella, P.F. Spano, M. Pizzi, NF- κ B in innate neuroprotection and age-related neurodegenerative diseases. *Front. Neurol.* **6** (2015). <https://doi.org/10.3389/fneur.2015.00098>
25. I.O. Korolev, L.L. Symonds, A.C. Bozoki, Predicting progression from mild cognitive impairment to Alzheimer's dementia using clinical, MRI, and plasma biomarkers via probabilistic pattern classification. *PLoS ONE* **11**, 1–25 (2016). <https://doi.org/10.1371/journal.pone.0138866>
26. M. Signaevsky, M. Prastawa, K. Farrell, N. Tabish, E. Baldwin, N. Han, M.A. Iida, J. Koll, C. Bryce, D. Purohit, V. Haroutunian, A.C. McKee, T.D. Stein, C.L. White, J. Walker, T.E. Richardson, R. Hanson, M.J. Donovan, C. Cordon-Cardo, J. Zeineh, G. Fernandez, J.F. Crary, Artificial intelligence in neuropathology: deep learning-based assessment of tauopathy. *Lab. Investig.*, 3–5 (2019). <https://doi.org/10.1038/s41374-019-0202-4>
27. A.S. Lundervold, A. Lundervold, An overview of deep learning in medical imaging focusing on MRI. *Z. Med. Phys.* **29**, 102–127 (2019). <https://doi.org/10.1016/j.zemedi.2018.11.002>
28. C. Salvatore, A. Cerasa, P. Battista, M.C. Gilardi, Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer's disease: a machine learning approach. **9**, 1–13 (2015). <https://doi.org/10.3389/fnins.2015.00307>
29. I.A. Illán, J.M. Górriz, J. Ramírez, F. Segovia, J.M. Jiménez-Hoyuela, S.J. Ortega Lozano, Automatic assistance to Parkinsons disease diagnosis in DaTSCAN SPECT imaging. *Med. Phys.* **39**, 5971–5980 (2012). <https://doi.org/10.1118/1.4742055>
30. A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in *Advances in Neural Information; Adv. Neural Inf. Process. Syst.*, 1097–1105 (2012). <http://arxiv.org/abs/1102.0183>

31. M.I. Razzak, S. Naz, A. Zaib, Deep learning for medical image processing: overview, challenges and future, pp. 1–30 (n.d.)
32. N. Tajbakhsh, J.Y. Shin, S.R. Gurudu, R.T. Hurst, C.B. Kendall, M.B. Gotway, J. Liang, S. Member, Convolutional neural networks for medical image analysis: full training or fine tuning?. vol. 35, pp. 1299–1312 (2016)
33. A. Fourcade, R.H. Khonsari, Deep learning in medical image analysis: a third eye for doctors. *J. Stomatol. Oral Maxillofac. Surg.* **120**, 279–288 (2019). <https://doi.org/10.1016/j.jormas.2019.06.002>
34. D.M. Dimiduk, E.A. Holm, S.R. Niezgod, Perspectives on the impact of machine learning , deep learning , and artificial intelligence on materials, processes, and structures engineering, pp. 157–172 (2018)
35. A.J. Stoessl, Neuroimaging in the early diagnosis of neurodegenerative disease, 1–6 (2012)
36. B. Ural, An improved computer based diagnosis system for early detection of abnormal lesions in the brain tissues with using magnetic resonance and computerized tomography images (2019)
37. M. Liu, D. Zhang, D. Shen, Hierarchical fusion of features and classifier decisions for Alzheimer's disease diagnosis. *Hum. Brain Mapp.* **35**, 1305–1319 (2014). <https://doi.org/10.1002/hbm.22254>
38. Google trends 2021, <https://trends.google.com/trends/explore?date=today%205-y&q=neurodegenerative%20diseases>
39. G. Prasad, S.H. Joshi, T.M. Nir, A.W. Toga, P.M. Thompson, D. Neuroimaging, I. Adni, Neurobiology of aging brain connectivity and novel network measures for Alzheimer's disease classification. *Neurobiol. Aging.* (2014). <https://doi.org/10.1016/j.neurobiolaging.2014.04.037>
40. Alzheimer's disease Neuroimaging Initiative (ADNI) 2021, <https://ida.loni.usc.edu/login.jsp?project=ADNI&page=HOME&logOut=true>

Design and Implementation of Stop Words Removal Method for Punjabi Language Using Finite Automata



Tanveer Singh Kochhar and Gulshan Goyal

Abstract With the ease in accessibility of Internet, data available online has become one of the main source of information. Large amount of data gets updated daily online. Although this data may be useful for research purposes, however, it cannot be used in its raw form. In general, unstructured data contains a lot of common irrelevant words which do not add to the semantic meaning of the document. These words are known as stop words, and removing them is an important requirement for efficient text processing as done in information retrieval systems and other natural language processing applications. A significant amount of research has been done for removing stop words in languages such as English, Chinese, Urdu, Arabic, Hindi. However, not enough work is done regarding removal of stop words in Punjabi language. Most of the available works utilize corpus-based methods for removing stop words, which tend to be time-consuming. Present paper proposes a method for removing stop words for Punjabi language using finite automata. The performance of the proposed method is compared with the classical method of stop words removal. The implementation results show that the proposed algorithm gives better results in terms of execution time.

Keywords Punjabi language · Text preprocessing · Stop word removal · Finite automata

1 Introduction

In the recent years, the use of the Internet to facilitate research has grown in popularity. Researchers have been using the freely available data on Internet to find patterns and extract meaningful knowledge for analytical purposes. Many natural language processing applications like information retrieval, text classification, text summarization, document clustering, sentiment analysis, and stemming require the data to be noise free. The removal of noise from a document is done in the data preprocessing

T. S. Kochhar (✉) · G. Goyal
Department of Computer Science and Engineering, Chandigarh College of Engineering and Technology, Chandigarh, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022
P. Verma et al. (eds.), *Advances in Data Computing, Communication and Security*,
Lecture Notes on Data Engineering and Communications Technologies 106,
https://doi.org/10.1007/978-981-16-8403-6_8

step, where the data is cleaned, reduced, and transformed into a format which is more suitable for employing different types of feature extraction methods [1]. One of the main sources of noise in a dataset is stop words.

For a word to be considered as a stop word, it must meet two requirements: First, it should be highly frequent in the document, and second, the statistical correlations with all of the classification groups have to be minimal [2]. Stop words are of very little or no relevance to the text processing tasks and account for the huge size of the corpus. Eliminating them from the text reduces the size of the corpus, saves the huge amount of space required for indexing, increases the classifier accuracy, and subsequently improves time complexity. Removal of these words can reduce the corpus size by 30–40% which makes text mining methods more efficient [3]. Therefore, removing such words is necessary before processing the data.

The most widely used natural languages in India like Hindi, Bengali, Punjabi, and other regional languages contain many sentences with stop words having no importance. These Punjabi stop words hardly contribute to the semantics of the sentence and decrease the performance of Punjabi information retrieval and text classification systems. For instance, consider the sentence ਹਰੀ ਹਾਕੀ ਖੇਡਣਾ ਪਸੰਦ ਕਰਦਾ ਹੈ, ਹਾਲਾਂਕਿ ਉਹ ਦੌੜਨ ਦਾ ਸ਼ੌਕੀਨ ਨਹੀਂ ਹੈ।(Hari likes to play hockey, however, he is not too fond of running.). Words such as ਹੈ and ਉਹ are stop words, and when eliminated from the sentence, the meaning of the sentence remains the same. Some other common stop words in Punjabi language include words such as ਅਤੇ, ਇਹ, ਹਨ, ਹੋਈ, ਨੂੰ, ਵਿਚ, ਵਿਚੋਂ.

Earlier methods of stop words removal use corpus-based pattern matching. In this method, a stop list is used which contains all the stop words for that language. The document to be processed is tokenized into words, and each word is compared with the stop word present in the stop list. If no match is found, then the word is added to the filtered document, else the word is considered to be a stop word. Though this method is simple to implement, it has a large execution time.

The present paper focuses on performing the same task of stop words removal but with the help of finite automata [4]. The proposed algorithm is much faster using finite automata-based method than the classical corpus-based method.

2 Related Work

Several studies have been done on stop words identification and reduction in various languages. Fox [5] generated a stop list for general text based on the Brown corpus. The addition and removal of words were done arbitrarily. Savoy [6] defined a stop list for French language by sorting the words based on their level of occurrence in the documents. Sinka and Corne [7] used the concept of word entropy to create two different stop lists, which performed better than the old stop lists for hard classification tasks.

For the Arabic language, Al-Shalabi et al. [8] proposed an algorithm for eliminating stop words based on a finite state machine. Alhadidi and Alwedyan [9] proposed a hybrid stop word elimination strategy based on a dictionary. El-Khair [10] created three types of stop words lists and compared its usage effect on Arabic information retrieval systems using different weightage schemes. The researchers concluded that the general stop list is recommended when dealing with different corpus, else the combined can also be used. Alajmi et al. [11] proposed a statistical approach for creating a stop list in Arabic using a corpus of 1000 words which performed better than the general list for text categorization tasks. For the Urdu language, Dar et al. [12, 13] proposed a DFA-based stop words removal algorithm similar to the work done by Al-Shalabi et al. [8], but no implementation has been done.

Zou et al. [14] suggested an automated approach for identifying stop words in Chinese language using mathematical models. Hao and Hao [15] specified a Chinese stop list based on the weighted Chi-squared statistic. Choy [16] also used the same method for generating a stop list from Twitter data. Yao and Ze-wen [17] compiled a stop list of 1289 Chinese words by combining the traditional stop list with stop words from various domains. For the Mongolian language, Zheng and Gaowa [18] proposed a word frequency-based approach for building a stop list.

Puri et al. [19] proposed an automated tool for the construction of Punjabi stop words list by combining two different stop lists. The method was tested on 10,000 Punjabi news articles taken from “Ajit” and concluded that the combined list will perform better during linear searches. Kaur and Saini [20] presented a stop list of 184 words for Punjabi Gurmukhi language identified from various online sources. Kaur and Buttar [21] implemented and compared dictionary-based method, frequency-based method, removing singletons-based method, and Punjabi words corpus-based method for stop words removal. They concluded that using these methods the size of the document can be reduced by 20–30%.

Jha et al. [22] implemented a DFA-based stop words removal algorithm. Their algorithm is able to filter out 200 documents in 1.77 s with an accuracy of 99% as compared to the classical method which does the same task in 3.4 s. With the assistance of linguistic experts, Siddiqi and Sharan [23] created a generic stop list of more than 800 stop words for Hindi language. Stop words removal algorithm and its implementation for Sanskrit language using dictionary are done by Raulji and Saini [24] using a generic stop list of 75 words. They were able to reduce an 87,000 Sanskrit words document size by approximately 13%. Behera [3] similarly implemented a FA-based stop words removal algorithm for English language which could filter 220 documents in 1.78 s with an accuracy of 99% in comparison with 3.3 s taken by the classical approach. Pimpalshende and Mahajan [25] proposed a DFA-based stop words removal algorithm for Devnagri text which can remove stop words from 20 documents in 1.98 s with 93% accuracy as opposed to 3.4 s taken by the classical method. Arora and Gandotra [26] proposed a frequency-based method of stop words identification for Dogri language. For the Bengali language, Haque et al. [27] proposed a corpus-based and a finite state automaton-based algorithm for identifying stop words in Bengali language. Their work shows an accuracy of 90% for

Table 1 Brief review of stop word removal algorithms

Sr. no.	Year	Author	Language targeted	Description
1	2016	Jha et al.	Hindi	1.77 s to remove stop words with 99% accuracy in comparison with 3.4 s taken by classical algorithm
2	2017	Pimpalshende et al.	Devnagri	1.98 s to remove stop words with 93% accuracy in comparison with 3.4 s taken by classical algorithm
3	2018	Behera	English	1.78 s to remove stop words with 99% accuracy in comparison with 3.3 s taken by classical algorithm
4	2020	Haque et al.	Bengali	90% accuracy in stop words detection

corpus-based method and an accuracy of 80% for FA-based method. Rajkumar et al. [28] proposed dictionary-based and frequency-based stop word removal methods for Tamil language using 1153 documents collected from Internet.

Table 1 summarizes stop word reduction strategies proposed by various researchers over the last five years.

From the review of literature and findings in Table 1, it is observed that not enough work has been done in stop words removal for Punjabi language, and even the work done for other languages has scope for improvement in accuracy and time taken to remove the stop words. Hence, there is a need for proposing a method of removing stop words for Punjabi language which proves to be faster.

3 Classical Corpus-based Stop Words Removal Approach

The classical corpus-based method takes in the Punjabi document as input and tokenizes this document into words. Then, each word of the document is compared with the Punjabi stop words present in the stop list. The words which do not match are added to the new document. Finally, the new document with the stop words removed is returned. Figure 1 shows the algorithm of this method.

Fig. 1 Classical corpus-based algorithm for stop words removal

```

Input: Punjabi text document containing stop words.
Output: Same document with stop words removed.

1. Start
2. Initialize:
   text= [] //To store the words present in the document
3. Perform:
   a. read the document and store each word in text
   b. for each word in text do
       for each stopword in the stoplist do
           check if word matches with stopword
           if there is no match
               add the word to the new_document
   c. write the new_document
4. Stop
    
```

4 Proposed Finite Automata-based Approach for Stop Words Removal

The present paper proposes a stop word removal approach using finite automata. A finite automata is formed from the given list of Punjabi stop words and exported to a JSON file. The finite automata constructed would have a transition diagram similar to Fig. 2 but with more number of states. Figure 2 shows the transition diagram for the finite automata accepting some of the Punjabi stop words such as ਕਰ, ਕਰਾ, ਕਰਦੇ, ਕਰਨ, ਕਰਿ, ਕਾ, ਕਿ, ਕਿਸ, ਕਿਸੇ. Table 2 shows the transition table for such an automata.

For removing stop words, the document and the JSON file are input to the program. The document is broken into words, and each word is passed to a function PROCESS_STRING() which depending upon the acceptance by the constructed finite automata, returns true or false. If false is returned, then the word is not a stop word and, therefore, is added to the new document. Figures 3 and 4 show the algorithm and flowchart for the proposed algorithm, respectively.

The PROCESS_STRING() function determines if the word is a stop word or not. For each alphabet in the word, it is checked whether a transition exists for that alphabet

Fig. 2 Transition diagram for the finite automata accepting the given Punjabi stop words

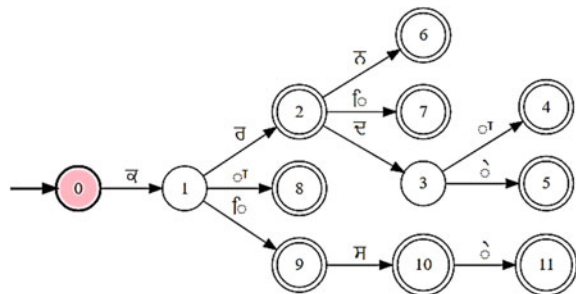


Table 2 Transition table for the finite automata accepting the given Punjabi stop words

	ਕ	ਰ	ਦ	ਨ	ਸ	ੌ	ੇ	ੌ
→0	1							
1		2				8		9
*2			3	6				7
3						4	5	
*4								
*5								
*6								
*7								
*8								
*9					10			
*10							11	
*11								

Fig. 3 Finite automata-based proposed algorithm for stop words removal**Input:** Punjabi text document containing stop words.**Output:** Same document with stop words removed.

1. Start
2. Initialize:
 - text= [] //To store the words present in the document
 - filtered_text= [] //To store the filtered words
3. Perform:
 - a. read the document, tokenize into words and store each word in text
 - b. for each word in text do
 - if PROCESS_STRING(word) == false // if function returns false
 - filtered_text.append(word)
 - c. for each word in filtered_text
 - add word to the new_document
 - d. write the new_document
4. Stop

and whether the current state has any transitions. If it does, then the current state is updated to the next state, else the function returns false. After successfully traversing the word, if the current state belongs to any of the final states, true is returned, else false is returned. The algorithm for the PROCESS_STRING() function is shown in Fig. 5. The flowchart for the same is shown in Fig. 6.

5 Implementation and Result Analysis

In the context of present paper, the classical corpus-based and the proposed finite automata-based stop word removal algorithms are implemented as per the specifications described in Table 3. For the corpus-based method, a stop list containing 133 Punjabi stop words have been used [29]. The algorithms have been tested on randomly selected news article taken from “BBC News Punjabi” [30] and “punjablibrary.com” [31]. Each test case is executed five times, and their average execution

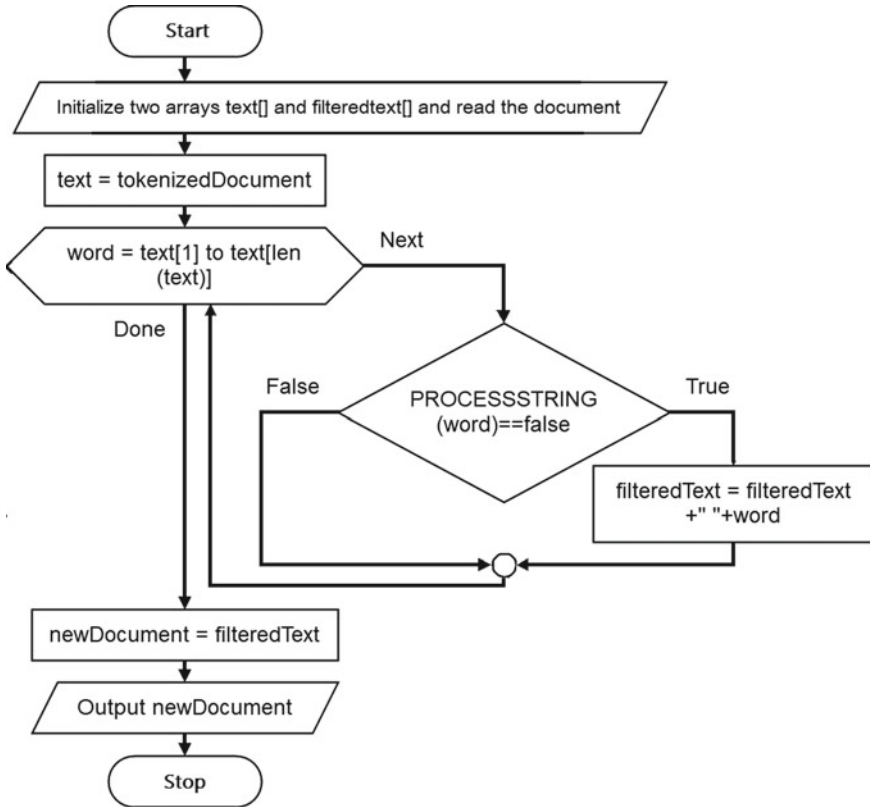


Fig. 4 Flowchart for proposed finite automata-based algorithm for stop words removal

Fig. 5 Algorithm for STOP_PROCESS function

```

Input: Punjabi word and finite automata.
Output: True or False.
1. Start
2. Initialize:
   current_state= initial state
3. Perform:
   a. for each alphabet in word do
      if transition on the current_state for the alphabet is not defined
         return false
         STOP
      else
         update current_state to the next state on the transition on the alphabet
   b. if current state in accepting state then
      return true
      STOP
   c. else
      return false
4. Stop
  
```

time is noted. The implementation results are summarized in Table 4. The performance comparison of time taken by both the algorithms is shown in Fig. 7. It is observed that the time taken by the proposed finite automata-based approach for stop word removal is comparatively less than the classical corpus-based method for stop words removal.



Fig. 6 Flowchart for the PROCESS_STRING function

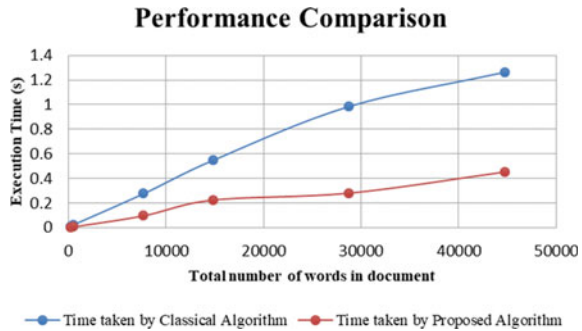
Table 3 Specifications of the simulation environment

Processor	2.27 GHz Intel Core i5 CPU
RAM	4.00 GB
Programming language	Python
No. of stop words in stop list	133

Table 4 Time taken by classical and proposed algorithm for filtering the documents

Total words in original document	Total words after removal of stop words	Time taken by classical algorithm(s)	Time taken by proposed algorithm(s)
199	137	0.0102	0.0052
511	354	0.0245	0.0091
7669	5059	0.2766	0.0984
14,814	9704	0.5488	0.2260
28,739	21,415	0.9867	0.2821
44,705	29,529	1.265	0.4531

Fig. 7 Performance comparison of the classical and proposed algorithm



6 Conclusion and Future Scope

Stop words are irrelevant words contained in any document that do not add to semantic meaning of the document. Removal of stop words is an important process in information retrieval systems and other natural language processing applications. In the present paper, a finite automata-based Punjabi stop word detection method is designed and implemented. The results are tested on randomly selected news article taken from “BBC News Punjabi” and “punjablibrary.com.” The results are compared with the classical corpus-based method constructed using a stop list containing 133 Punjabi stop words. From the results, it is concluded that removing stop words from Punjabi documents is much faster using finite automata-based method than the classical corpus-based method, and the time taken to remove stop words using the proposed algorithm is approximately four times less than the classical algorithm. In the future, the results can be tested on extended dataset. Further, performance aspects can be improved so as to make the stop word removal method more effective and efficient.

References

1. R. Feldman, J. Sanger, Categorization, in *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data* (Cambridge University Press, New York, 2016), p. 68
2. R.B. Myerson, Fundamentals of social choice theory. *QJPS*. **8**(3), 305–337 (2013). <https://doi.org/10.1561/100.00013006>
3. S. Behera, Implementation of a finite state automaton to recognize and remove stop words in English text on its retrieval, in *2018 2nd ICOEI* (IEEE, 2018). <https://doi.org/10.1109/icoei.2018.8553828>
4. J. Martin, Finite automata and the languages they accept, in *Introduction to Languages and the Theory of Computation* (McGraw-Hill, New York, 2011), p. 45
5. C. Fox, A stop list for general text. *SIGIR Forum*. **24**(1–2), 19–21 (1989). <https://doi.org/10.1145/378881.378888>
6. J. Savoy, A stemming procedure and stop word list for general French corpora. *J. Am. Soc. Inf. Sci.* **50**(10), 944–952 (1999). [https://doi.org/10.1002/\(sici\)1097-4571\(1999\)50:10%3c944::aid-asi9%3e3.0.co;2-q](https://doi.org/10.1002/(sici)1097-4571(1999)50:10%3c944::aid-asi9%3e3.0.co;2-q)

7. M.P. Sinka, D.W. Corne, Towards modernised and Web-specific stoplists for web document analysis, in *Proceedings IEEE/WIC International Conference on Web Intelligence* (2003). <https://doi.org/10.1109/wi.2003.1241221>
8. R. Al-Shalabi et al., Stop-word removal algorithm for Arabic language, in *Proceedings 2004 ICICT: From Theory to Applications* (IEEE, 2004). <https://doi.org/10.1109/iccta.2004.1307875>
9. B. Alhadidi, M. Alwedyan, Hybrid stop-word removal technique for Arabic language. Egypt. Comput. Sci. J. **30**, 35–38 (2008)
10. I.A. El-Khair, Effects of stop words elimination for Arabic information retrieval: a comparative study. IJCIS. **4**, 119–133 (2006)
11. A. Alajmi, E.M. Saad, R.R. Darwish, Article: toward an ARABIC stop-words list generation. Int. J. Comput. Appl. **46**(8), 8–13 (2012)
12. K.S. Dar et al., An efficient stop word elimination algorithm for Urdu language, in *2017 14th ECTI-CON* (IEEE, 2017). <https://doi.org/10.1109/ecticon.2017.8096386>
13. S. Kamran et al., Stop words elimination in Urdu language using finite state automaton. Int. J. Asian Lang. Process. **27**, 21–32 (2017)
14. F. Zou et al., Automatic construction of Chinese stop word list, in *Proceedings of the 5th WSEAS ICACS, Hangzhou, China* (2006), pp. 1010–1015
15. L. Hao, L. Hao, Automatic identification of stop words in Chinese text classification, in *2008 ICCSSE* (IEEE, 2008). <https://doi.org/10.1109/csse.2008.829>
16. M. Choy, Effective listings of function stop words for Twitter. IJACSA. **3**, 6 (2012). <https://doi.org/10.14569/ijacsa.2012.030602>
17. Z. Yao, C. Ze-wen, Research on the construction and filter method of stop-word list in text preprocessing, in *2011 4th ICICTA* (IEEE, 2011). <https://doi.org/10.1109/icicta.2011.64>
18. G. Zheng, G. Gaowa, The selection of Mongolian stop words, in *2010 IEEE ICICIS* (IEEE, 2010). <https://doi.org/10.1109/icicisys.2010.5658841>
19. R. Puri, R.P.S. Bedi, V. Goyal, Automated stopwords identification in Punjabi documents. IJES. **8**(June) (2013)
20. J. Kaur, J.R. Saini, Punjabi stop words, in *Proceedings of the ACM Symposium on Women in Research 2016—WIR'16* (ACM Press, 2016). <https://doi.org/10.1145/2909067.2909073>
21. J. Kaur, Stopwords removal and its algorithms based on different methods. IJARCS. **9**(5), 81–88 (2018). <https://doi.org/10.26483/ijarcs.v9i5.6301>
22. V. Jha, et al., HSRA: Hindi stopword removal algorithm, in *2016 International Conference on Microelectronics, Computing and Communications (MicroCom)* (IEEE, 2016). <https://doi.org/10.1109/microcom.2016.7522593>
23. S. Siddiqi, A. Sharan, Construction of a generic stopwords list for Hindi language without corpus statistics. IJACR. **8**(34), 35–40 (2018). <https://doi.org/10.19101/ijacr.2017.733030>
24. J.K. Raulji, J.R. Saini, Stop-word removal algorithm and its implementation for Sanskrit language. Int. J. Comput. Appl. **150**(2), 15–17 (2016). <https://doi.org/10.5120/ijca2016911462>
25. A. Pimpalshende, A.R. Mahajan, Test model for stop word removal of Devnagari text documents based on finite automata, in *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)* (IEEE, 2017). <https://doi.org/10.1109/icpcsi.2017.8391797>
26. B. Arora, S. Gandotra, Automated stop-word list generation for Dogri corpus, in *IJAST*, vol. 28 (2019), pp. 884–889
27. R.U. Haque et al., Bengali stop word and phrase detection mechanism. Arab. J. Sci. Eng. **45**(4), 3355–3368 (2020). <https://doi.org/10.1007/s13369-020-04388-8>
28. N. Rajkumar, et al., Tamil stop word removal based on term frequency, in *Advances in Intelligent Systems and Computing* (Springer Singapore, 2020), pp. 21–30. https://doi.org/10.1007/978-981-15-1097-7_3
29. K.P. Johnson, (2020). <https://github.com/cltk/cltk/blob/master/src/cltk/stops/pan.py>. Accessed 5 May 2021
30. BBC.com (2021). <https://www.bbc.com/punjabi>. Accessed 21 May 2021
31. PunjabiLibrary.com (2021). <https://punjabilibrary.com/news/>. Accessed 21 May 2021

Meticulous Presaging Arrhythmia Fibrillation for Heart Disease Classification Using Oversampling Method for Multiple Classifiers Based on Machine Learning



Ritu Aggarwal and Prateek Thakral

Abstract An arrhythmia is a problem with the rate or rhythm of the heartbeat. Variation in the heartbeat is measured by the time interval between the consecutive heartbeats known as HRV. If the arrhythmia is too slow or too fast, then it is identified as an irregular rhythm. This condition sometimes originates the tachycardia. The early detection and diagnoses of this disease make it consistent and effective to choose drugs related to arrhythmia. In this paper, the oversampling methods are used for cardiac arrhythmia arterial fibrillation such as SMOTE for nominal, random minority oversampling with replacement, adaptive synthetic sampling approach for imbalanced learning oversampling. The five machine learning classifiers are used that includes NB, SVM, DT, KNN, and LR. These oversampling methods are used to set the imbalanced data. Electrocardiogram (ECG) is one of the best ways to identify the heart rate, stroke, and heart disease by their electrical signals (electrodes, leads). The proposed methodology improves the performance in terms of best results that shows average accuracy for 87.04% achieved using the random minority oversampling with replacement. By using SMOTE method for nominal data, the accuracy achieved is 93.45%. Accuracy achieved using ADASYN method is 94.1% for multiple classifiers.

Keywords Arrhythmia · Fibrillation · HRV · ECG classification · Random forest · Logistic regression · ADASYN

1 Introduction

Machine learning (ML) is the branch of artificial intelligence that acts as being programmed. Due to the rapid growth of technology in the medical field, machine learning works as a domain of challenge. With the help of a machines, learning deploys a large amount of data that deals with a huge amount of patient history

R. Aggarwal (✉) · P. Thakral
Maharishi Markandeshwar Engineering College, Mullana, Ambala, Haryana, India

P. Thakral
Jaypee University of Information Technology, Solan, Himachal Pradesh, India

data [1–3]. Arrhythmia is occurred due to a disorder in the heart rhythm (electrical pulses) that is called as heart stroke, heart disease, or SCD [4, 5]. If SCD is an early diagnosis, it is used to choose and prescribe heart-related arrhythmia drugs which reduce the deaths in individuals [1].

An ECG arrhythmia is measured by the electrical impulse of the heart and could be analyzed and measured by the alterations and heart rhythm disorders in the ECG waveform that are evidence of heart-related problems [6, 7]. ECG is based on non-invasive arrhythmia that is detected by 12 lead, 10 lead ECG signals [8]. These electrodes are placed according to the position of the chest of patients. Ten electrodes are placed on different parts of the body surface like six in chest and four in limbs [3, 9]. These electrodes diagnose effective treatment for cardiovascular disease. In the advancement of technology, a holter machine could be used to detect heart disease [10]. The preprocessing of ECG signals is not an easy task [2, 11]. With help of machine learning algorithms, cardiovascular disease is detected that classified and deployed different types of diseases [12–14]. In this proposed work, arrhythmia disease classification is detected at early stages by using the MIT-BIH PhysioNet toolkit dataset. In this current study, information regarding heartbeats (normal and abnormal) is easily obtained and implemented using Python Jupyter notebook.

The rest of the paper is organized as follows: Sect. 2 discusses the literature review, Sect. 3 consists of the proposed system and materials, Sect. 4 contains experiment results and discussion, whereas the conclusion is done in Sect. 5.

2 Related Work

Avanzato and Beritelli [8] proposed an automated heart disease recognition technique using ECG signals based on the CNN model. The overall structure/architecture of CNN has been discussed in detail. Dataset used is divided into three classes: “Normal Class,” “Premature Ventricular Contraction Class,” and “Atrial Premature Beat class.” PhysioNet dataset is used for performance evaluation. With the help of results obtained from the confusion matrix, the performance of the proposed methodology is evaluated by applying various statistical classification functions such as sensitivity, specificity, and test accuracy. The proposed technique had high accuracy and had low complexity of implementation. Wang et al. [10] proposed a new technique for the detection of atrial fibrillation using two-lead ECG signals. In the proposed methodology, ANN is used along with wavelet packet transform (WPT) and correlation functions for feature extraction and classification. Dataset used is ECG signals from the MIT-BIH database. The credibility and robustness of the work were also guaranteed by using various statistical analyses and parameter tuning strategies. The results, thus, obtained are compared with various other existing algorithms such as SVM, KNN, fuzzy classifier that use the same MIT-BIH dataset and found to be much better. However, the proposed technique can detect ECG segments of 10 s maximum. Liaqat et al. [9] developed a novel framework for the accurate detection of AF using various ML and DL techniques. Six different models were built such as

SVM, CNN, MLP, LR, LSTM which were either deep learning approaches or feature-based approaches. Two standard benchmark datasets PhysioNet dataset and MIT-BIH Atrial Fibrillation dataset were used to implement the proposed framework. Various evaluation metrics such as precision, accuracy, F-measure were used to calculate the overall performance of the proposed approach. The experimental results obtained after applying the proposed framework showed that the deep learning algorithms such as CNN and LSTM gave much better output in terms of more accurate and less error-prone chances as compared to machine learning approaches such as SVM, LR. Although deep learning approaches do not require any feature engineering like ML algorithms, still real-time approach to detect AF can be developed where labeled data is not required at all.

Hannun et al. [15] proposed a method using DL approach to classify and use the 12 rhythm classes for ECG. The ECG dataset is constructed over a large range. DL approach is used to perform classification on ECG dataset, and by using 12 classes, the rhythm is calculated. This study shows the several limitations by which single lead ECG records are obtained from monitor.

3 Proposed Methods and Material

In this proposed study, the following dataset and method are used to implement cardiovascular disease.

3.1 Dataset

The performance of the dataset is evaluated by MIT-BIH AR and AHA. It is a standard dataset that is used to evaluate arrhythmia classifiers. ECG is an electrical activity which could be measured by using the electrode. The identification of variation in heart rate reflects the health of human health. The recorded ECG waveform could be detected by ECG analysis. The UCI has a standard dataset that is used in this proposed methodology. The dataset contains 452 patients and 279 features set. Out of these, 206 of which are linear valued, and the rest are nominal. It has 1–16 classes of arrhythmia. Class 1 is considered normal, and Classes 2–15 are considered as abnormalities or abnormal. These feature different attributes such as age, height, sex, gender. The ECG samples record at 360 HZ. Many missing values are present in this dataset. Each record contains the two leads of ECG (As given in Arrhythmia dataset Table 1).

3.2 Proposed Methodology

In this proposed work, arrhythmia is detected using different machine learning classifiers. This proposed model is implemented on a Python Jupyter notebook using various libraries and tools for the arrhythmia dataset. Figure 1 the graphic abstract

Table 1 Arrhythmia dataset [1]

Class	Number of instances
Normal	230
IC	52
Anterior MI	10
Inferior MI	10
Sinus tachycardia	15
Sinus bradycardia	34
VPC	5
SPC	4
LBBB	10
RBBB	45
FDAV	1
SDAV	0
TDAV	0
Left ventricular hypertrophy	6
Atrial fibrillation	8
Others	22

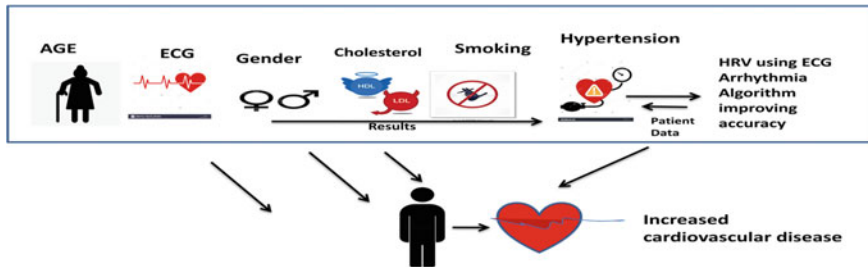


Fig. 1 Graphical representation of proposed methodology

representation of the proposed work.

Firstly, age of a person is calculated and then the heart disease is detected using ECG record. The patient history is taken by their records such as cholesterol, smoking, hypertension gender. The patient data is then transferred with the help of electrodes used on a limb and chest. The HRV test is DSE by which artificially stress is given to the heart, if the results measures find ok, then the chance of risk of heart disease is less and detected. With the help of a machine, a learning classifier could improve the accuracy applied to the arrhythmia dataset for ECG.

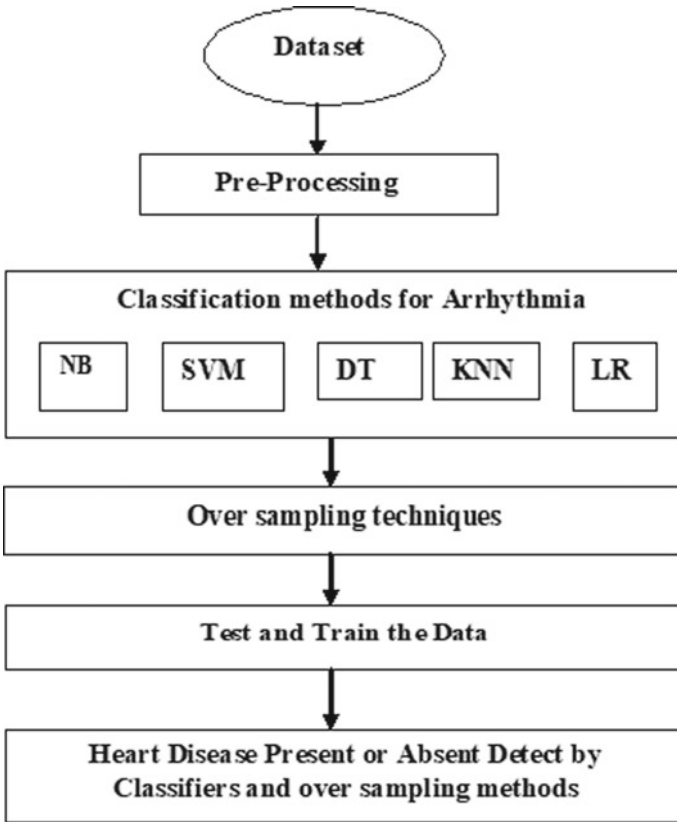


Fig. 2 Flowchart of proposed work

3.3 Arrhythmia ECG Methodologies for Proposed Work

This proposed work uses an arrhythmia arterial fibrillation dataset along with an algorithm. The first step is to take the dataset and preprocesses the data. The machine learning algorithms are used for classification which helps to train and test the data. Then, the last step using ML classifiers oversampling methods is implemented on the given dataset, to analyze the arrhythmia (heart disease) is present or absent (Fig. 2).

4 Experimental Results and Discussions

For this current work, the results were obtained by the Python Jupyter notebook using the arrhythmia dataset. Previous work of the three oversampling methods is used; random minority oversampling with replacement, SMOTE for nominal,

Table 2 Performance of various classifiers random minority oversampling with replacement

Algorithm	Features accuracy (%)	PT	Accuracy for 120 feature (%)	PT	Accuracy for 14 features (%)	PT
Logistic regression	59	0.055	82.6	0.018	81	0.103
SVM	92.89	0.622	92.8	1.366	92.63	7.128
DT	95.87	8.858	95.1	9.311	95.11	7.186
NB	97.45	0.612	97.5	0.871	98.19	2.152
DT + NB	98.38	7.315	98.5	9.94	96.13	14.866
DT + KNN	78.66	0.676	91.5	0.761	90.68	0.275

Table 3 Performance for SMOTE for nominal using oversampling technique

Algorithm	Features accuracy (%)	PT	Accuracy for 120 feature (%)	PT	Accuracy for 14 features (%)	PT
Logistic regression	84.5	0.082	86	0.054	79.04	0.67
SVM	94.4	0.519	95.9	0.394	95.85	4.556
DT	98.4	5.646	98.9	1.88	85.39	6.781
NB	99.4	8.347	99	0.53	98.19	0.156
DT + NB	99.1	0.186	98	0.124	98.355	0.165
DT + KNN	85.7	0.762	86.7	0.04	88.72	0.264

Table 4 Performance for various classifiers using adaptive synthetic sampling approach for imbalanced learning oversampling

Algorithm	Features accuracy (%)	PT	Accuracy for 120 feature (%)	PT	Accuracy for 14 features (%)	PT
Logistic regression	81.6	0.083	89.3	0.027	78.8	0.026
SVM	93.7	0.691	94.7	0.368	95.4	18.898
DT	94.7	8.962	96.4	8.215	95.4	6.467
NB	99.3	0.765	98.6	0.398	98.5	0.158
DT + NB	98.1	2.916	97.6	0.873	80.6	1.284
DT + KNN	97.2	9.732	98.2	8.461	97.5	25.99

ADASYN. Their results are shown in [Table 2](#) Performance of various classifiers random minority, [3](#) Performance for SMOTE for nominal, [4](#) Performance for various classifiers using adaptive synthetic and figures show results using graph [Figs. 3](#) Results for random minority, [4](#) Results using SMOTE, [5](#) Results using ADASYN technique. The below resulted table shows the performance of this model using various machine learning classifiers: LR, SVM, DT, NB, DT + KNN, DT + NB.

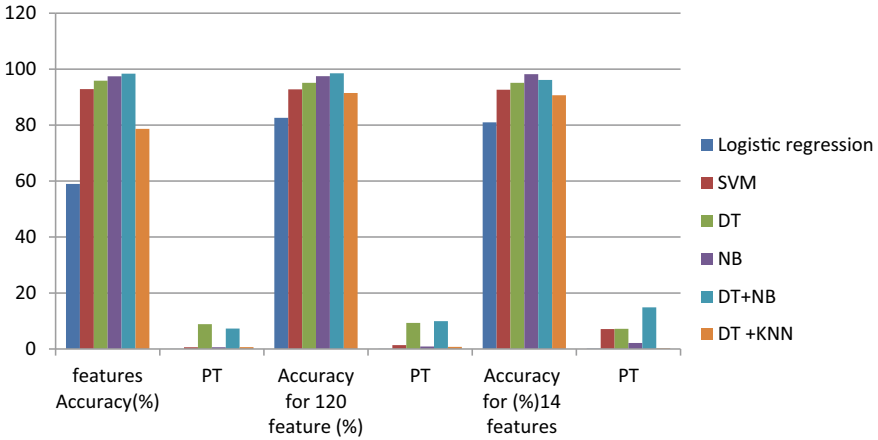


Fig. 3 Results for random minority oversampling with replacement

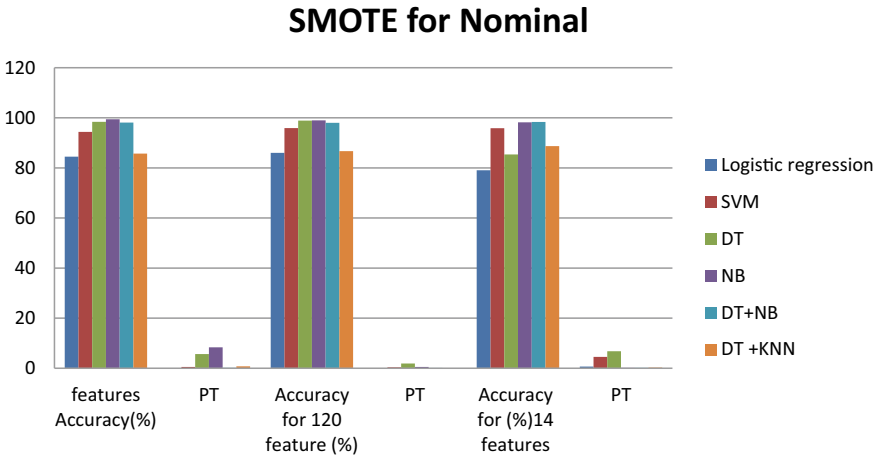


Fig. 4 Results using SMOTE for nominal method

The most effective features accuracy is 98.38 achieved using the random minority oversampling with replacement. For using method SMOTE for nominal, the accuracy achieved 99.1%, for 120 features 98%, for 14 features obtained 98.3%, and for using ADASYN method accuracy achieved 98.1%, for 120 features 97.6, for 14 features obtained 80.6%. The best method for this current in terms of accuracy rate is the SMOTE method.

Adaptive synthetic sampling approach for imbalanced learning Oversampling

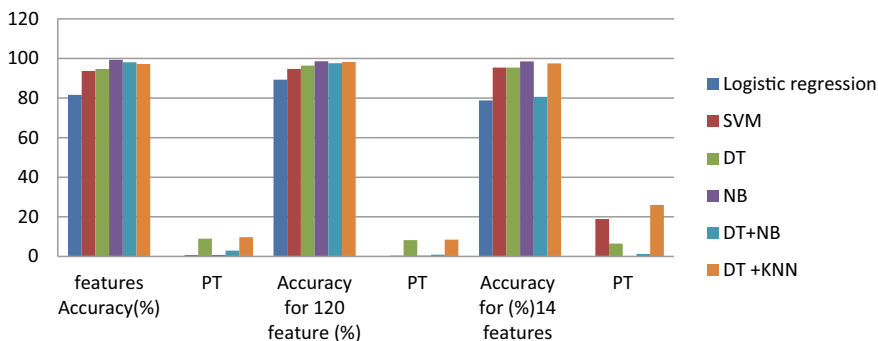


Fig. 5 Results using ADASYN technique

5 Conclusion

In this proposed study, we diagnose and classify the arrhythmia disease with the help of their dataset. This is a standard dataset taken from UCI/Kaggle. Using proposed methodology, early detection of arrhythmia disease is possible. Prediction time is calculated against each classifier and the oversampling methods to be used. The training and testing are used for arrhythmia disease implementation. The most effective features accuracy is 98.38 achieved using the random minority oversampling with replacement. For using method SMOTE for nominal, the accuracy achieved 99.1%, for 120 features 98%, for 14 features obtained 98.3%, and for using ADASYN method accuracy achieved 98.1%, for 120 features 97.6, for 14 features obtained 80.6%. The best method for this current in terms of accuracy rate is the SMOTE method for DT + NB, but if considered for DT + KNN, the ADASYN method is best. In the future, more dataset values and features could be used for implementing this work.

References

1. F. Beritelli, G. Capizzi, G. Lo Sciuto, C. Napoli, F. Scaglione, Rainfall estimation based on the intensity of the received signal in an LTE/4G mobile terminal by using a probabilistic neural network. *IEEE Access* (2018)
2. A. Mustaqeem, S.M. Anwar, M. Majid, Multiclass classification of cardiac arrhythmia using improved feature selection and SVM invariants. *Hindawi Comput. Math. Meth. Med* (2018). Article ID 731049
3. R. Guo, Projective synchronization of a class of chaotic systems by dynamic feedback control method. *Nonlinear Dyn.* **90**(1), 5364 (2017)

4. R.S. Andersen, A. Peimankar, S. Puthussertpady, A deep learning approach for real-time detection of atrial fibrillation. *Expert Syst. Appl.* **115**, 465–473 (2019)
5. S. Mondal, P. Choudhary, Detection of normal and abnormal ECG signal using ANN, in *Advances in Intelligent Systems and Computing* (2019), pp. 25–37
6. R. Avanzato, F. Beritelli, F. Di Franco, V.F. Puglisi, A convolutional neural networks approach to audio classification for rainfall estimation, in *Proceedings of the 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, Metz, France* (2019)
7. S.S. AbdElMoneem, H.H. Said, A.A. Saad, J. Phys. Conf. Ser. **1447**; in *Fourth International Conference on Advanced Technology and Applied Sciences Cairo, Egypt* (2019)
8. R. Avanzato, F. Beritelli, Automatic ECG diagnosis using convolutional neural network. *Electronics* **9**, 951 (2020). <https://doi.org/10.3390/electronics9060951>, pp.1-14
9. S. Liaqat, K. Dashtipour, Z. Adnan, A. Kamran, R. Naeem, Detection of atrial fibrillation using a machine learning approach. *Information* **11**(12), 549. <https://doi.org/10.3390/info11120549>
10. J. Wang, P. Wang, S. Wang, Automated detection of atrial fibrillation in ECG signals based on wavelet packet transform and correlation function of a random process. *Biomed. Sig. Process. Control* **55** (2020). Article ID 101662
11. Z. Xiong, ECG signal classification for the detection of cardiac arrhythmias using a convolution recurrent neural network, *Physiol. Meas.* **39** (2018). Article ID 0940069
12. A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P.C. Ivanov, Predicting paroxysmal atrial fibrillation/flutter. *PhysioNet Comput. Cardiol. Challenge*. Accessed 12 Sept 2020
13. G.D. Clifford, C. Liu, B. Moody, I. Silva, Q. Li, A. Johnson, R.G. Mark, AF classification from a short single lead ECG recording: the PhysioNet/computing in cardiology challenge (2017)
14. C. Liu, F. Liu, Y. Liu, L. Zhao, X. Zhang, C. Ma, S. Wei, J. Li, E.N. Kwee, An open-access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *J. Med. Imag. Health Inf.* **8**(7), 1368–1373 (2018)
15. A.Y. Hannun, P. Rajpurkar, M. Haghpanahi, G.H. Tison, C. Bourn, M.P. Turakhia, Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Med.* **25**(1), 65 (2019)

Unsupervised Modeling of Workloads as an Enabler for Supervised Ensemble-based Prediction of Resource Demands on a Cloud



Karthick Seshadri, C. Pavana, Korrapati Sindhu,
and Chidambaran Kollengode

Abstract Understanding the resource demands and managing the resources for future workloads is a challenging task in the cloud. Typically, applications receive dynamic and time-varying workloads. Forecasting future workload and subsequently inferring future resource requirements will aid in better capacity planning and resource utilization. In this paper, we propose an unsupervised approach to predict the number of utilization profiles required to model a workload. An ensemble-based workload forecasting is proposed to predict future workloads. For forecasting, three models, namely ARIMA, XGBoost, and LSTM, are ensembled. Both forecasting and classification are used to predict the number of resources required so that the problem of over-provisioning and under-provisioning of the resources can be addressed during capacity planning and provisioning. The error rate is decreased by 15% with ensemble model when compared with individual models for prediction.

Keywords Workload characterization · Cloud computing · Time series models · Capacity planning · Resource provisioning · Ensemble models

1 Introduction

Organizations prefer to migrate their applications to a cloud environment mainly to reduce the number of computing and storage resources required to run their applications. Virtualization of resources helps in running multiple Virtualized Operating Systems (OS) on top of a host OS to ensure a healthy utilization of the resources.

K. Seshadri (✉) · C. Pavana · K. Sindhu
Department of Computer Science and Engineering, National Institute of Technology,
Tadepalligudem, Andhra Pradesh, India
e-mail: karthick.seshadri@nitandhra.ac.in

K. Sindhu
e-mail: sindhukorrapati.sclr@nitandhra.ac.in

C. Kollengode
Data and AI Platforms, LinkedIn, Bangalore, Karnataka, India

However, the number of Virtual Machines (VMs) to be provisioned for an application is difficult to decide due to the dynamic nature of the users who interact with these applications. To cater to the dynamism in the number of VMs required, researchers attempt to develop algorithms to make a cloud “elastic.” Cloud elasticity is defined as the ability of the cloud environment to adapt to the changes in the workload through adaptive algorithms that perform resource provisioning or de-provisioning such that the resources provisioned exactly fit the application’s resource requirement at a given time instant. Therefore, a data analytic framework that characterizes the workloads through statistical models and predicts the resource requirements of the host applications on a cloud is a justified need for data centers.

In the existing research attempts [1, 2], the utilization profiles of different resources required to execute a workload are considered in modeling the workload. However, the number of unique utilization profiles needed to model the workload and the characteristics of these profiles are typically manually curated. Another demerit arguably shared by many of the existing research attempts is that the workload models do not possess sufficient representational power to encode both the short-term and long-term dependencies typically observed in the workload stream; this translates into a poor predictive ability of the model when forecasting the resource requirements into the future. This feature is mandatory for the data center administrators to carry out capacity planning and effective management of the virtualized resources in the data center.

The framework proposed in this paper addresses the above-cited demerits of the existing cloud provisioning frameworks with the following quad-folded approach:

- (i) A Gaussian framework is initially leveraged to infer a mixture model that characterizes the utilization profiles in a sequence of workloads received at a data center over a time interval.
- (ii) A window-based feature engineering is performed by treating the workloads corresponding to each utilization profile as a sequence/stream to extract features from the temporal, spatial, and spectral domains.
- (iii) To predict the most likely utilization profile of an incoming window of workloads, a tree-based classifier is proposed to be used, after a thorough evaluation of the performance of the state-of-the-art classifiers for this prediction task.
- (iv) Design and evaluation of an ensemble of time series models to predict the future workload after factoring in both short- and long-term correlations in the workload patterns.

The key contributions of this paper are the following:

- (i) The paper proposes an unsupervised approach to infer the number and components of the utilization profiles required to characterize a data center workload.
- (ii) A novel feature engineering and ensemble modeling are proposed, which work in synergy to encode both short and long-term correlations inherent in a typical workload and thereby providing a better forecast.

The rest of this paper has the following layout: Sect. 2 discusses the existing methods for workload characterization and prediction. In Sect. 3, we illustrate with appropriate reasons, the suitability of various workload characterization and prediction techniques to address the research problem. We also explain in detail the proposed method for workload prediction. In the subsequent section, we present the experimental results obtained with inferences. We conclude with future research pointers to extend this thread of research, in the last section of this paper.

2 Related Work

Various methods for cloud workload characterization and prediction have been developed using machine learning and deep learning models in recent years. For characterizing workloads, clustering approaches, namely K-means and Gaussian Mixture Model (GMM), attempt to determine the cluster to which a particular workload belongs [3] and classified into different categories [4]. In prior research attempts, such as [5] ensemble-based clustering algorithms have been analyzed for their suitability in workload characterization.

In Rahmanian et al. [6] the workload prediction problem is addressed using ensemble models. The different models ensemble are Autoregressive Integrated Moving Average (ARIMA), second moving average model (SMA), exponential moving method (EMA), autoregressive model (ARM), neural network (NN), Support Vector Machine (SVM), Markov models, and Bayesian models as the component models. Manish Kumar et al. [7] used deep learning models, namely Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), for predicting the user demands in peak travel times. Experiments are done on the Uber dataset. Song et al. [8] proposed a LSTM model for predicting the host load. The framework proposed in Radhika and Sudha Sadasivam [9] uses a recurrent neural network with LSTM to predict the future workloads. Baig et al. [10] proposed a method for accurate estimation of resource utilization, using a model for heterogeneous workloads.

3 Proposed Method

In this section, we describe the methodology adopted for workload characterization and resource requirement forecasting. The overall architecture of the proposed method is shown in Fig. 1. The proposed approach encompasses the following steps:

- (i) An azure public dataset which analyzes the CPU utilization at different time stamps on a cloud was used for our analysis, each data point in the dataset is normalized using min-max normalization.
- (ii) Gaussian-based unsupervised method is used to fit the mixture of Gaussians that generated the data, and the optimal number of mixture components

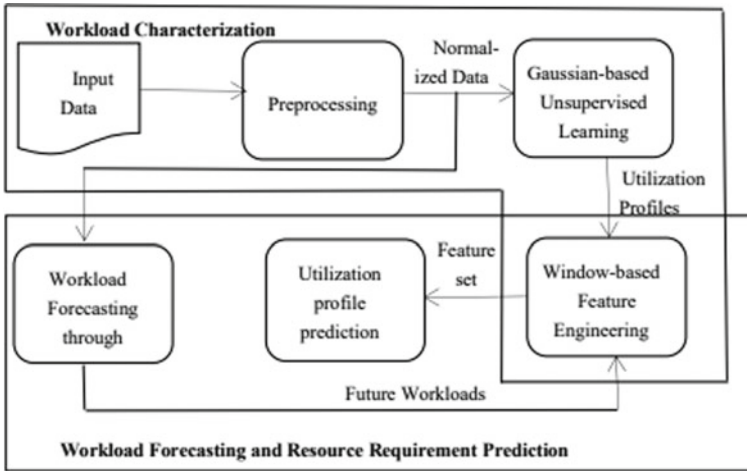


Fig. 1 The architecture of the proposed model

is decided based on the Akaike's Information Criteria (AIC) and Bayesian Information Criteria (BIC).

- (iii) Each data point in the dataset is labeled with corresponding mixture component. The mixture component describes the utilization profile of resource utilization to execute a workload.
- (iv) The data points assigned to a mixture component are treated as a time-series comprising multiple windows and a window-based feature engineering was performed to extract a feature vector.
- (v) A classifier is built using the features engineered, to model the relationship between the features and the mixture component.
- (vi) Ensemble-based forecasting is done to predict the upcoming workloads, and then, using the classifier, the resource requirement can be predicted.

3.1 Pre-processing

To perform our experiments, we have used the azure public dataset which comprises CPU usage statistics at different time stamps in a data center. Each data point in the dataset is normalized using min-max normalization.

3.2 *Gaussian-based Unsupervised Learning*

Once the data is preprocessed, the next step is to identify the number and parameters of the utilization profiles required to model the workload. Initially, the time dimension is ignored while fitting the Gaussians. Gaussian-based clustering technique is used to model the generator of the dataset as a mixture of a finite number of Gaussian distributions. Three different Gaussian mixture models have been tried, namely GMM, Gaussian HMM, and Deep GMM. Each Gaussian component in the mixture maps to the utilization profile needed to execute the workload.

All the above-cited Gaussian mixture models require the number of mixture components to be manually supplied. In this paper, AIC and BIC are used to find the number of mixture components [11]. AIC estimates the relative amount of information loss of a given model. Therefore, the model with a low AIC score will fit the data better. AIC will not penalize a complex model, but BIC penalizes complex models by assigning a high score to such models.

3.3 *Window-based Feature Engineering*

After identifying the number of Gaussians required to represent the workload, we associate a Gaussian mixture component to each data point, such that the component associated has the highest likelihood for generating the data point. Let C_i represents the i th Gaussian mixture component, and the data points that belong to C_i are treated as a time series. The data points within a time series are then segmented into windows. In our experiments, the window size is set to 20. Window size is initialized with a value equal to $(\text{number of samples})^{1/(\text{number of input workloads})}$. Subsequently, the following temporal, statistical, and spectral features are extracted from each window.

Temporal Features

Temporal features describe the long-term dynamic characteristics of workload over time like autocorrelation, centroid, mean absolute differences, maximum and minimum peaks, entropy, peak to peak distance, area under the curve, slope, and zero-crossing rate.

- (i) **Autocorrelation:** It measures the correlation between different values of the same attribute observed during different time stamps and is defined similar to correlation co-efficient.
- (ii) **Mean Absolute Difference:** It can be calculated as the arithmetic mean of the absolute value of all pair-wise differences between data points in a window.

Statistical Features

Statistical features describe the mathematical structure of data through various statistical measures like histogram, interquartile range, variance, and Estimator of Cumulative Distribution Function (ECDF).

- (i) **Histogram:** A histogram represents the frequency of a variable.
- (ii) **Interquartile range:** The interquartile range (IQR) denotes the observations falling between the 25th and 75th percentile observations.
- (iii) **Empirical Cumulative Distribution Function (ECDF):** ECDF is the distribution function associated with the empirical measure of a sample.

Spectral Features

Spectral features are frequency-based measures on the wavelets like fundamental frequency, median frequency, wavelet absolute mean, and wavelet variance.

3.4 Building a Classifier for Prediction of Utilization Profiles

After extracting the features, a dataset is created with the window identifier (w_i), features extracted ($f_{i1}, f_{i2}, \dots, f_{ik}$), and the class label (C_i), which is the identifier of the mixture component associated with the data points in the window. The structure of dataset with the features extracted is as shown in Fig. 2. A classifier model is trained with 80% of the tuples in the dataset. Different classifier models were employed, namely decision tree, random tree, and deep forest. The classifier providing better accuracy is selected as the model to be used for predicting the utilization profile of a new workload.

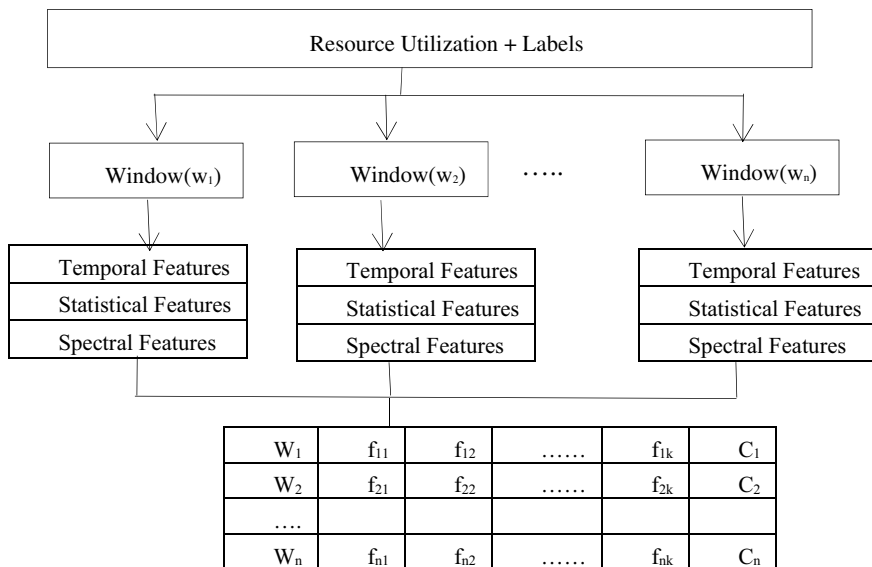


Fig. 2 Dataset structure

Workloads in a window				Predicted
t-20	t-19	t-1	t
t-19	t-18	t	t+1
..

Fig. 3 Forecasting process

3.5 Workload Forecasting

The purpose of workload forecasting is to predict the next incoming workload based on past observations. For workload forecasting, three different sequence methods, namely ARIMA model, eXtreme gradient boosting (XGBoost), and LSTM, are ensembled. To predict the workload at time step “t,” the observations at the time steps $t - 1$, $t - 2$, ..., and $t - 20$ are considered. Initially, each model in the ensemble is trained with 80% of the tuples in the dataset and the next workload is predicted.

Weights are assigned to the models in the ensemble based on their performance. Once the forecasting of a window of twenty observations is done, we can extract features by treating these observations as a window, and then, the utilization profile of a workload is predicted using the classifier. Fig.3 depicts the workload forecasting process.

4 Experimental Procedure

In this section, we describe the experiments carried out for workload characterization and resource requirement forecasting.

4.1 Experimental Setup

Experiments are carried out on a twelve core Intel (R) Xeon (R) W-2265 series processor running at 3.50 GHz with 32 GB RAM. We have used free open-source packages like pandas, keras, Statsmodels, and matplotlib available in Python for implementing the proposed method. Experiments are performed on azure public dataset [12].

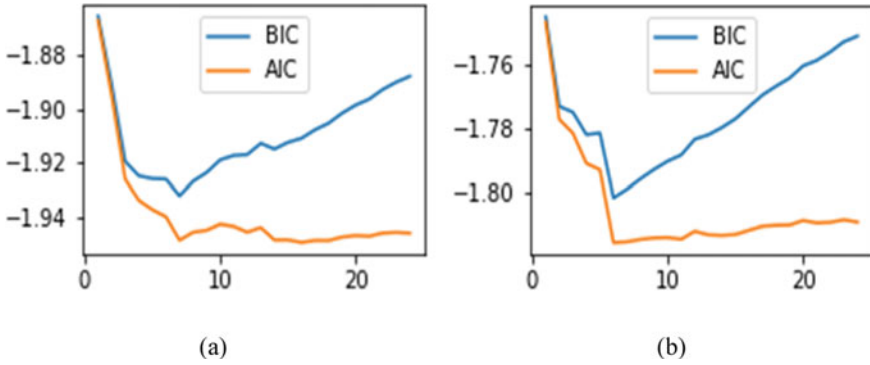


Fig. 4 Number of utilization profiles for **a** MinCPU **b** MaxCPU

4.2 Analyzing the Gaussian-based Framework

The first step in workload characterization is analyzing the mixtures of Gaussians that have generated the data. By performing generative modeling through each of the methods, namely GMM, Gaussian HMM, and deep GMM, GMM required eight components and other methods required seven components each to fit the dataset. We have generated AIC and BIC values for 25 different models where each model is characterized by a different number of Gaussian distributions.

To validate the number of mixture components obtained by using GMM, Gaussian HMM, and deep GMM, AutoGMM is leveraged. AutoGMM determines the number of components dynamically optimized through agglomerative clustering and regularization. In our analysis, we have selected seven gaussian components post validation by AutoGMM and mapped them to seven classes in the dataset. Thus, the number of classes can be selected in an unsupervised way using the AIC and BIC criteria for the attributes MinCPU and MaxCPU as shown in Fig. 4a, b, respectively. A knee could be observed at seven, hence we require seven different resource requirement profiles to model the dataset, with each profile corresponding to a class label. After deciding the number of Gaussian mixture components (number of class labels), the labeling for MinCPU and MaxCPU, respectively, is as shown in Fig. 5a, b.

4.3 Building a Classifier Model

Once the dataset is created with the features extracted from each window, we built a classifier to predict the utilization profile of a workload. The accuracies of the three classifiers, namely decision tree, random tree, and deep forest are found to be 82.9%, 87.9%, and 94.5%, respectively.

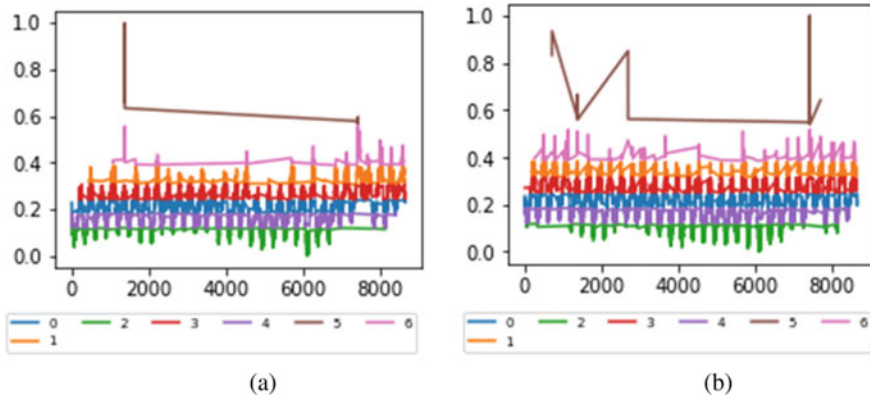


Fig. 5 Labeled workload **a** MinCPU **b** MaxCPU

It is observed that the deep forest model performed better than the other two; hence, in our experiments, deep forest classifier is leveraged for classification of workloads and to subsequently infer their utilization profiles.

4.4 Workload Forecasting

To create an ensemble model which predicts the next incoming workload, ARIMA, LSTM, and XGBoost are ensemble with random weights in the ranges $[0, 0.2]$, $[0.4, 0.6]$, and $[0.4, 0.6]$, respectively. Predictions made by XGBoost and LSTM are similar and accurate, so a relatively larger weight is assigned to XGBoost and LSTM when compared to the ARIMA model.

Figure 6a–c shows the predictions made by LSTM, XGBoost, and ARIMA models, respectively, when the resource demands imposed on data center are high. The overall ensemble model performs well for both short-term and long-term predictions. Table 1 provides the overview and merits of ARIMA, LSTM and XGBoost models. Mean Square Error (MSE) and Mean Absolute Error (MAE) of the workload forecasting models is shown in Table 2. Table 2 asserts that the predictive accuracy of the ensemble model is better than its individual components.

5 Conclusion and Future Work

In this paper, we have proposed a model which accurately predicts the resource demands in future, thereby addressing the problems of over-provisioning and under-provisioning of resources in a cloud. An unsupervised approach has been adopted to characterize the number and parameters of the utilization profiles required to model

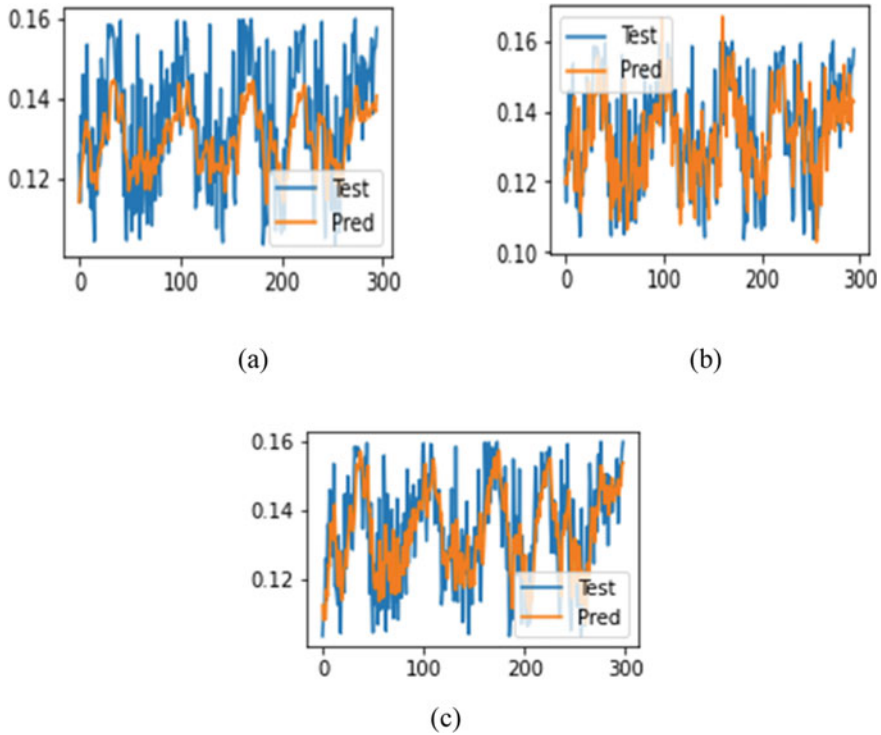


Fig. 6 a LSTM b XGBoost c ARIMA predictions for MaxCPU

Table 1 Description of workload forecasting methods

Ensemble component	Description	Advantages
ARIMA	ARIMA is a linear regression model that uses historical data to make logs that act as future predictors	Short-term workload forecasting can be done with a smaller number of parameters
XGBOOST	It performs optimized gradient boosting on an ensemble of decision trees, creating a robust model	Combines weaker predictors for making a robust predictor. After characterization, XGBOOST can extrapolate similar workloads in less time
LSTM	It is a variant of RNN without the vanishing gradient problem that can retain the state from one iteration to the next by using their own output as input for the next step	Due to the memory components of the model, it has a capacity to store large gaps in the workload trends. The fluctuations in the series can be captured across a longer time span with this component

Table 2 Comparison of workload forecasting methods

S. No.	Method	Mean squared error (MSE)			Mean absolute error (MAE)		
		MinCPU	MaxCPU	AvgCPU	MinCPU	MaxCPU	AvgCPU
1	ARIMA	0.014	0.018	0.016	0.113	0.120	0.120
2	XGBOOST	0.013	0.016	0.017	0.103	0.090	0.115
3	LSTM	0.014	0.015	0.017	0.108	0.112	0.120
4	Ensemble	0.008	0.004	0.005	0.076	0.084	0.097

a data center’s workload. A deep forest tree-based classifier is subsequently used to predict the utilization profile of an incoming workload. A future research pointer will be to leverage deep networks for automatic feature engineering rather than hand-curating the features and to evaluate its relative performance with the approach presented in this paper.

Funding This work has been sponsored by LinkedIn under a research grant for the project entitled “A Scalable Resource Requirement Prediction and Provisioning Framework for Elastic Cloud.”

References

1. C. St-Onge, S. Benmakrelouf, N. Kara et al., Generic SDE and GA-based workload modeling for cloud systems. *J. Cloud Comp.* **10**(6) (2021)
2. A. Ganapathi, Y. Chen, A. Fox, R. Katz, D. Patterson, Statistics-driven workload modeling for the Cloud, in *Proceedings 26th International Conference on Data Engineering Workshops (ICDEW)* (2010), pp. 87–92
3. E. Patel, D.S. Kushwaha, Clustering cloud workloads: K-Means vs Gaussian mixture model. *Procedia Comput. Sci.* **171**, 158–167 (2020)
4. E. Ergüner Özkoç, Clustering of time-series data, in *Data Mining–Methods, Applications and Systems* (IntechOpen, 2020), pp. 1–19
5. S. Ismaeel, A. Al-Khazraji, A. Miri, An efficient workload clustering framework for large-scale data centers, in *Proceedings 8th International Conference on Modeling Simulation Applied Optimization (ICMSAO)* (2019), pp. 1–5
6. A.A. Rahmanian, M. Ghobaei-Arani, S. Tofighy, A learning automata-based ensemble resource usage prediction algorithm for cloud computing environment. *Futur. Gener. Comput. Syst.* **79**, 54–71 (2018)
7. M. Kumar, D.K. Gupta, S. Singh, Extreme event forecasting using machine learning models, in *Advances in Communication and Computational Technology. Lecture Notes in Electrical Engineering*, vol. 668, eds. by G. Hura, A. Singh, L. Siong Hoe (Springer, Singapore, 2021). https://doi.org/10.1007/978-981-15-5341-7_115
8. B. Song, Y. Yu, Y. Zhou, Z. Wang, S. Du, Host load prediction with long short-term memory in cloud computing. *J. Super Comput.* **74**, 6554–6568 (2018)
9. E.G. Radhika, G. Sudha Sadasivam, An RNN-LSTM based flavor recommender framework in hybrid cloud, in *Proceedings 17th International Conference on Machine Learning and Applications (ICMLA)* (2018), pp. 270–277
10. S. Baig, W. Iqbal, J.L. Berral, A. Erradi, D. Carrera, Adaptive prediction models for data center resources utilization Estimation. *IEEE Trans. Netw. Serv. Manage.* **16**, 1681–1693 (2019)

11. T. Hastie, R. Tibshirani, J. Friedman, Model assessment and selection, in *The Elements of Statistical Learning* (Springer, 2009), pp. 219–257
12. Azure Public Dataset. Available at <https://github.com/Azure/AzurePublicDataset>. Accessed 10 June 2021

A Technique to Find Out Low Frequency Rare Words in Medical Cancer Text Document Classification



Falguni N. Patel, Hitesh B. Shah, and Shishir Shah

Abstract A vast amount of digital medical documents are increasing day by day, and there is need of automatic text document classification. Medical research persons, doctors, and medical community search or classify their relevant documents. The documents can be medical research papers, articles, reports, surveys, etc. In this paper, we have investigated that tradition classification method applied on medical data and removed rare low frequency words that degrade performance of classifiers. We find that rare words are important in medical domain and study existing methods to find rare words. The available methods are fixed statistical calculation-based threshold value for all dataset or sample collection. So, we proposed a method for rare word finding using dynamic threshold calculation based on term frequency as well as inverse documents frequency and medical dictionary words matching concept. We have taken two real medical text dataset and applied three text classifiers kNN, NB, and SVM. The results shown that our method finds right rare words. Considering only rare words gives same or nearer accuracy of all features in classification. It also shows that removing rare words degrades performance of classifiers in most of the cases specific in medical domain.

Keywords Medical text data · Text classification · Rare terms · Dynamic threshold · Low frequency words

F. N. Patel (✉)
GTU, Ahmedabad, Gujarat, India

H. B. Shah
Department of EC, GCET, Ahmedabad, Gujarat, India
e-mail: hiteshshah@gcet.ac.in

S. Shah
University of Houston, Houston, USA
e-mail: sshah@central.uh.edu

1 Introduction

With the rapid growth of computerized medical information as text literature, text classification become demanding process to properly step-wise manage and analyze data. The medical text data can be short sentence or large documents [1]. Text classification means categorize or give a pre-class label from predefined labels to unknown/new text sample/instance using trained model which is built from training labeled dataset [2]. Text classification applications in medical domain are biomedical sentence classification, biomedical literature classification, cancer document classification, disease classification, adverse drug events (ADE) classification, etc. [3, 4]. In this paper, we are concentrated on large medical articles or research papers for cancer articles classification.

Using supervised machine learning approach, medical document classification performed using different phases like text documents preprocessing phase which include tokenizing, case conversion, stop word removal, word stemming/lemmatization, term-document matrix conversion, etc., feature extraction/feature selection/feature reduction phase, model building and testing phase, and result evaluation phase [3, 5–8]. This traditional medical text classification process flowchart is shown by us in Fig. 1.

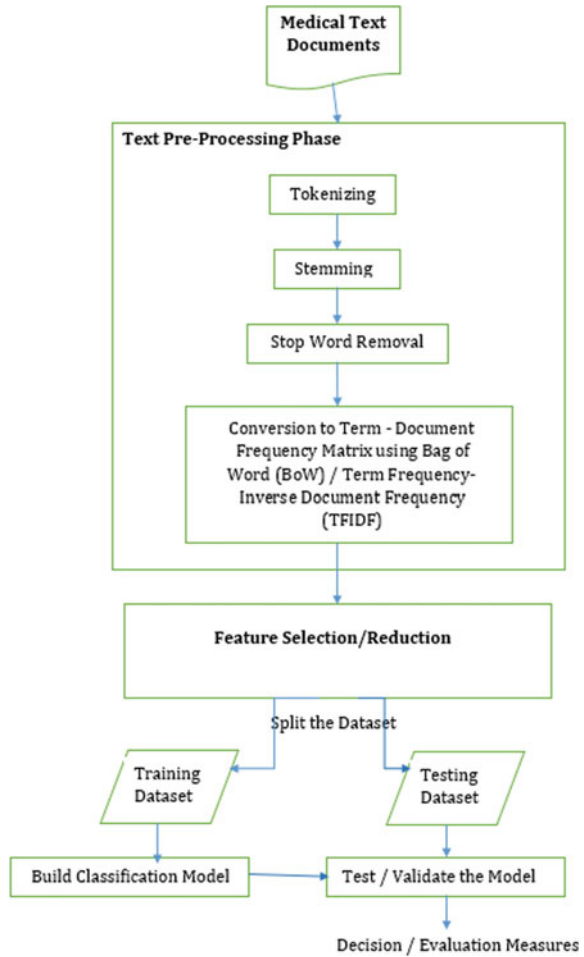
1.1 Challenges of Medical Text Data in Classification Task

Medical text data has many issues regarding to classification process, and they are as follow [6, 9]:

- (1) Medical abbreviations/Acronyms terms identification
- (2) Polysemy and synonymy words
- (3) Non-grammatical terms
- (4) Special medical terms with prefix, root and suffix terms
- (5) Large number of medical terms and Terminology
- (6) Spelling error terms
- (7) Rare terms/Low frequency terms, etc.

As above list of issues is related to medical text data, our paper is mainly concentrated on rare terms/word which are low frequency words. Generally, state of art or any traditional text classification process finds less frequent words called rare words, which are removed with static threshold. But in medical domain, if same traditional approach applied, it may decade the performance of classifiers because low frequency words/rare terms can be special medical terms or abbreviations. So, we cannot directly remove those rare terms. In paper, authors shown that rare words are important for medical text data classification [2, 10, 11]. They show that using only rare terms considered for classification and all features for classification gives

Fig. 1 Traditional text classification process



same or nearer accuracy of classification. If we remove rare terms, it degrades accuracy in most of cases. In next sections, we discussed different approaches or types of methods available for static threshold calculation in literature.

2 Background

In literature, there are number of researches performed for define static threshold which are used for different purpose like text classification, clustering, summarization, etc., purpose. Some of the authors define or calculate threshold based on statistical- and frequency-based graph methods.

In the paper [12], authors define threshold by calculating percentage using ratio. The ratio is number of matching terms to the class and number of terms present in the document. This is a static threshold. Author selects the important terms and applies similarity measure for research paper and patent article for classification. They used threshold value for assigning research paper to a specific class. Authors define a threshold using mean and standard deviation static measures which are used for selecting proper topic names for article/paragraph means topic modeling [13]. In the research paper [14], authors try to develop efficient text summarization for people who do not sufficient time to read whole book. For visually challenged people, they also have problem to read whole book also. In the preprocessing phase, they have used sentence-based total TFIDF and selected some sentences with high total value using threshold which are useful for summarization. The threshold is calculated with equation. Padma et al. [15] develop automatic text summarization using the weight of sentences. This weight is calculated using statistical and linguistic features. They define threshold using average measure of all sentences weight to select good sentences. In paper [16, 17], author first calculates feature selection score for each feature and calculates threshold iteratively. For document frequency-based feature selection method, initially calculate accuracy for low, average, and high number of features. Here, whose accuracy is higher than those bounded features are selected as interval with low, mid, and high. These steps are performed till the accuracy is met or maximum iterations are met. Authors used this threshold with different feature selection methods for binary class Chinese document classification. Roy et al. [18] define a threshold based on graph. This graph represents term frequency count versus global frequency count or threshold. In graph, they find that for low global frequency, a greater number of term frequency count and this graph line are declined as global frequency increase. Authors select a point where the graph decline. That point selected as threshold and used for finding association between genes from biomedical literature. Christian et al. [19] used percentage measure as threshold calculation for text summarization. Authors select 30% of features among all features for text summarization. In research paper [20], authors determine threshold based on clustering which use local and global distance of intra and inter cluster representative. They discussed other six threshold methods like maximum, average, fixed local distance, weighted local distance, standard deviation, and maximum deviation. Authors proposed threshold formula, $T = (1/LD) + (1/GD)$ for text clustering purpose.

3 Research Gap

As shown in literature, there is a need of proper threshold calculation method for text classification to find rare terms specific for medical text data. Another thing is as dataset changes threshold value should be changed that mean dynamic threshold determination is major requirement for classification. This threshold determination method is useful for finding important rare terms in medical data which can be further

handled for improving classifiers performance [2, 17, 21]. This proposed method can be useful for feature selection in other data classification task also. Therefore, in next section, we have discussed our proposed dynamic threshold calculation technique for medical text data classification.

4 Proposed Algorithm for Rare Word Finding

In medical text classification, our proposed algorithm for finding rare words includes two phases: first is dynamic threshold value calculation for selecting rare word and second is medical dictionary word matching [22]. The steps of our proposed algorithm for calculation of dynamic threshold are as follow:

4.1 Dynamic Threshold Determination for Finding Rare Words

Terminology

p = Number of documents, q = Number of features, $i = 1, 2, 3 \dots p$ document, $j = 1, 2, 3 \dots q$ feature, TF_{ij} = Term frequency of i th document and j th feature, IDF_j = Inverse document frequency of j th feature, AVG = Average term, MU = Multiplication term, S_j = Summation of term frequencies of feature j , SS = Summation of all multiplication terms MU , DT = Dynamic threshold.

Algorithm Steps

1. Find the sum of term frequency S_j of feature j with all 1 to p documents.
2. Find the $F_{AVG(j)}$ average of term frequencies of feature j : S_j is divided by p .
3. Compute $F_{IDF(j)}$ of feature j and perform the multiplication.

$$MU_j = F_{AVG(j)} * F_{IDF(j)}$$
4. Repeat the steps 1–4 for calculating multiplication terms MU_j for all features $j = 1, 2, 3 \dots q$.
5. Find the sum of all multiplication terms MU_q of feature called SS .
6. Find dynamic threshold DT , by averaging SS is divided by q .
7. Stop.

Above steps of proposed algorithm find the dynamic threshold for medical text data which is useful to select or find rare words. In addition to this algorithm, more filtering rare words can be found using medical term dictionary. This is useful step to identify domain-specific word matching in medical domain.

4.2 Dictionary-Based Rare Words Matching

For finding more important related to medical domain, we have used medical terminology dictionary which include 125,000 approximate terms and abbreviation are listed. The dictionary is created from GitHub [23], Google search, etc., sources. We have converted our dataset words in lower-case in preprocessing step. So, before matching with dictionary, all terms and abbreviations are converted to lower-case and then matching is performed. The above algorithm is tested with medical text data for document classification in next section. The whole section covered with different dataset, classifiers, and evaluation parameter measures with results.

5 Experimental Results

In this section, we used two real medical text datasets to investigate the performance of proposed approach for document classification. We shall start with dataset introduction, classifiers, and performance measures.

5.1 Datasets

For medical text data classification, full text research papers or articles with class labels dataset are downloaded from PubMed [24, 25]. Here, we are considering two real datasets. The first dataset has three classes (prostate, lung, and breast cancer). Dataset having around 29,437 instances [24]. Another dataset is downloaded from medical journals, NCBI collection and manually data labels provided. Dataset includes around 7500 samples with three classes like colon, thyroid, and lung cancer [25]. Dataset is cleaned and preprocessed for classification with TFIDF and BoW feature representation /extraction/selection method [5–7, 14, 15].

5.2 Rare Words

Medical text data is preprocessed and converted to document term matrix with TFIDF or BoW data representation. Here, we considered first dataset with sample size 1500 documents and applied dynamic threshold algorithm with dictionary-based word matching and got dynamic threshold value which is 12 as given in Table 1. The table shows rare word with term frequency less than equal to 12:

Table 1 List of rare words identified with term frequency threshold (12)

Bronchus (4)	CT (3)	Diagnosed (6)	Imaging (5)	Human (3)
Patient (5)	Sign (4)	Right (4)	Surgery (5)	Tramadol (3)
Suspected (3)	Lesion (4)	Pleural (3)	Identification (3)	Treatment (4)

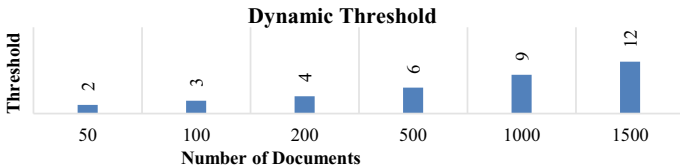


Chart 1 Dynamic threshold determined with different sample size

5.3 Threshold Values Versus Sample Size

In dataset, as sample size increase, features are also proportion increase. In classification task, number of samples in dataset vary in any dataset. So, we have tried different number of samples and determined dynamic threshold. Following Chart 1 shown that how threshold values are varying as number of samples are varied.

5.4 Proposed Method Evaluation

We have implemented the proposed algorithm using Python3.8 with supported libraries like machine learning, NLTK, etc. We have used traditional $TF \leq 3$, $IDF \leq 2$, percentage threshold (here, 30%) with TFIDF, BoW traditional approach. We have used three text classifiers which are best for text classification like k-nearest neighbor (kNN), Naïve Bayes (NB), and support vector machine (SVM) [5–8, 21]. For evaluation of all compared methods, the performance measures like accuracy, precision, recall, F1-measure, and execution time are used. We have selected features as rare terms/words/feature only and except rare words in two experiments. We have results with two group of data as explained below:

5.4.1 Evaluation of Results with Only Rare Words

As shown in following Table 2 and Chart 2, all methods of exist fixed threshold values and proposed method is compared here. For experimental purpose, we considered datasets which have selected random 1500 samples out of all samples for BoW results. Medical text data is preprocessed and further divided into training set and testing set for classification. The results shown that only rare terms are important

Table 2 Only rare words, 1500 samples with accuracy, precision, recall, F1-measure

Methods \ Classifiers	Accuracy			Precision			Recall			F1-measure			Total exe. time (Sec.)	Rare ws (Out of 34,056 words)
	KNN	SVM	NB	KNN	SVM	NB	KNN	SVM	NB	KNN	SVM	NB		
	BoW_AllTerms	81	97	97	81	97	97	81	97	97	81	97		
BoW_OnlyTF_rare (TF <= 3)	62	87	92	72	87	92	61	87	92	59	87	92	0.9073	26,763
BoW_OnlyTF_30%_rare	39	48	50	66	80	70	40	48	48	30	42	45	0.6292	10,217
BoW_OnlyIDF_rare (IDF <= 2)	51	60	69	70	80	77	48	60	69	43	59	69	0.872	18,223
BoW_Proposed_Thres_rare	62	89	90	66	90	90	62	90	90	62	89	90	17.0268	7102 (Dict)

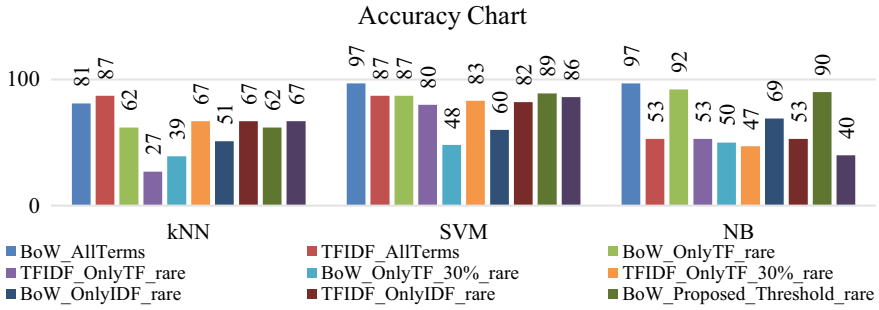


Chart 2 Accuracy of only rare word/features with sample size 1500 and BoW, TFIDF methods

in medical domain and give same performance as original data with all features in most of cases. With only rare terms (7102), proposed method is compared with other methods, and our method is outperformed.

5.4.2 Evaluation of Results with Excluding Rare Words

Result is shown in Chart 3 which as per traditional flow of text classification where rare words are removed. As per Chart 3, in some cases, rare terms' removing degrades performance and that proves rare words are very important. For experimental purpose, we considered dataset which has selected random 50 samples with BoW. Here, rare words in less quantity give same or nearer accuracy of all features method's accuracy, which is good example that rare words are important. Therefore, execution time of classification in proposed method is less. But our algorithm takes little second to find threshold and dictionary matching.

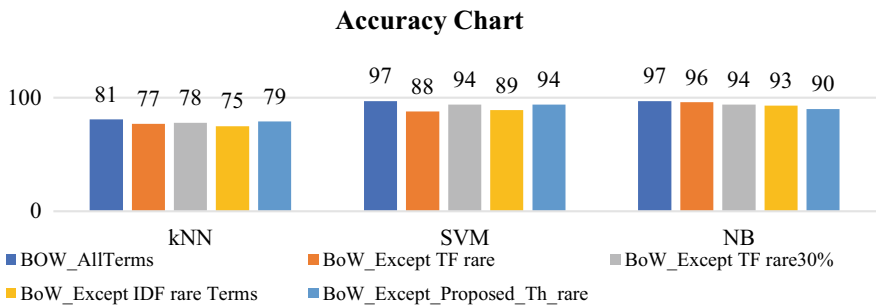


Chart 3 Accuracy of excluding rare words with sample size 50 and BoW method

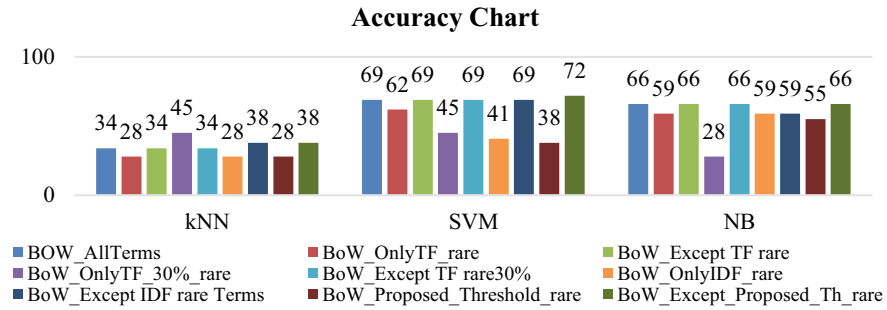


Chart 4 Only rare words and excluding rare words with accuracy measure and BoW method

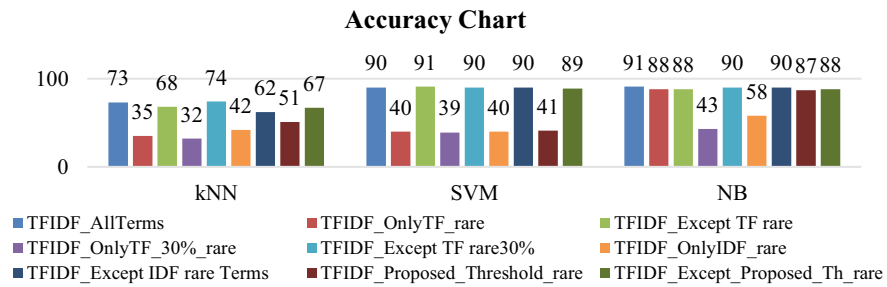


Chart 5 Only rare words and excluding rare words with accuracy measure with TFIDF method

5.4.3 Results Evaluation for Second Dataset

The second real dataset has considered 50 samples. As per Charts 4 and 5, results shown that with TFIDF, it performed better than BoW. It also shown that only rare words give nearer performance same as all features data. Our method finds best rare words. Except rare terms in any method, degrade accuracy in most of case. It shows that rare words are important in medical domain.

6 Conclusion

For medical text classification task, traditional flow of classification is applied directly in most of cases. Medical data is complex and has many domains specific words which is challenging task that is our focus. Rare words in medical domain are important. In literature, to find rare words, some statistical calculation-based static threshold value determination methods are available for text summarization, clustering, etc., so we proposed rare word finding method using dynamic threshold determination and medical dictionary-based word matching combined used. The proposed method

is compared with traditional approach with static threshold, and the results shown that dynamic method is well performed in most of case.

References

1. H.S. Yahia, A.M. Abdulazeez, Medical text classification based on convolutional neural network: a review. *Int. J. Sci. Bus. IJSAB Int.* **5**(3), 27–41 (2021)
2. X. Yan, J. Bien, Rare feature selection in high dimensions. *J. Am. Stat. Assoc.* (2020) <https://doi.org/10.1080/01621459.2020.1796677>
3. Al-D.I. Obaidat, M. Lee, Unstructured medical text classification using linguistic analysis: a supervised deep learning approach. in *2019 IEEE/ACS 16th International conference (AICCSA)* (2019), pp. 1–7, <https://doi.org/10.1109/AICCSA47632.2019.9035282>
4. L. Qing, W. Linhong, D. Xuehai, A novel neural network-based method for medical text classification. *Future Internet* **11**(12), 255 (2019). <https://doi.org/10.3390/fi11120255>
5. P.V. Arivoli, T. Chakravarthy, Document classification using machine learning algorithms—a review. *IJSER, ISSN (Online)* **5**(2), 2347–3878 (2017)
6. U. Naseem, M. Khushi, S.K. Khan, K. Shaukat, M.A. Moni, A comparative analysis of active learning for biomedical text mining. *Appl. Syst. Innov.* **4**(1), 23 (2021). <https://doi.org/10.3390/asi4010023>
7. R. Jindal, R. Malhotra, A. Jain, Techniques for text classification: literature review and current trends. *Webology* **12**(2) (2015)
8. R.T.W. Lo, et al., Automatically building a stopword list for an information retrieval system. *J. Dig. Infor. Mgmt.* **3**(1) (2005)
9. A. Holzinger, J. Schantl, M. Schroettner et al., in *Biomedical Text Mining: State-of-the-Art, Open Problems and Future Challenges*. Springer Lecture Notes in Computer Science, vol. 8401. Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-43968-5_16
10. M. Tahrawi, The role of rare terms in enhancing the performance of polynomial networks based text categorization. *J. Intell. Learn. Syst. Appl.* **05**, 84–89 (2013). <https://doi.org/10.4236/jilsa.2013.52009>
11. M. Tahrawi, The significance of low frequent terms in text classification. *Int. J. Intell. Syst.* **29** (2014). <https://doi.org/10.1002/int.21643>
12. G. Bathla, R. Jindal, Similarity measures of research papers and patents using adaptive and parameter-free threshold. *IJCA, ISSN 0975–8887* (2011)
13. L. Skorkovska, *Dynamic Threshold Selection Method for Multi-label Newspaper Topic Identification*. LNAI, vol. 8082, pp. 209–216 (Springer-Verlag Berlin Heidelberg, 2013)
14. S. Basheer, et al., Efficient text summarization method for blind people using text mining techniques. *Int. J. Speech Technol.* 1–13 (2020)
15. E. Padma Lahari, D.V.N. Siva Kumar, S. Prasad, Automatic text summarization with statistical and linguistic features using successive thresholds. *2014 IEEE Int. Conf. Adv. Commun. Control Comput. Technol.*
16. Li, Yanling, and Li Song, Threshold determining method for feature selection. in *2009 Second International Symposium on Electronic Commerce and Security*, vol. 2. IEEE (2009)
17. E. Marchiori, *Class Dependent Feature Weighting and K-Nearest Neighbor Classification* (Springer, 2013)
18. R. Roy, R. Homayouni, M.W. Berry, A.A. Pureskiy, *Nonnegative Tensor Factorization of Biomedical Literature for Analysis of Genomic Data*. https://doi.org/10.1007/978-3-642-45252-9_7.70
19. H. Christian, M. Agus, D. Suhartono, Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). *ComTech* **7**(4), 285–294 (2016)
20. N. Ishtayeh, in *Similarity Threshold Determination for Text Document Clustering*. Thesis of Master in CS (Zarqa University, Jordan, 2014)

21. J. Huang, Y. Wei, J. Yi, M. Liu, An improved kNN based on class contribution and feature weighting. *IEEE* (2018)
22. B. Settles, ABNER: an open-source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics* **21**, 3191–3192 (2005)
23. <https://github.com/glutanimate/wordlist-medicalterms-en>
24. https://figshare.com/articles/dataset/SparkText_SampleDataset_19681Abstract.zip
25. PubMed: www.pubmed.ncbi.nlm.nih.gov

Image-Based Spammer Analysis Using Suitable Features Selection by Genetic Algorithm in OSNs



Somya Ranjan Sahoo, Asish Kumar Dalai, Sanjit Ningthoujam, and Saroj Kumar Panigrahy

Abstract Online social networks brought about to change the fundamental of the people about the way they generate and share information. Over the last decades, gratuitous messages in the form of spam have become one of the most unplayful situations for social network users. In this paper, we report a novel framework to support precise and vigorous spam images filtering. The framework is developed based on multiple properties and features extracted by our crawler from different level of granularity, pointing to express more discriminative contents for effective spam image detection. Besides, a stochastic method used for function optimization (Genetic algorithm) based on natural genetics for suitable features selection. It can facilitate more accurate for identification of spam images by using sample training sets. By using some sample test collections, the proposed framework is compared with other state-of-the-art techniques. Our result presents a remarkably higher detection rate with a different outlook.

Keywords Online social networks · PSO · Facebook · Machine learning

1 Introduction

We are living in the present-day world where it is literally impossible to imagine the existence of life without the Internet. Earlier people used to communicate to each other through letters, telephones, and fax machines. All these types of communication medium are point to point because there is single receiver or the single sender. Hence, it is quite obvious that possibility of these media influencing the mass population is very minimal. But in the present-day world, Internet has taken the place of

S. R. Sahoo (✉) · A. K. Dalai · S. Ningthoujam · S. K. Panigrahy
VIT-AP University, Amaravati, Andhra Pradesh 522237, India

A. K. Dalai
e-mail: asish.d@vitap.ac.in

S. Ningthoujam
e-mail: sanjit.n@vitap.ac.in

these medium. Geographical borders had reduced a lot. They use the various social networking platforms like Facebook, Twitter, and Instagram to communicate to each other. At the present scenario, 2 billion [1] people are the active users of the site Facebook. There was a lot of work going on in order to determine the text-based spams, and a lot systems are already developed which can recognize a text-based spam with a decent accuracy rate. But detecting an image-based is still not so much developed area. Spam can be used to steal the confidential information of the people by redirecting them to the other pages and then stealing their confidential information. Image spams are developed and sent to the users in large numbers. Since all the images of a particular batch look almost similar, it is quite difficult to identify a spam image. Earlier Honeypots are used to detect an image-based spam, but these are not reliable. So there is an urgent need of developing a system which contains extensibility, high accuracy rate, and high efficiency. Spammers like Honeypots work on low-level features like shape, texture, and color. Since number of features are limited in Honeypot, that is why the accuracy of Honeypots is really low. In order to increase the accuracy, more OCR is appended with these features [2]. In the OCR, text is retrieved from the image which indirectly increases the number of features and that is why the accuracy automatically increased.

There is a need for optimization technique for better suitable feature selection by using every feature associated with images. But gathering features associated with the image is a great obstacle in this solution. In this paper, we use an optimization technique (Genetic algorithm) to analyze and select the best suitable features for detecting spam images in social network platform through machine learning and deep learning-based solutions.

The rest of the paper is organized as follows. Section 2 describes various work related to spam account detection. Section 3 describes the overall architecture and process flow to achieve the desired objective. Section 4 highlights the resultant analysis with various machine learning algorithms and genetic algorithm for suitable feature selection including the comparative analysis with existing approaches. Finally in Sect. 5, we discuss the conclusion and future research direction related to our topic.

2 Related Work

Authors in [3], used a technique consisting of principal component analysis (PCA) and support vector machine (SVM). The accuracy of their improved dataset is 70%. Authors in [4] use the global features of an image which includes color and histograms of the image, and the method used for classification is probabilistic boosting tree. The low-level features of the image are used in this. According to their experimental results, basic file properties of the system were able to detect almost 80% of spam images, and by using histogram analysis, this system was able to detect almost 84% of the spam images with a very low false positive rate. Authors in [5] big data processing is combined with the fingerprint techniques to detect the spam emails. Authors in [6], maintained the privacy preserving process by encrypting the information in cloud

environment. The effective fuzzy keyword search problem is solved by this suggested method. Author in [7] made a hybrid technique by combining the local and global features of an image. They also used OCR for retrieving the text. Author in [8], in this approach, suggested mail avenger model. This model works on the network layer and based on blacklist, whitelist, and graylist. Blacklist contains the information about that DNS names which we are sure that these DNS are definitely blocked. Whitelist contains the information about the DNS names which are legitimate. Graylist contains the DNS names which we are not sure about that they are legitimate or not.

Authors in [9] made an anti-phishing technique which uses the Chinese image spamming lexical features to detect phishing-based Web sites. Authors in [10] used visual-based features like email headers, image metadata, and classifiers used which are linear discriminative analysis (LDA) and random forests. Authors in [11] use the both low-level and high-level features of the image. The decision tree classifier method is used with these features for classifying an image into spam/ham. This method increased the accuracy rate as compared to the previous methods. Authors in [12] studied collectively the features in this approach. The features like aspect ratio, image area, compression ratio, bit depth, etc., are combined in this approach. Authors in [13] proved that the accuracy is increased by making a histogram of the image because it is easy to extract many features from the histogram like RGB values and entropy of the image.

The author uses the histogram with the existing systems and found that the accuracy increases a lot because the features used to train the dataset is increased. Author in [14] suggested an approach in which the image is taken partially step by step. First the partial image is generated along X-axis, and then, other partial image is generated along Y-axis called Sobel-X and Sobel-Y, respectively. Then, Laplacian image is generated by combining both Sobel-X and Sobel-Y.

3 Proposed System Architecture

In this section, we discuss the basic framework for the proposed approach. A crawler extracts various images from different social networking sites. Our feature extraction algorithm extracts various features of images and genetic algorithm for suitable feature selections. Figure 1 depicts all the basic architecture of our proposed framework. Spam images are the images which have a large amount of text, memes, or the cartoons, but ham images are the normal images of the person, animals, objects like car, chair, and a lot others. Another kind of spam includes those types of images which imbibes the links within them. These links redirect the user to the other site where the data or the confidential information of the people is stolen.

A good spam detection system must focus on the three parameters namely accuracy rate, efficiency, and extensibility, and all of these features are already contained in text-based spammers but since image spammers are quite new as compared to text-based spammers and this is only reason that these characteristics are lacking in the image spammers. Hence, there is a great scope of innovation in image spammers.

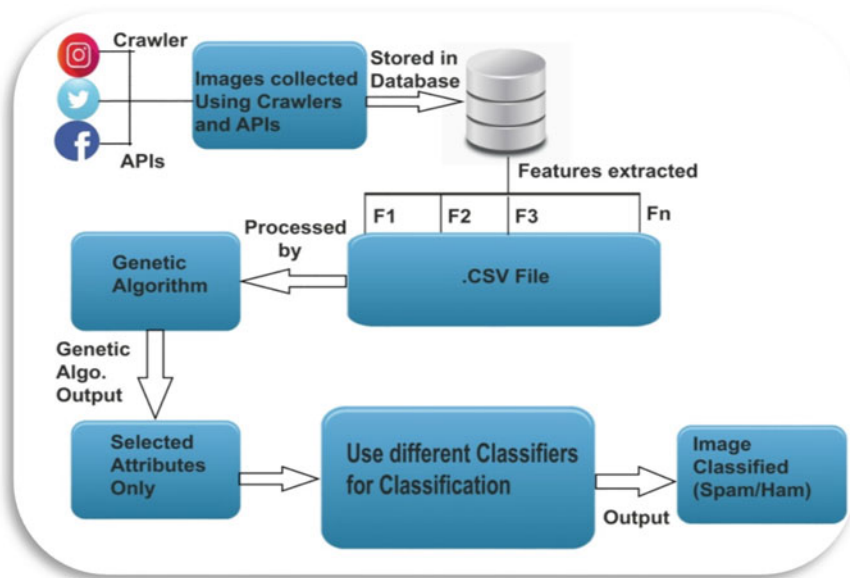


Fig. 1 Spammer detection framework

Since we are working on the images, we cannot directly use them for classifying an image into spam/ham. Therefore, we extracted some of the features from the images and then used those features for further classification. The features which are used in this project for classifying an image into spam/ham are depicted in Table 1.

So, in this paper, a framework for selecting the best suitable optimal set of features corresponding to the fake image detector has been presented. The complete framework is formulated based on genetic algorithm and various machine learning classifiers, including extraction of various features from different images for analysis. Various features extracted by our feature extraction algorithm is shown in Fig. 2a, b.

3.1 Genetic Algorithm for Suitable Feature Selection

A genetic algorithm (GA) formally stated as a stochastic approach for selecting best suitable features based on feature analysis depicted in Fig. 3. The best features which are present in the current population got selected, and the offspring which are produced in the next generation will get better features from the current one. It is also possible in one of the cases that the population which is present in current got vanished and completely new population which is mutated will appear in the next generation.

Table 1 Communicating entities in the proposed approach

Various features	Explanation
Width	It treats as the horizontal expansion of the image
Height	It is represented as a vertical expansion of the image
Size of image	The size of the image calculated from the disk space or file size it occupies
Aspect ratio	It is calculated from the ratio between the height and width of the image
Image area	The area of the image calculated by multiplying image height and width of the image
Compression ratio	Ratio of original image size to the reduced size of the image
Bit depth	Bit depth = the unique colors present in the color pallet of the image in the form of 0's and 1's
Total pixel	Total number of pixels present in an image
Saturation	It represents the color intensity of the image. We consider if it is 1, then the color is pure red, and if it is 0, then it is pure white
Entropy	It identified through the degree of randomness of an image
Average color	The average of the RGB represents the average of the color
Edge detection	Change in the abrupt value of an image
Color histogram	It represents the graph between RGB values
Sobel-X	Partial information of the image about the edge along X-axis
Sobel-Y	Partial information of the image about the edge along Y-axis
Hue	It identifies the pure color of an image, e.g.: (Yellow-1/6, Green-1/3)
Value	It represents the lightness of the image. Ex: (0-Black, 1-White)
Laplacian	It can calculate by combining Sobel-X and Sobel-Y of an image
Entropy of red	Randomness of red color in an image
Entropy of blue	Randomness of blue color in an image
Entropy of green	Randomness of green color in an image

3.2 Steps for Genetic Algorithm

The whole process of genetic algorithm works in five different steps. Each steps have its own existence related to the data content for operation shown above in Fig. 3.

Initialization: Every population consists of the individuals and every individual have some features. If a particular feature is present in individual A, then it is denoted by 1 but that peculiar feature is absent, and then, it is denoted by 0. It basically talks about the presence as well as absence of all the features in the whole population.

Fitness Function: It determines how much fit is a particular individual. Some features are given more importance in comparison with the other features. Here, weights are assigned to all the features along the profit values. Hence, it will be more beneficial in the second generation; it is more profitable feature which is having the less weight that is selected. For example: It is based on probability. In first iteration,

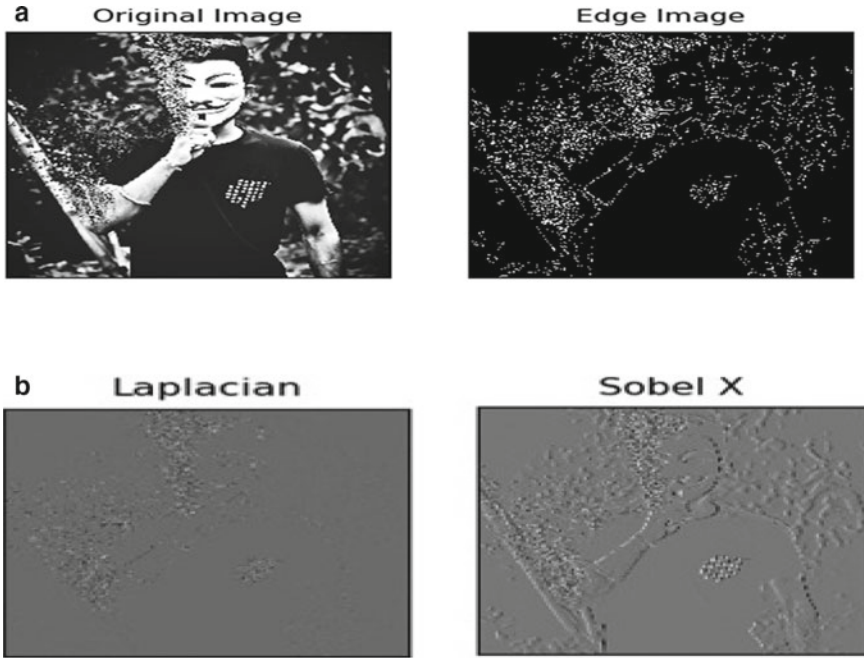


Fig. 2 a Original image and edge of an image, b Laplacian and Sobel-X of an image

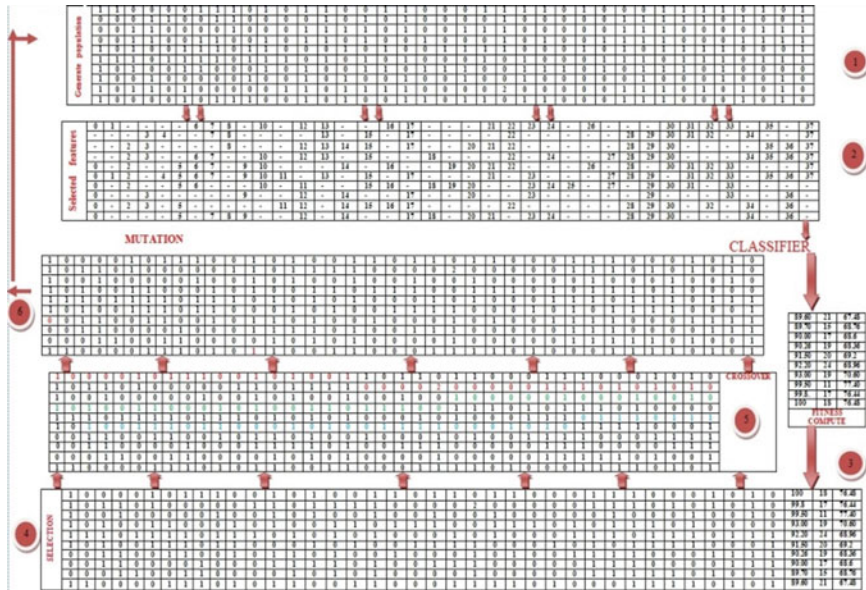


Fig. 3 Genetic algorithm for selecting features

Fig. 4 Crossover

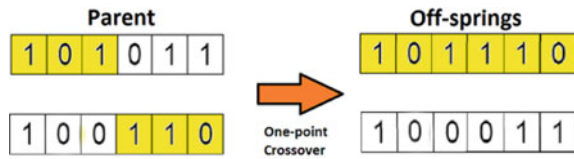


Fig. 5 Mutation



some features are selected, and in further iterations, some different features given more importance and that are selected.

Selection: Fitness value is given for the whole population so for breeding the individuals which are having high fitness value got selected. The individual whose fitness function value is low which is dropped out. For example: In our dataset, hall of fame is maintained which contains the index of the individual having the maximum fitness.

Crossover: It is the step in which parents are selected, so that they can breed further for producing out the next generation. For example: in our dataset from the hall of fame individual list, probability again comes, and from that list, only individuals' mate to produce the next generation is depicted in Fig. 4.

Mutation: Sometimes it happens that the new offspring which get produced have different features from the parents. This is because of the fact that some values of features get changed during the duration when they are inherited from the parents. As a result, features of offspring are different from the parent. For example: in our dataset, two individuals having specific features are mated, but sometimes offspring contains the different features from the parents depicted in Fig. 5.

The complete genetic algorithm can be coded in Python using the combination of inbuilt functions and by providing some other inputs like mutation probability.

4 Experimental Analysis and Result

The dataset which we made is processed with the genetic algorithm. After applying genetic algorithm to the collected dataset, we filtered out the best suitable features. The collected best suitable features are applied to which various classifiers in machine learning environment. It can be seen very clearly that after using genetic algorithm, accuracy rate has always increased compared to original dataset depicted in Table 2.

Table 2 Comparison of results of various classifiers

Sr. no.	Different machine learning-based classifiers	Accuracy without genetic algorithm (%)	Number of features after 50th iteration in GA	Accuracy with genetic algorithm (%)	Number of features after 50th iteration in GA	Accuracy with genetic algorithm (%)
1	Random forest classifier	94.3	19	98.6	15	98.9
2	Decision tree classifier	92.6	19	97.28	15	97.87
3	Logistic regression classifier	93.98	19	97.21	15	98.23
4	Gradient booster classifier	94.76	19	98.01	15	98.64
5	Naïve Bayes classifier	92.95	19	95.37	15	95.98
6	Bagging classifier	94.03	19	94.16	15	95.67
7	SVM	96.35	19	98.98	15	99.17

4.1 Graphical Analysis of Results

The graphical representation of various experimental analysis depicted in Figs. 6 and 7. We observed that after 100th iteration, the accuracy is more as compared to

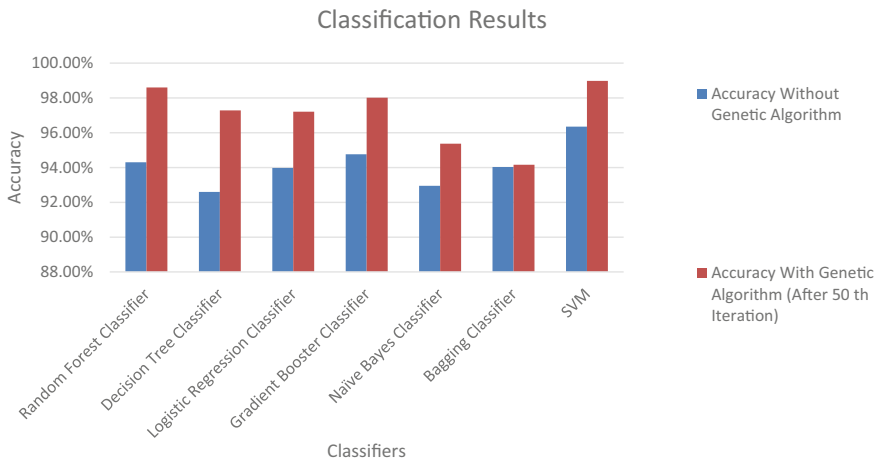


Fig. 6 Comparative analysis of result with and without genetic algorithm with 50th iteration

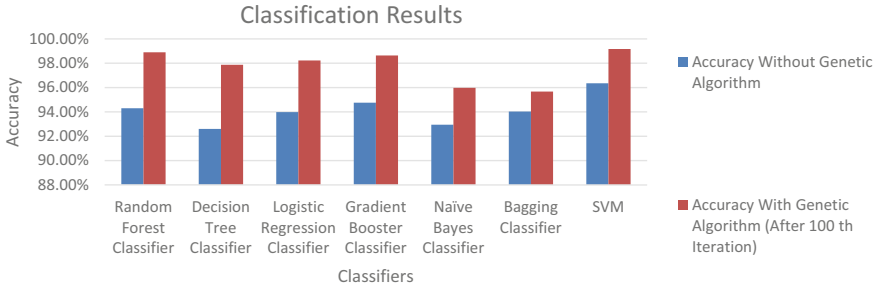


Fig. 7 Comparative analysis of result with and without genetic algorithm with 100th iteration

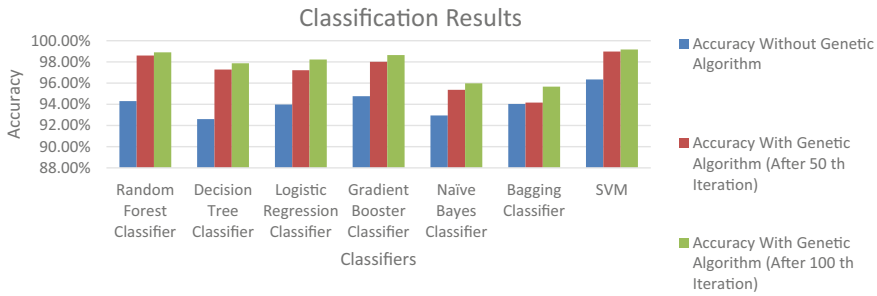


Fig. 8 Comparative analysis of results with and without genetic algorithm with 50th and 100th iteration

50th iteration result. Also, the genetic algorithm-based machine learning analysis gives more accurate result compared to without GA. The comparative analysis of various results through machine learning classifiers and genetic algorithm-based feature selection with number of iteration is depicted in Fig. 8.

4.2 Comparative Analysis with Other Existing Approaches

Till now, if we consider all the work that had been done on detecting spam images in social network platform, then we found out that maximum of ten features are used for classifying an image as spam. But, in our paper, we used all the high-level and low-level features of various images, and also seven different classifiers are used which gives an additional edge to the accuracy. The comparative analysis of our framework with other existing approach is given in Table 3.

Table 3 Comparative analysis of our approach with existing approaches

Sr. no.	Research paper	Accuracy (%)	Tools and technique used for evaluation
1	[6]	70	PCA and SVM based technique used
2	[9]	84	Histogram analysis method is used for extracting feature, and then, classifiers like SVM used for classification
3	[15]	89.44	Features such as color and gradient orientation histogram are basis of classification, and the classifier used is probabilistic boosting tree (PBT)
4	Our proposed scheme	99.17	Our proposed approach analyzes various characteristics of the image as features, genetic algorithm for suitable feature selections, and various machine learning algorithms to analyze spammer content in social network environment

5 Conclusion and Future work

A spam image detection system is developed which will help a lot in detecting that whether a particular image is spam or ham image. A spam image is having more text in comparison with a normal image. A genetic algorithm is also implemented which will help a lot in increasing the accuracy rate by selecting out the optimal number of features from all the features of the dataset. If we select all the features, then we have to compromise with the complexity of the program, but if we select a limited number of features, then there might be the chances that false positive rate of the image increases. Hence, genetic-based algorithm is of utmost importance in order to increase the accuracy of the spam/ham image detection system. Once optimal features are selected, then various classifiers are implemented which predict that whether the given image is ham/spam image. Since a total of six classifiers are used, so it will give overall clarity that whether a particular image is spam/ham because if majority of classifiers say that an image is spam then probability of being a spam image is more for that instance of image.

References

1. M. Fire, R. Goldschmidt, Y. Elovici, Online social networks: threats and solutions. *IEEE Commun. Surv. Tutorials* **16**(4), 2019–2036 (2014)
2. F. Gargiulo, A. Penta, A. Picariello, C. Sansone, Using heterogeneous features for anti-spam filters. in *19th International Workshop on Database and Expert Systems Application, 2008. DEXA '08*. IEEE (2008), pp. 670–674
3. B.B. Gupta, S.R. Sahoo, *Online Social Networks Security: Principles, Algorithm, Applications, and Perspectives* (CRC Press, 2021)
4. S.R. Sahoo, B.B. Gupta, Classification of spammer and non-spammer content in online social network using genetic algorithm-based feature selection. *Enterp. Inf. Syst.* **14**(5), 710–736

- (2020)
5. P. Parekh, K. Parmar, P. Awate, *Spam URL Detection and Image Spam Filtering using Machine Learning* (2018)
 6. J. Chen, H. Zhao, J. Yang, J. Zhang, T. Li, K. Wang, An intelligent character recognition method to filter spam images on cloud. *Soft. Comput.* **21**(3), 753–763 (2017)
 7. S.R. Sahoo, B.B. Gupta, Popularity-based detection of malicious content in facebook using machine learning approach. in *First International Conference on Sustainable Technologies for Computational Intelligence* (Springer, Singapore 2020), pp. 163–176
 8. Mail Avenger (2006). <http://www.mailavenger.org>
 9. P. Parekh, K. Parmar, P. Awate, Spam URL detection and image spam filtering using machine learning. *Comput. Eng.* (2018)
 10. S.R. Sahoo, B.B. Gupta, Security issues and challenges in online social networks (OSNs) based on user perspective. in *Computer and Cyber Security* (Auerbach Publications 2018), pp. 591–606
 11. He, P., Wen, X., & Zheng, W. (2009, June). A simple method for filtering image spam. in *IEEE Eighth IEEE/ACIS International Conference on Computer and Information Science, 2009. ICIS* (2009), pp. 910–913
 12. C. Wang, F. Zhang, F. Li, Q. Liu, Image spam classification based on low-level image features. in *IEEE 2010 International Conference on Communications, Circuits and Systems (ICCCAS)* (2010), pp. 290–293
 13. S.R. Sahoo, B.B. Gupta, Multiple features based approach for automatic fake news detection on social networks using deep learning. *Appl. Soft Comput.* **100**, 106983 (2021)
 14. C. Xu, Y. Chen, K. Chiew, An approach to image spam filtering based on base64 encoding and N-Gram feature extraction. in *IEEE 2010 22nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, vol. 1 (2010), pp. 171–177
 15. Y. Gao, M. Yang, X. Zhao, B. Pardo, Y. Wu, T.N. Pappas, A. Choudhary, Image spam hunter. in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP* (2008), pp. 1765–1768

An Improved Firefly Algorithm Based Cluster Analysis Technique



Manju Sharma and Sanjay Tyagi

Abstract Unsupervised machine learning approach like cluster analysis finds a large number of applications in different engineering domains. A variety of meta-heuristic algorithms have been proposed in the literature for clustering. Firefly is one of the most commonly used meta-heuristic algorithm as it has efficient capability of automatic subdivision of population and natural capability of dealing with multimodal optimization. But due to more dependency on local solution for movement, it generally leads to premature convergence. In this paper, an improved variant of firefly algorithm is proposed by introducing a new position updating equation for movement of firefly by using the idea of best solution for global search. A mutation operator is also incorporated in the basic firefly algorithm to enhance its convergence speed and exploration capability. The proposed firefly algorithm is simulated and compared with standard firefly algorithm on standard 13 benchmark functions. Moreover, the efficiency of the proposed firefly algorithm is also tested by adopting it as a clustering technique. The performance is tested on seven real-life datasets and also compared with various state-of-the-art meta-heuristic clustering techniques. The computation outcomes showed that the proposed algorithm is better in finding the optimal cluster center with minimum intra-cluster distance, along with fast convergence speed. Results are also verified quantitatively using Friedman, Wilcoxon, and post-hoc pairwise Nemenyi tests.

Keywords Clustering · Firefly algorithm · Mutation · Meta-heuristic

1 Introduction

Clustering is a popular unsupervised machine learning technique that assembles the objects into clusters so as to maximize the cohesiveness between the objects of same

M. Sharma (✉)
Government College for Women, Karnal, India

S. Tyagi
Department of Computer Science and Applications, Kurukshetra University, Kurukshetra, India

cluster and minimize the uniqueness between the objects of different clusters [1]. Most of the real-life problems can either be transformed or can be acknowledged to this type of problem [2]. Therefore, at present time, cluster analysis techniques are getting more attention from researchers from different fields like feature selection, image processing, healthcare, e-learning, gene expressions, process monitoring, etc. [3].

Cluster analysis is an NP hard problem [4]. Consequently, generating optimal quality solution for large scale problems is very difficult and expensive. Through literature survey, it is observed that a numerous meta-heuristic algorithms (like particle swarm optimization (PSO) [5], genetic algorithm (GA) [6], differential evolution (DE) [7], firefly algorithm (FA) [8], etc.) have been designed by researchers for clustering problems. These algorithms are independent of problem size, solution space and are capable to adapt according to the problem domain. Still, many issues are found associated with these techniques in finding the optimal solution for clustering problems like premature convergence, imbalanced exploitation, and exploration phases, convergence speed, etc. Many researchers have addressed these issues by designing some variants or by using some hybrid versions of other meta-heuristic algorithms for efficient clustering results [9, 10].

The key contribution of this paper is to produce a new variant of firefly algorithm (FA). The firefly algorithm was introduced in 2009 by Yang [11]. FA has powerful exploration capability as compared to other similar algorithms; however, it also suffers from local optima trapping problem and slow convergence rate. Consequently, to make FA algorithm more robust and efficient for analysis problems, some improvements are incorporated in classic firefly algorithm as:

- A new position updating (movement of firefly) equation is proposed.
- A mutation operator is also introduced to explore the optimum solution.

The remaining sections are systematized as follows. Section 2 provides the details regarding clustering problem. Section 3 describes the basic firefly algorithm and its proposed improvements. Improved firefly algorithm as a cluster analysis technique has been introduced in Sect. 4. Simulation results and statistical tests are described in Sect. 5. Section 6 summarizes the entire work and its future views.

2 Background

Consider a dataset X having n data points, $\{X = P_1, P_2, \dots, P_n\}$ the clustering model divides the datasets among k clusters or partition $\{C_1, C_2, \dots, C_k\}$ such that: (i) Cluster should not be empty. (ii) Each data point must belong to only single cluster [12]. To measure the similarity between the two data points, Euclidean distance is the most commonly used similarity measure and is defined as:

$$D(P_i, P_j) = \sqrt{\sum_{m=1}^d |P_{im} - P_{jm}|^2} \quad (1)$$

where $D(p_i, p_j)$ is the distance between the object i and j . p_{im} represents the j th attribute of i th point, such that for each point $p_{im}(i= 1, 2, 3... n, j = 1, 2, 3 \dots d)$, n indicates the number of instances in a dataset and d represents the attributes of each instance. In this paper, the Euclidean distance is calculated between each data point/object of a dataset to the cluster centers. Then, the object is allocated to the cluster with minimum value of Euclidean distance. Consequently, the center of a cluster is substituted with the calculated average of all objects belong to the corresponding cluster. Sum of squared error (SSE) is used as measurement criteria for validation index. SSE is defined as:

$$SSE = \sum_{i=1}^k \sum_{P \in C_i} ed^2(Z_i, P) \quad (2)$$

where Z_i (centroid of cluster) is the average of distance of all data points belonging to cluster C_i and is defined as:

$$Z_i = \frac{1}{N_i} \sum_{P \in C_i} P \quad (3)$$

Here, N_i specifies number of data points which belongs to cluster C_i , P is a vector term that specifies all the data points of C_i , and ed^2 specifies the Euclidean distance.

3 Firefly Algorithm and Its Proposed Variant

3.1 Traditional Firefly Algorithm

Firefly algorithm developed by Yang [11] is a population-based swarm intelligence algorithm that simulates the behavior of fireflies. The three idealized rules regarding fireflies are (i) as the fireflies are unisex, hence, each firefly can attract any other firefly. (ii) The important characteristics of firefly is that they glow brighter generally to share their foods as well as to prevent from the predators. (iii) Attraction is based on the brightness, and as the distance increases, brightness decreases. Initially, all the fireflies (as agents) are randomly distributed in the search space. Each firefly emits some amount of light through bio-luminescence process. The attraction between the fireflies is proportional to their glowing capacity. Fitness function value indicates the brightness of the firefly. As the distance increases, attractiveness decreases. When

there is no brighter firefly in the neighbor of a particular firefly, then it follows the random movement. The two important phases in firefly algorithm are:

1. Light intensity variation: Light emission is based on the value of objective function. Let x_i is the position of i^{th} firefly. Then, its light intensity is proportional to its fitness value, i.e.,

$$I_i = f(X_i) \quad (4)$$

2. Attractiveness and movement of firefly: Each firefly has its own attractiveness (β) value that usually varies with the distance between the two fireflies and is defined as:

$$\beta = \beta_0 e^{-\gamma r^2} \quad (5)$$

where r is the Euclidean distance between the two fireflies, β_0 is the initial attraction coefficient, and γ is the light absorption coefficient. The movement of firefly i from position X_i toward more attractive (brighter) firefly j at position X_j specifies as:

$$X_i(t+1) = X_i + \beta_0 e^{-\gamma r^2} (X_j - X_i) + \alpha \varepsilon \quad (6)$$

here α is a random number (0–1) and ε is the randomness from the Gaussian distribution, and these two terms represent the random walk of a firefly.

3.2 Proposed Improvements in Firefly Algorithm (IFA)

- In basic firefly algorithm, brighter firefly (local to the iteration) influences the movement of all other fireflies, and the movement is irrespective of the global optima that may lead to premature convergence. In order to remove this weakness, a new movement variant of firefly algorithm is introduced as follows:

$$X_i(t+1) = X_i + \beta_0 e^{-\gamma r^2} (X_j - X_i) + \alpha \varepsilon e^{-\gamma r^2} (X_{\text{gbest}} - X_i) \quad (7)$$

If the fitness of firefly (j) is better than firefly (i), then firefly (i) will move toward firefly (j), the movement explained in the second part of above equation. But if the fitness is not better, then instead of random movement, firefly moves according to the global best solution as shown in third part of Eq. 7.

- In order to further enhance the diversity in the searching phase in the basic algorithm, a mutation operator is also introduced in the proposed algorithm. The new solution generated after each inner loop is further enhanced by using a mutation operator having mutation probability 0.1.

4 Improved Firefly Algorithm (IFA) as Clustering Technique

In this research, proposed improved firefly approach is used for data clustering. Each firefly is representing as a string of real numbers and is considered as the center of clusters (k). Each firefly in d dimension is represented in the form of matrix ($k * d$), where k is the number of clusters and d is the number of instance of a dataset from a clustering problem.

IFA Clustering

Input:

Define number of cluster centers(k) Population size(n_{pop}), $MaxItr, \alpha, \beta, \gamma$.

Output:

Optimum Cluster Center, Intra cluster distance

Begin

Randomly initialize the fireflies population with k number of cluster centers.

Set the light intensity of each firefly by evaluating fitness function. using Eq 4.

Set $itr=0$;

While ($itr < MaxItr$) **do**

For all fireflies ($i=1, 2, \dots, N$) **do**

For all other fireflies ($j=1, 2, \dots, N$) **do**

 If ($f(x_i) < f(x_j)$) then // consider problem as minimization problem

 Move the i^{th} firefly towards j^{th} firefly according to Eq7.

 End if

End For loop

 Calculate fitness of new solution and if new solution is better than old one, accept.

 Apply mutation on new solution with mutation probability 0.1.

 Update the cluster centers based on updated positions of the new solution.

 Update the light intensity using Eq 4.

End For loop

Rank all fireflies & find the global optimum solution;

Increment $itr=itr+1$;

Reduce α by a random factor.

End while

 Partition the data points of the given dataset according to the optimal cluster centers given by global best solution

Return optimal cluster centers.

End

Algorithm 2 Pseudocode of improved firefly-based clustering algorithm.

Table 1 Comparison between FA and IFA algorithm for benchmark functions

Algorithm function	FA		IFA		Wilcoxon
	Mean	Std. dev	Mean	Std. dev	p-value
F1	0.2459	0.0465	4.15E-16	2.78E-16	1.83E-04
F2	2.1402	0.1274	1.51E-10	5.33E-11	1.83E-04
F3	0.0123	0.0032	2.56E-23	6.63E-23	1.83E-04
F4	0.0607	0.0094	4.47E-22	5.51E-22	1.83E-04
F5	7.3225	2.2549	4.4256	2.8281	0.0022
F6	0.2597	0.0182	2.53E-16	1.68E-16	1.83E-04
F7	0.1927	0.0454	0.0068	0.0029	1.83E-04
F8	-3.50E + 03	180.9159	-3.63E + 03	217.1619	0.1403
F9	80.841	10.7406	38.3059	10.0622	1.83E-04
F10	0.7505	0.0605	4.58E-09	2.67E-09	1.83E-04
F11	0.0164	0.0045	0.0071	0.0145	0.0027
F12	0.0858	0.1757	0.0829	0.0818	0.9097
F13	0.0491	0.0059	0.0044	0.0057	1.83E-04

5 Results and Discussions

5.1 Comparison Between FA and IFA Using Standard Benchmark Functions

Firstly, the proposed (IFA) approach has been executed in MATLAB(R2013a) and compared with standard firefly algorithm using 13 standard real encoded benchmark functions given in paper [13]. Functions F1-F7 are unimodal and F8-F13 are multimodal functions. Each algorithm has been simulated independently for 20 runs, and results are given in Table 1. It is clear from the simulation results that the proposed IFA outperforms standard FA. In order to validate the supremacy of IFA over FA, Wilcoxon test has been performed using level of confidence = 0.05. It is clear from p-value of Wilcoxon test that the proposed IFA is statistically significant for all the functions except F12.

5.2 Performance Evaluation as Clustering Technique

PSO, GA, FA, and DE are the powerful meta-heuristic cluster analysis techniques. These approaches are capable to generate solution in reasonable amount of time. Still, there is no assurance of providing optimal solutions. In this research, the performance of IFA is further checked as clustering technique by using seven standard datasets from UCI repository [14]. The detailed description (attributes, instances,

Table 2 Dataset properties

Dataset	Instances	Attributes	Clusters
Iris	150	4	3
CMC	1473	9	3
Wine	178	13	3
Vowel	871	3	6
Thyroid	215	5	3
Cancer	683	9	2
Glass	214	9	6

no. of classes) of datasets is given in Table 2. Furthermore, the IFA based clustering is compared with other well-known clustering techniques like GA, PSO, DE, FA, and k-means algorithms in terms of intra-cluster distance and cluster quality. Each algorithm has run 20 times independently. Results are taken as the best, mean, standard deviation and worst value of intra-cluster distance of different datasets. Table 3 revealed that the proposed IFA algorithm provides better mean value of intra-cluster distances for all the datasets.

It is clear from the results that the IFA provides enhanced quality clusters than the other algorithms. Moreover, the convergence behavior of the proposed IFA is also checked between different approaches. Figure 1 shows the convergence behavior that clearly indicates that the IFA has better convergence speed and preventing the problem regarding local optima.

5.3 Statistical Analysis of Proposed Clustering Algorithm

The quantitative performance verification of proposed algorithm has been carried out here using statistical Friedman test [15]. This test specifies the difference, if any, in performances of compared algorithms and proposed algorithm. The two hypothesis are framed. (H_0 : Null hypothesis, H_1 : Alternate hypothesis), i.e., H_0 : represents No significant difference; H_1 : represents significant difference.

Table 4 shows the ranking and average ranking of algorithms with Friedman test. The observations support the proposed IFA with better ranking (i.e., 1.0) than other compared algorithms. As reported in Table 5, statistical value of results is 19.65306 for 0.05 as level of confidence and degree of freedom, and the critical value equals to 5 and 11.0705, respectively. With p -value of 0.001452, it rejects the null hypothesis and further supports that there is significant difference in the performances of proposed and other compared algorithms. Subsequently, a post-hoc pairwise Nemenyi (multiple comparison) test is conducted to discern which of the pairs have significantly differences. The results illustrated in Table 6 reveal the substantial difference in the performance of proposed and other compared algorithms.

Table 3 Comparison based on intra-cluster distance

Dataset		Algorithms					
		IFA	PSO	GA	DE	FA	K-means
Iris	Best	96.6555	96.667	96.8605	97.7523	96.677	97.3259
	Mean	96.6555	100.354	100.9961	100.4732	97.0924	99.9902
	STD	2.92E-14	6.4163	6.6501	1.8676	1.1753	8.2007
	Worst	96.6555	120.2707	124.325	103.5754	100.5301	123.969529
Wine	Best	16,292.19	16,310.35	16,315.1844	16,304.5024	16,294.7782	16,555.6794
	Mean	16,292.45	16,340.55	16,349.3869	16,317.5689	16,301.52873	16,894.9222
	STD	0.5170	23.0442	44.2227	7.5632	6.36363	696.9479
	Worst	16,294.33	16,402.2	16,467.6176	16,330.4457	16,320.8715	18,294.8465
CMC	Best	5532.185	5704.635	5577.5848	5558.1338	5565.7569	5543.51194
	Mean	5532.195	5865.415	5719.02962	5580.34577	5643.52444	5544.03169
	STD	0.010805	119.0598	88.4269913	13.3465202	55.78170507	0.7872082
	Worst	5532.221	6208.916	5908.5102	5599.7119	5760.1204	5545.2005
Vowel	Best	148,967.2	149,041.1	152,324.37	155,536	149,041.1318	149,398.66
	Mean	149,494.3	149,578.7	159,353.94	164,875	149,835.2749	152,004.998
	STD	566.,9103	1267.833	3,005.45	4496.96	814.1339896	3435.75423
	Worst	150,156.9	153,051.9	166,882.56	170,533.52	150,840.0079	159,220.773
Thyroid	Best	1866.466	1915.768	1905.95306	1876.97548	1868.293903	1983.07595
	Mean	1885.642	2069.051	1985.16789	1902.20504	1886.403881	1990.94573
	STD	9.63438	113.0255	76.7519689	13.495368	9.714267096	6.01127397
	Worst	1890.207	2273.129	2142.86976	1926.01866	1896.00117	2001.63582
Cancer	Best	2964.387	2990.019	2971.8524	2966.4831	2964.491	2988.42781
	Mean	2964.392	3193.091	3050.3839	2981.6634	2964.587337	2988.42781
	STD	0.023255	162.8198	85.0111537	13.5045685	0.100049878	9.3312E-13
	Worst	2964.491	3685.731	3283.7769	3012.1405	2964.8486	2988.42781
Glass	Best	210.43	215.09	243.95	269.46	253.79	215.470412
	Mean	214.5311	230.031	248.974	283.694	273.045	223.621477
	STD	2.620445	14.83343	4.46757254	5.977748	11.11695127	11.8951246
	Worst	218.94	251.51	256.45	289.12	286.2	246.91

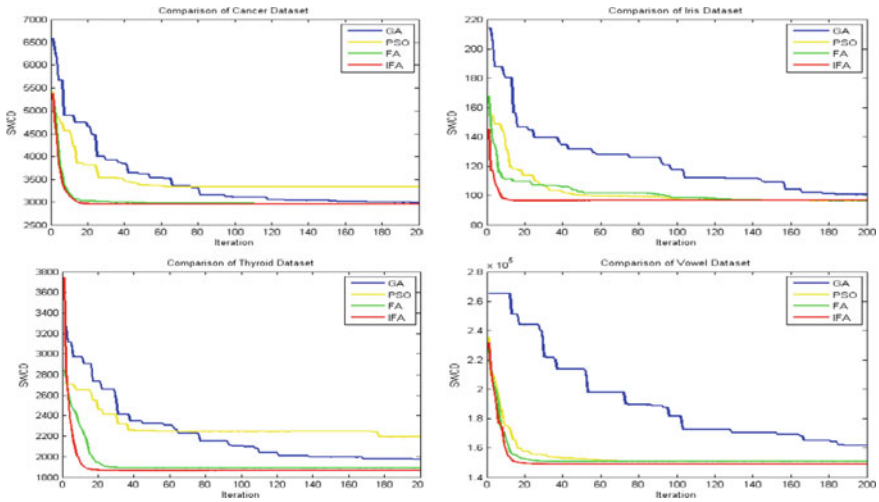


Fig. 1 Convergence behavior of algorithms for different datasets

Table 4 Friedman test-based ranking of algorithms using average intra-cluster distance

Algorithms/Datasets	IFA	PSO	GA	DE	FA	K-means
Iris	1	4	6	5	2	3
wine	1	4	5	3	2	6
CMC	1	6	5	3	4	2
Vowel	1	2	5	6	3	4
Thyroid	1	6	4	3	2	5
Cancer	1	6	5	3	2	4
Glass	1	3	4	6	5	2
Average ranking	1.00	4.43	4.86	4.14	2.86	3.71

Table 5 Statistical test result of Friedman test

Test	Statistical value	Critical value	Hypothesis	<i>p</i> -value
Friedman	19.65306	11.0705	Rejected	0.001452

Table 6 Nemenyi post-hoc test results (unadjusted *P* values) using intra-cluster distance

	IFA	PSO	GA	DE	FA
PSO	0.007985				
GA	0.001601	0.998171			
DE	0.020756	0.999745	0.980292		
FA	0.428918	0.617519	0.342180	0.793037	
K-means	0.072404	0.980292	0.863453	0.998171	0.956477

6 Conclusions and Future Scope

In this work, an improved firefly algorithm is proposed to solve the real-world clustering problems. To remove the local optima in standard firefly algorithm as well as to improve the convergence speed, a mutation operator and modified random movement equation are proposed as improved firefly algorithm. Cluster center-based encoding scheme is used for cluster analysis approach, and the proposed approach is compared with other clustering algorithms. It is observed that the proposed improved firefly clustering algorithm outperforms other compared algorithms in terms of better convergence speed and cluster quality. The results are also verified by using Friedman and Wilcoxon tests that clearly indicates the significance of proposed algorithm quantitatively. In future, the proposed algorithm can be used for other NP hard problems like protein synthesis, image segmentation, etc.

References

1. J.A. Hartigan, *Clustering Algorithms*, 1st edn. (Wiley, New York, 1975)
2. O. Maimon, L. Rokach (eds.), *Soft Computing for Knowledge Discovery and Data Mining* (Springer, New York, 2008)
3. P. Shabanzadeh, R. Yusof, An efficient optimization method for solving unsupervised data classification problems. *Comput. Math. Methods Med.* 10 (2015). <https://doi.org/10.1155/2015/802754>
4. W.J. Welch, Algorithmic complexity: three NP-hard problems in computational statistics. *J. Stat. Comput. Simul.* 15(1), 17–25 (1982). <https://doi.org/10.1080/00949658208810560>
5. D.W. van der Merwe, A.P. Engelbrecht, Data clustering using particle swarm optimization, in *The 2003 Congress on Evolutionary Computation, 2003. CEC '03*, vol. 1 (2003), pp. 215–220. <https://doi.org/10.1109/CEC.2003.1299577>
6. U. Maulik, S. Bandyopadhyay, Genetic algorithm-based clustering technique. *Pattern Recogn.* 11 (2000)
7. W. Kwedlo, A clustering method combining differential evolution with the K-means algorithm. *Pattern Recogn. Lett.* 32(12), 1613–1621 (2011). <https://doi.org/10.1016/j.patrec.2011.05.010>
8. H. Malik, N.-U.-Z. Laghari, D.M. Sangrasi, Z.A. Dayo, Comparative analysis of hybrid clustering algorithm on different dataset. in *2018 8th International Conference on Electronics Information and Emergency Communication (ICEIEC)* (2018), pp. 25–30. <https://doi.org/10.1109/ICEIEC.2018.8473568>
9. W.A. Khan, N.N. Hamadneh, S.L. Tilahun, J.M.T. Ngotchouye, A review and comparative study of firefly algorithm and its modified versions. *IntechOpen* (2016). <https://doi.org/10.5772/62472>
10. M. Sharma, J.K. Chhabra, Sustainable automatic data clustering using hybrid PSO algorithm with mutation. *Sustain. Comput. Inf. Syst.* 23, 144–157 (2019). <https://doi.org/10.1016/j.suscom.2019.07.009>
11. X.-S. Yang, Firefly algorithms for multimodal optimization. *Stochastic Algorithms: Found. Appl.* 169–178 (2009). https://doi.org/10.1007/978-3-642-04944-6_14
12. R. Xu, D. Wunsch, *Clustering*. John Wiley & Sons (2008)
13. M. Khishe, M.R. Mosavi, Chimp optimization algorithm. *Expert Syst. Appl.* 149, 113338 (2020). <https://doi.org/10.1016/j.eswa.2020.113338>

14. UCI Machine Learning Repository: Data Sets. <https://archive.ics.uci.edu/ml/datasets.php>. Accessed on 28 Apr 2021
15. S.W. Scheff, Chapter 8—nonparametric Statistics. in *Fundamental Statistical Principles for the Neurobiologist*, ed. by S.W. Scheff (Academic Press, 2016), pp. 157–182. <https://doi.org/10.1016/B978-0-12-804753-8.00008-7>

Stock Market Analysis of Beauty Industry During COVID-19



Satya Verma, Satya Prakash Sahu, and Tirath Prasad Sahu

Abstract COVID-19 has significant influence on the financial market. This paper aimed to explore the COVID-19 scenario analysis for stock market of beauty industry. Stock data of Estée Lauder Companies (EL), Revlon Inc. (REV) and Coty Inc. (COTY) is considered for this purpose. Deep learning models (LSTM and CNN) are utilized for the stock price prediction of beauty companies during COVID-19 era. LSTM and CNN, both the model worked well for the stock price prediction; however, LSTM performed better in all cases. Lockdown scenario along with the stock data is taken for the analysis purpose. Study shows that beauty industries got affected during initial spread of the virus, but now recovering.

Keywords COVID-19 · Stock market · Beauty industry · Deep learning · LSTM · CNN

1 Introduction

COVID-19 as pandemic not only hampered people's life but their health and wealth too. During this pandemic, many countries imposed lockdown for different duration as per the situation. People changed their lifestyle and purchasing behaviour as well. Consumer behaviour is more towards the purchasing of essential goods instead of non-essential one. People also realized the significance of healthcare and hygiene care products apart from basic essential needs. This directly affected some of the industrial sectors. Healthcare and pharmaceutical sectors are in huge demand, whereas fashion and beauty industry suffers [1, 2]. Increase in work from home culture, restrictions in social gathering/functions and lack of tourist activities encourage the people to spend less on the fashion and beauty products. This directly impacts the stock prices of the fashion and beauty industries.

Beauty industry covers the production of personal care products and makeup products. Demand of personal care products are intact during COVID-19 scenario.

S. Verma (✉) · S. P. Sahu · T. P. Sahu

Department of Information Technology, National Institute of Technology, Raipur, India

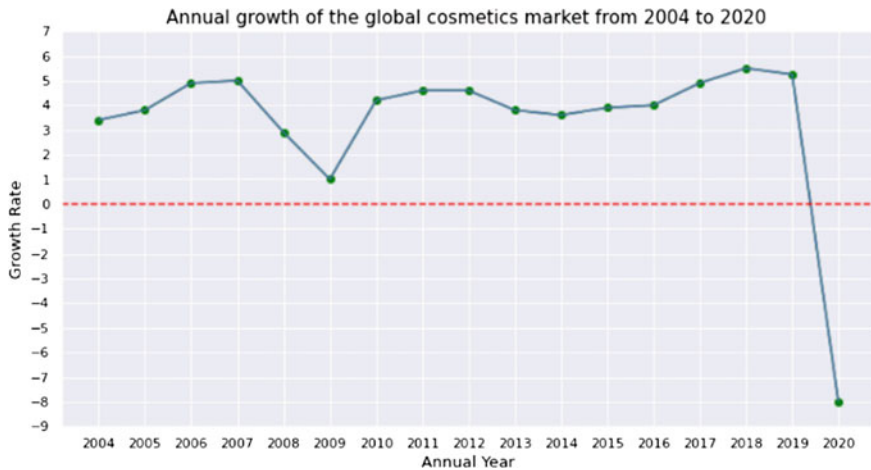


Fig. 1 Annual growth rate of cosmetic market

Although demand in makeup product has been reduced due to the fact that consumer does not find it necessary in present scenario. Consumer purchasing behaviour can be favourable or unfavourable towards any sector, so is for beauty industry [1]. As per the study of [3], there is 20–30% downfall in the sales of beauty products. Figure 1 gives the annual growth rate of the cosmetic market. The data is taken from the website <https://www.statista.com>. We can clearly observe that there is a significant downfall in the year 2020. This kind of consumer behaviour leads to the imbalance in demand and supply. Imbalance in demand and supply will impact the stock market behaviour. There is lot of work done on the financial time series forecasting [4]. Also a lot of studies are based on COVID-19 and its impact on beauty industry [5, 6]. Lot of work is done in literature to predict COVID-19 cases, COVID-19 mortality rate and impact of COVID-19 on social and economic aspects by using machine learning and artificial intelligent-based approaches [7, 8]. Researchers generally have not considered stock market prediction for beauty industry.

This study focuses to analyse the impact of COVID-19 on stock market covering beauty industries. The remaining part of the paper is organized as follows. Section two provides the relevant literature. Section three as methodology of the proposed study, in which stock price prediction and analysis of beauty industry is done with the help of deep learning models. Section four discusses the results. And section five gives the conclusion and future work.

2 Literature Review

Recent development based on machine learning and deep learning techniques have leveraged the performance of the artificial intelligence (AI)-based applications [9]. Stock market forecasting is one such application where deep learning is giving good

performance [10, 11]. Accurate prediction of stock prices is always centre of attraction for the researchers and investors. This section briefs the work done in past which are relevant to stock price/trend movement.

Gua et al. [12] did stock market forecasting by applying adaptive support vector regression (SVR) on stock data. Author did parameter tuning of SVR by using particle swarm optimization (PSO). Author evaluated the model with root mean square error (RMSE), mean absolute percentage error (MAPE) and mean absolute deviation (MAD). Author reported that adaptive SVR gave better performance than the SVR and back propagation neural network (BPNN) for the datasets (SH600006, SH600016, SH600026, SH600036 and SH600056) obtained from Shanghai Stock Exchange. Basak et al. [13] utilized tree-based ensemble approach. For stock trend prediction, random forest (RF) and XGBoost classifiers were applied by the author. Author evaluated the model with accuracy, precision, recall, specificity, F-score, Brier score and area under curve (AUC) for the stock dataset of Apple Inc. and Facebook Inc. The trading window size was chosen as 3, 5, 10, 15, 30, 60 and 90 days, respectively. Author received best accuracy score of 93.02 and 94.76 for Apple and Facebook stocks, respectively, for the window size 90. Cao et al. [14] combined complex network technique with support vector machine (SVM) and K-nearest neighbour (KNN) for trend prediction of stock indices. The dataset of DJIA, S&P500 and NASDAQ indices were considered. Author reported more than 70% prediction accuracy for all three indices.

Nikou et al. [15] did forecasting of iShares MSCI United Kingdom with deep learning model, i.e. long short-term memory (LSTM). Author compared performance of LSTM model with artificial neural network (ANN), SVR and RF. The reported mean absolute error (MAE), mean square error (MSE) and RMSE value for LSTM model by the author is 0.210350, 0.093969 and 0.306543. Chung and Shik Shin [16] proposed multichannel convolutional neural network (CNN) with optimized network. Author optimized the model with genetic algorithm (GA). Author predicted stocks of KOSPI Index and reported prediction accuracy of 73.74%. Pang et al. [17] proposed LSTM with auto-encoder for prediction and analysis of stock market. Author predicted stock prices of the three stocks listed in Shanghai A-share composite index and the index itself. Author reported 57% prediction accuracy for the index prediction.

Štifanić et al. [18] considered COVID-19 confirmed cases along with stock data of commodity (crude oil) and stock indices (DJIA, S&P 500, NASDAQ). Author proposed integrated bi-directional LSTM model and wavelet transform for the prediction purpose. Khattak et al. [19] used LASSO regression for the prediction of European stock indices considering COVID-19 duration. Goh et al. [20] analysed movement of Indonesian stocks with fast Fourier transform (FFT) and linear regression for COVID-19 duration. Ghosh and Chaudhuri [21] did stock market analysis for pre- and post-COVID-19 duration. Author did feature engineering with bootstrapping (FEB) and proposed FEB-stacking and FEB-DNN for prediction of stock market.

In literature, it is identified that analysis of beauty industries during COVID-19 is lacking. Also, as per the work done in literature, it is evident that deep learning

models are good enough for the prediction task. This paper utilized the deep learning model for the analysis of COVID-19 impact on beauty industry.

3 Methodology

In this section, we discuss proposed methodology for the stock market analysis of the beauty industry (Fig. 2).

3.1 Data Collection

Stock market data for leading cosmetic companies (Estée Lauder Companies, Revlon Inc., Coty Inc.) are collected from the website of Yahoo finance. Stock data of three and half years (Dec-2018 to May-2021) are collected. The reason for selection of this particular duration is to have balanced dataset for normal time and pandemic time. At the same time, lockdown scenario is also considered. During Mar-2020 to Jun-2020, many countries in the world imposed lockdown to prevent the spread of the COVID-19. Still some geographical regions are having lockdown as per the pandemic situation. In this research, it is assumed that lockdown started from March, 2020, to have impact analysis on beauty stocks.

3.2 Preprocessing and Feature Engineering

After data collection, missing values are to be dealt before going further. To deal with the missing stock data, mean values are to be considered to replace the missing stock data. Initially, stock data has only six values/features, i.e. high, low, open, close, adj. close and volume. In order to predict better, it is needed to have more number

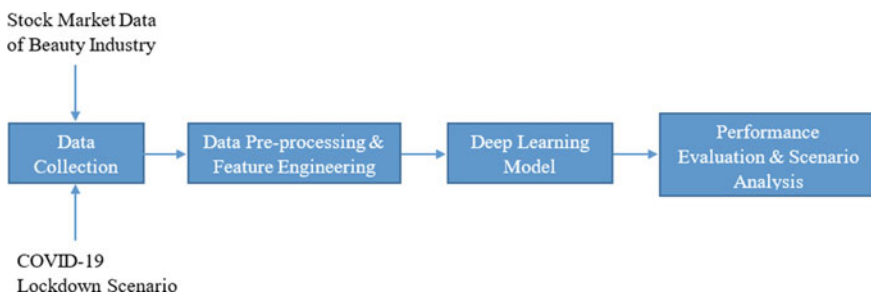


Fig. 2 Process flow

of features. Technical indicators (features) are to be extracted from historical stock data. Extracted technical indicators are Simple Moving Average (SMA), Exponential Moving Average (EMA), Moving Average Convergence Divergence (MACD), Relative Strength Index (RSI), Commodity Channel Index (CCI), On Balance Volume (OBV), Momentum (MOM) and Bollinger Bands (BB). Along with these technical indicators, a new column lockdown is added to analyse COVID-19 impact. After feature extraction, it is required to know the importance of the features and rank them accordingly. This step is needed to select the important features and discard the unimportant ones. XGBoost (eXtreme Gradient Boosting) is considered for feature ranking and therefore feature selection.

3.3 Deep Learning Model

Deep learning is a kind of ANN with multiple processing layers. There are number of deep learning models like deep multilayer perceptron (DMPLP), recurrent neural network (RNN), LSTM, CNN, deep belief network (DBN), auto-encoders (AE) and restricted Boltzmann machine (RBM) [4]. This paper considered LSTM and CNN for the beauty industry’s stock price forecasting. Implementation of LSTM and CNN is done with Keras.

3.3.1 LSTM

LSTM belongs to class of RNN. LSTM model stores information in the memory cells for particular duration. Memory cell is composed with the three gates, i.e. input, output and forget. Gates are used to control the information flow. Sigmoid function is used to trigger the gate. Generalized mathematical equations of the gates are:

$$F_t = \sigma(W_F X_t + U_F H_{t-1} + B_F) \tag{1}$$

$$I_t = \sigma(W_I X_t + U_I H_{t-1} + B_I) \tag{2}$$

$$O_t = \sigma(W_O X_t + U_O H_{t-1} + B_O) \tag{3}$$

$$C_t = F_t * C_{t-1} + I_t * \sigma_C(W_C X_t + U_C H_{t-1} + B_C) \tag{4}$$

$$H_t = O_t * \sigma_H(C_t) \tag{5}$$

where X_t —input vector; F_t, I_t, O_t —activation function of forget gate, input gate; and output gate, respectively; H_t —output vector; C_t —cell state; σ —sigmoid function; σ_C, σ_H —hyperbolic tangent functions; W, U —weight matrix; B —bias.

Bi-directional LSTM is implemented with 1000 epochs, 32 batch size and dropout rate of 0.2.

3.3.2 CNN

CNN belongs to the class of deep neural network (DNN), most widely used in image processing. CNN consists of number of convolutional layers. Convolutional layers are obtained through convolutional operation. CNN layers consists of convolutional layer, max-pooling layer, fully connected MLP and dropout layer.

$$S(t) = (X * w)(t) = \sum_{a=-\infty}^{\infty} X(a)w(t - a) \quad (6)$$

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(m, n)K(i - m, j - n) \quad (7)$$

$$Z_i = \sum_j W_{i,j} X_j + B_i \quad (8)$$

$$Y = \text{softmax}(Z) \quad (9)$$

$$\text{softmax}(Z_i) = \frac{\exp(Z_i)}{\sum_j \exp(Z_j)} \quad (10)$$

where t —time; S —feature map; w —kernel; X —input; I —input image; K —kernel; (m, n) —dimensions; a, i, j —variables; W —weight, B —bias, Z —output of a neuron; Y —output.

Equations 6 and 7 give the convolution operation for 1 and 2 dimension. Equation 8 gives the architecture of ANN. Softmax function is used to get the output through Eqs. 9 and 10. CNN is implemented with 500 epochs, 128 batch size and dropout rate of 0.2.

3.4 Performance Evaluation

Mean absolute percentage error (MAPE), mean absolute error (MAE) and root mean square error (RMSE) are considered for performance evaluation.

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (\text{Actual}_t - \text{Predicted}_t)^2} \tag{11}$$

$$MAPE = \frac{1}{n} \sum_{t=1}^n \frac{|\text{Actual}_t - \text{Predicted}_t|}{|\text{Actual}_t|} \times 100 \tag{12}$$

$$MAE = \frac{1}{n} \sum_{t=1}^n |\text{Actual}_t - \text{Predicted}_t| \tag{13}$$

where Actual_t —Actual price on t th day, Predicted_t —predicted price on t th day and n —number of data values.

4 Results

The proposed work is implemented in Python. Considered dataset of the three beauty companies are visualized in Figs. 3, 4 and 5. As per the analysis, it is found that all companies faced downfall in the first quarter of 2020 at the time of COVID-19 first wave. During first quarter of 2021, there is a huge leap in closing price. Also companies encountered a spike in the volume of stocks traded in the first quarter of 2021 after the COVID-19 era (lockdown) has started. Daily return also showed uncertainty during second quarter of 2020 due to peak of the COVID-19 all over the world. Even though the beauty industry has been affected due to COVID-19 pandemic, the companies started to recover since first quarter of 2021 as people

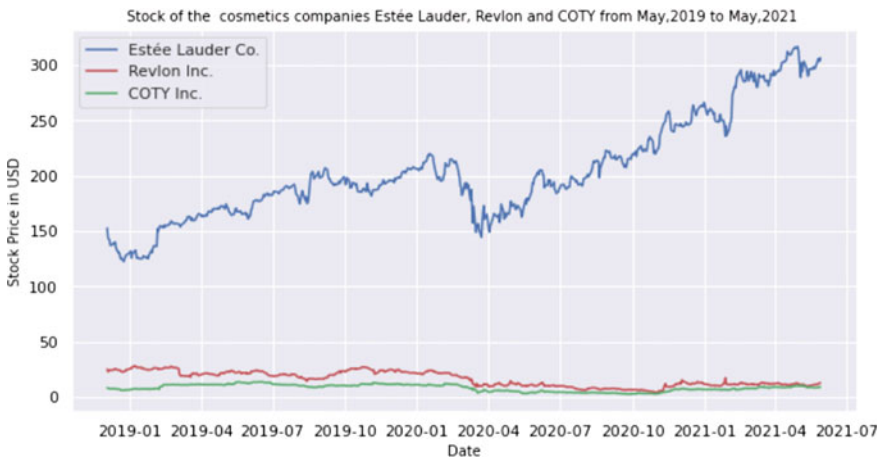


Fig. 3 Close price trend of EL, REV and COTY during 2019–2021

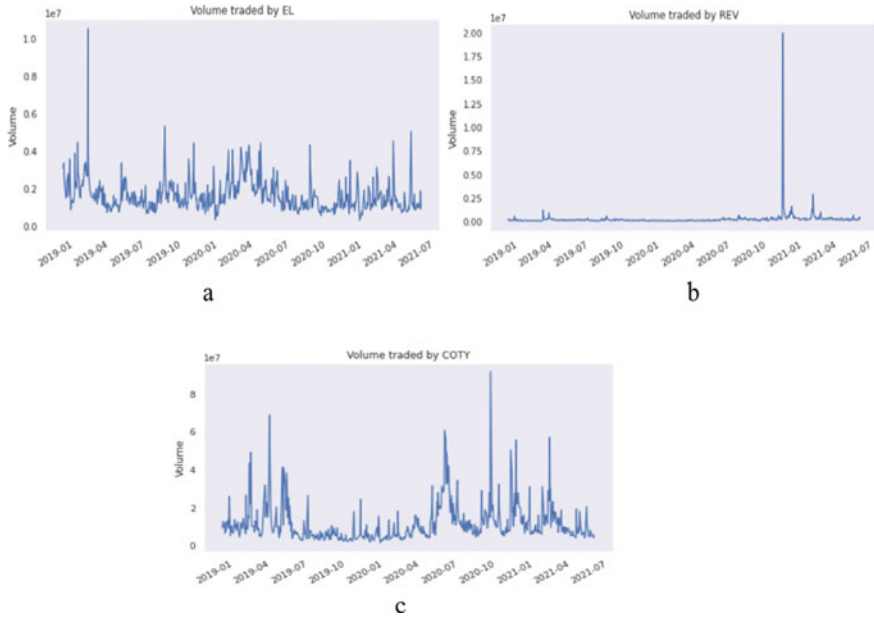


Fig. 4 Volume traded of EL (a), REV (b) and COTY (c) during 2019–2021

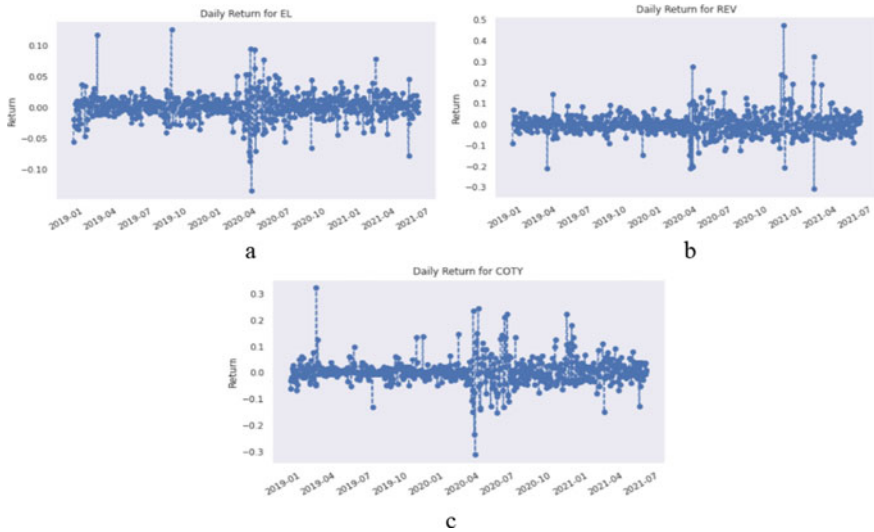


Fig. 5 Daily stock return of EL (a), REV (b) and COTY (c) during 2019–2021

Table 1 Performance evaluation of the proposed approach

Deep learning model	Stock data	MAPE	MAE	RMSE
LSTM	EL	17.19919	0.08673	0.07840
	REV	13.62354	0.02439	0.03729
	COTY	24.21362	0.04136	0.04974
CNN	EL	25.66412	0.13449	0.15757
	REV	19.29612	0.09684	0.13099
	COTY	18.61184	0.07285	0.09323

The bold values signify the best prediction outcome that is obtained for the company REV through the LSTM model

started taking care of their health and wellness. In addition, during pandemic, the online retail purchases have been increased due to the difficulty faced by the people to shop physically in retail stores. It seems from the data pattern that the industry will be reinstated.

Considering the analytical part of the dataset in mind, it is attempted to find out whether deep learning models are able to do the accurate prediction during COVID-19 era.

The prediction results are given Table 1 and visualized in Fig. 6. While comparing the two models (LSTM and CNN), it is clear that for our dataset, LSTM provides the better prediction results except MAPE for COTY. The graph plot for LSTM model is only given here, since it provides the good results in comparison to CNN.

5 Conclusion

This paper analysed the stock market scenario during COVID-19 for beauty industries. It is found that market of beauty industry suffered in 2020 mainly because of lockdown situation. Stock prices of such beauty companies have also suffered. Deep learning (LSTM and CNN) models are applied to predict the stock market prices of leading beauty industries under COVID-19 scenario. Both the model gives good results; however, LSTM outperformed in all the three cases. LSTM provides best prediction result for REV with 13.62354, 0.02439 and 0.03729 as MAPE, MAE and RMSE, respectively. CNN provides best prediction for COTY with 18.61184, 0.07285 and 0.09323 as MAPE, MAE and RMSE. There is a significant increase in stock prices of Estée Lauder Co. during 2021 in comparison with the other two companies. This work can be extended with the parameter optimization of the deep learning models.

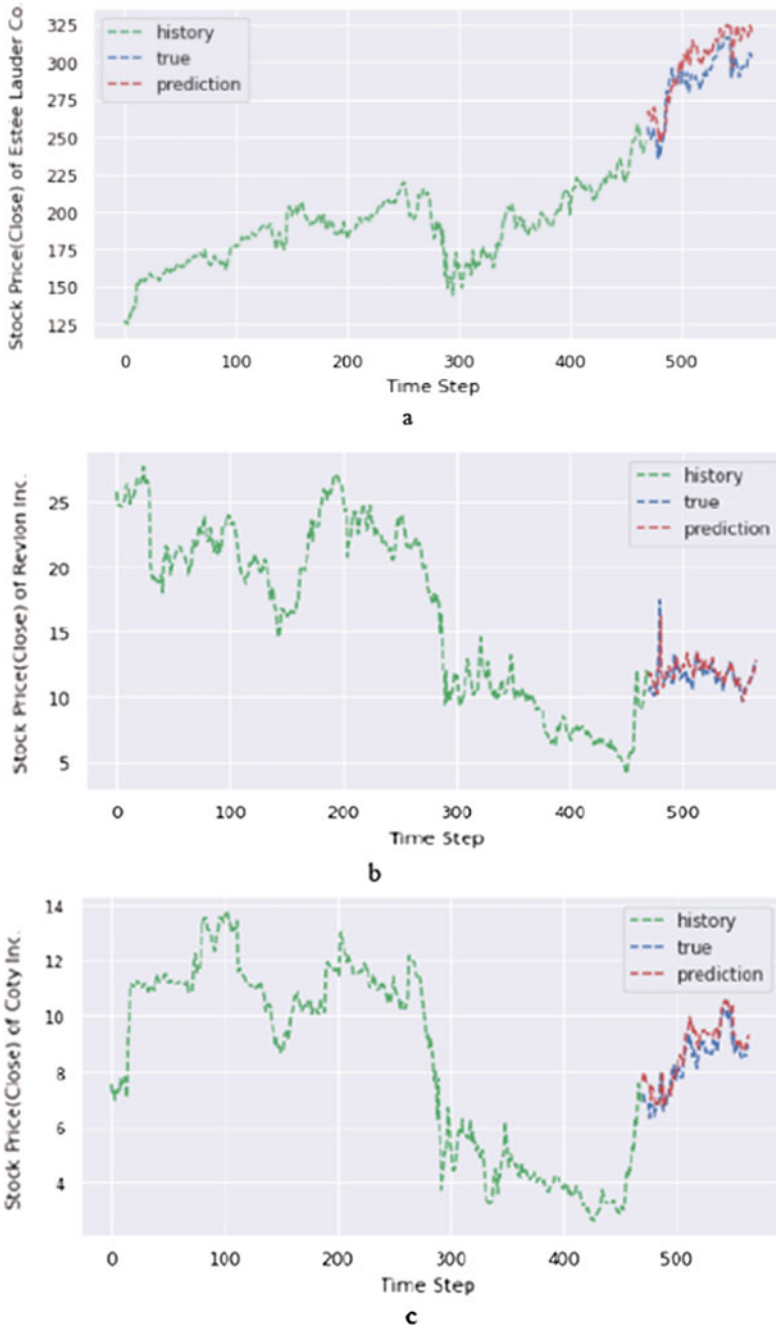


Fig. 6 LSTM prediction graph for Estée Lauder Co (a), Revlon (b) and Coty Inc. (c)

References

1. A. Sharma, D.M. Mehta, Effect of covid-19 consumer buying behaviour towards cosmetics: study based on working females pjaee. *Palarch's J. Archaeology Egypt/Egyptol.* **17**, (9), 5155–5175 (2020), Available: <https://www.archives.palarch.nl/index.php/jae/article/view/4802>
2. S. Akter, Changes in consumer purchasing behavior due to COVID- 19 pandemic. *J. Mark. Consum. Res.* (2021). <https://doi.org/10.7176/jmcr/77-04>
3. E. Gerstell, S. Marchessou, J. Schmidt, E. Spagnuolo, How COVID-19 is changing the world of beauty. *McKinsey Company Consum. Packag. Goods Pract.* **1**(May), 1–8 (2020)
4. O.B. Sezer, M.U. Gudelek, A.M. Ozbayoglu, Financial time series forecasting with deep learning: a systematic literature review: 2005–2019. *Appl. Soft Comput. J.* **90**, 106181 (2020). <https://doi.org/10.1016/j.asoc.2020.106181>
5. T.D. Pikoos, S. Buzwell, G. Sharp, S.L. Rossell, The COVID-19 pandemic: psychological and behavioral responses to the shutdown of the beauty industry. *Int. J. Eat. Disord.* **53**(12), 1993–2002 (2020). <https://doi.org/10.1002/eat.23385>
6. P. Shekam, L. Singh, V.K. Dixit, Impact of Covid-19 on personal care service industry. *J. Crit. Rev.* **7**(15), 3932–3939 (2020)
7. M. Singh, S. Dalmia, Prediction of number of fatalities due to Covid-19 using machine learning. in *2020 IEEE 17th India Council International Conference INDICON 2020* (2020). <https://doi.org/10.1109/INDICON49873.2020.9342390>
8. V. Kumar, D. Singh, M. Kaur, R. Damaševičius, Overview of current state of research on the application of artificial intelligence techniques for COVID-19. *PeerJ. Comput. Sci.* **7**, e564 (2021). <https://doi.org/10.7717/peerj-cs.564>
9. K.A. Shi Dong, P. Wang, A survey on deep leaning architectures and its applications. *Comput. Sci. Rev.* **40** (2021). <https://doi.org/10.1016/j.cosrev.2021.100379>
10. R. Singh, S. Srivastava, Stock prediction using deep learning. *Multimed. Tools Appl.* **76**(18), 18569–18584 (2017). <https://doi.org/10.1007/s11042-016-4159-7>
11. M. Nabipour, P. Nayyeri, H. Jabani, A. Mosavi, E. Salwana, S. Shahab, Deep learning for stock market prediction. *Entropy* **22**(8) (2020). <https://doi.org/10.3390/E22080840>
12. Y. Guo, S. Han, C. Shen, Y. Li, X. Yin, Y. Bai, An adaptive SVR for high-frequency stock price forecasting. *IEEE Access* **6**, 11397–11404 (2018). <https://doi.org/10.1109/ACCESS.2018.2806180>
13. S. Basak, S. Kar, S. Saha, L. Khaidem, S.R. Dey, Predicting the direction of stock market prices using tree-based classifiers. *North Am. J. Econ. Financ.* **47**(June), 552–567 (2019). <https://doi.org/10.1016/j.najef.2018.06.013>
14. H. Cao, T. Lin, Y. Li, H. Zhang, Stock price pattern prediction based on complex network and machine learning. *Complexity* **2019** (2019). <https://doi.org/10.1155/2019/4132485>
15. M. Nikou, G. Mansourfar, J. Bagherzadeh, Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms. *Intell. Syst. Accounting, Financ. Manag.* **26**(4), 164–174 (2019). <https://doi.org/10.1002/isaf.1459>
16. H. Chung, K. Shik Shin, Genetic algorithm-optimized multi-channel convolutional neural network for stock market prediction. *Neural Comput. Appl.* **32**(12), 7897–7914 (2020). <https://doi.org/10.1007/s00521-019-04236-3>
17. X. Pang, Y. Zhou, P. Wang, W. Lin, V. Chang, An innovative neural network approach for stock market prediction. *J. Supercomput.* **76**(3), 2098–2118 (2020). <https://doi.org/10.1007/s11227-017-2228-y>
18. D. Štifić, J. Musulin, A. Miočević, S. Baressi Šegota, R. Šubić, Z. Car, Impact of COVID-19 on forecasting stock prices: an integration of stationary wavelet transform and bidirectional long short-term memory. *Complexity* **2020** (2020). <https://doi.org/10.1155/2020/1846926>
19. M.A. Khattak, M. Ali, S.A.R. Rizvi, Predicting the European stock market during COVID-19: a machine learning approach. *MethodsX* **8**(December 2020), 101198 (2021). <https://doi.org/10.1016/j.mex.2020.101198>

20. T.S. Goh, H. Henry, A. Albert, Determinants and prediction of the stock market during COVID-19: evidence from Indonesia. *J. Asian Financ. Econ. Bus.* **8**(1), 001–006 (2021). <https://doi.org/10.13106/jafeb.2021.vol8.no1.001>
21. I. Ghosh, T.D. Chaudhuri, Feb-stacking and feb-dnn models for stock trend prediction: a performance analysis for pre and post covid-19 periods. *Decis. Mak. Appl. Manag. Eng.* **4**(1), 51–84 (2021). <https://doi.org/10.31181/dmame2104051g>

An OWA-Based Feature Extraction and Ranking for Performance Evaluation of the Players in Cricket



Khalid Anwar, Aasim Zafar, Arshad Iqbal, and Shahab Saquib Sohail

Abstract The popularity of cricket match has increased in recent years across many nations. With the new tools and technologies taking place in every sphere of life, the importance of mathematical model to predict various aspects in a match has been in demand and a topic of relevance for the researchers. The new coming cricket teams from different region of the globe face a challenge of selecting best players. Players are selected by assessing their performance which involves a lot of subjectivity. In this work, we have proposed a fuzzy aggregation approach for assessing the performance. The player's score is the basis for performance assessment and ranking. The results have been shown that suggests that the proposed approach reduces the subjectivity by considering different parameters to calculate the player score. It is envisaged that the proposed approach can serve as benchmark for leading cricket teams as well as new coming teams in selecting best XI for them especially in a tournament environment. Moreover, to the best of our knowledge, we are first to use fuzzy aggregation for suggesting best players in cricket who are currently playing and actively engaged in the match.

Keywords Cricket · Ranking · Performance evaluation · OWA · Fuzzy aggregation

1 Introduction

Cricket is one of the most popular sport of modern era. It was started by the people of Great Britain, and they propagated it to other parts of the world. The original form of cricket is test cricket which is played for five days with an average of ninety overs

K. Anwar (✉) · A. Zafar

Department of Computer Science, Aligarh Muslim University, Aligarh, India

A. Iqbal

KA Nizami Centre for Quranic Studies, Aligarh Muslim University, Aligarh, India

S. S. Sohail

Department of Computer Science and Engineering, School of Engineering Sciences and Technology, Jamia Hamdard, New Delhi 110062, India

bowled per day, and each team gets two innings to bat. It is slow and consumes more time. To cope with the needs of fast moving world, one day cricket (ODI) was started, and in the early twenty-first century, T-20 cricket has evolved. Commercialization of sports in late twentieth century has opened several paths for business and employment. The young generation of twenty-first century is enthusiastic toward cricket as profession and to earn for their life. This has not only increased the popularity of cricket but also increased the competition. This kind of popularity and competition has created new challenges of selecting best out of many talented cricketers for the cricket administrators and selectors. They try to assess the performance of players and accordingly rank them to make the selection procedure easier. However, assessing the performance and ranking of players is a subjective task due to variation in playing conditions, strength of opposition, etc.

A deep study of the concerned field indicates that many academicians and scholars have proposed solutions to minimize the subjectivity involved in the performance assessment and ranking of cricket players. Rohde et al. [1] have applied the idea of opportunity cost and supernormal profit for economic ranking of all batsman in test match cricket. They have considered only two batting features, i.e., runs scored in career and batting average to calculate the score and rank of batsman. In another paper [2], the authors have developed a criteria for comparing the batting performance and have used it to select best XI batsman of 2003 World Cup. The features used in the calculation are batting average and batting strike rate. Singh et al. [3] have developed a system using fuzzy logic to assess the performance of batsman. They have considered batting strike rate, number of fours and sixes, team strength, and opposition strength as the parameters to assess the performance. Their main focus is to analyze the performance of batsman in a particular match. In another work, Ahmed et al. [4] have used multi-objective optimization approach for team selection for Indian Premier League (IPL). They have used batting average as the key feature for measuring the batting performance of batsman. The authors of [5] have developed a two-stage technique using regression and OWA method to measure the batting parameters in T-20 cricket. They have investigated and suggested that strike rate is most important feature to assess the batting performance. They have evaluated their proposed work on the dataset of batsman from IPL-2011. Akhtar et al. [6] have proposed another technique of player ranking by analyzing the comparative performance of players in different sessions of a test match by applying the multinomial logistic regression. The technique categorizes the bowlers and batsman into different classes and gives a mixed ranking of players by analyzing their performance in different sessions of a test match. The calculation process is complex, and it may have some biasness. For example, if a test match is being played on a flat batting-friendly pitch, then the batsman will get a better rank using this technique. Similarly, if a match is played on a green fast bowling favoring condition, then fast bowlers will be on top of the ranking chart, similarly Asian spin-friendly conditions will favor the spinners in the ranking process. In another recent work, Jayanth et al. [7] have used features like batting average, batting strike rate, milestone reaching ability, aggressiveness, and out rate to calculate the ranking index of the batsman. They have also calculated the rank of bowlers. By utilizing the ranking of batsman and bowlers, they have recommended

the team which may win the game and predicted the match outcome. Another recent work by Prem Kumar et al. [8] have applied factor analysis method to rank batsman and bowlers in ODI cricket. They have tried to remove subjectivity by calculating the batsman score based on certain features and then accordingly assign a rank to the batsman. The limitation with this work is that it does not consider the experience of a batsman in calculating the rank. If in any particular session or series, a new batsman played more matches due to various reasons like team of experience batsman played less matches or experience batsman took rest due to heavy work load of various format or get injured. Then, the less experienced batsman will get better rank even if he has less caliber of winning the matches as compared to the experienced batsman.

In this paper, we have used fuzzy aggregation operator to assess the performance of batsman and to rank them. We have reduced the subjectivity involved in assessing the value of player by considering important batting features and systematically preprocessing the data, assigning weights to features, and calculating player score. We hope that proposed approach will cater the need of cricket administrators, selectors, and sponsors in assessing the performance and ranking of batsman in different forms of cricket. The proposed approach may also be used to assess the bowling performance by considering the bowling parameters.

The rest of this paper is organized as follows. Section 2 is about OWA, its application, weight assignment, and score calculation followed by Sect. 3 which explains our proposed approach. Section 4, discusses the data collection, preprocessing, and feature selection approach in detail. Section 5 is devoted to result and discussion, followed by performance evaluation in Sect. 6 and conclusion in Sect. 7.

2 OWA

OWA is a well-known multi-criteria decision-making operator which is based on fuzzy. It was introduced by R. Yager in the year 1988 to combat the problem of uncertainty in decision making [9]. OWA is finding its use in various real-life problems. It has been used in recommender systems for opinion mining and collaborative filtering [10], for recommendation of books [11]. It also finds its use in GIS applications [12], in urban development [13], multi-sensor data fusion [14], ranking of football players [15], and recommending books for university students [16, 17]. Recently, it has been applied for queue modeling in healthcare [18].

The ordered weighted averaging (OWA) can be defined as the mapping of n dimension associated over a vector W of n weights.

OWA: $R^n \rightarrow R$, where

$$W = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ \vdots \\ w_n \end{pmatrix}$$

And satisfies the conditions,

- (a) $w_i \in [0,1]$
- (b) $\sum w_i = 1$

Mathematically, OWA is written as:

$$OWA(c_1, c_2, c_3 \dots c_n) = \sum w_i d_i$$

where d_i is obtained by arranging the sequence of criteria c in descending order. We have used the equation to calculate the weights ‘ W_k ’ for OWA operator.

$$W_k = \{Q(k/n) - Q((k - 1)/n)\},$$

where $k = 1, 2 \dots n$.

Function $Q(r)$ for relative quantifier can be calculated by Eq. (1) as:

$$Q(r) = \begin{cases} 0 & \text{if } r < a \\ \frac{r-a}{b-a} & \text{if } a \leq r \leq b \\ 1 & \text{if } r > b \end{cases} \tag{1}$$

where $Q(0) = 0, \exists r \in [0, 1]$ such that $Q(r) = 1$, and a, b and $r \in [0, 1]$.

Example: If number of criteria $n = 5$ and parameters $a = 0.3$ and $b = 0.8$, then the corresponding weight vector

$$W = \begin{pmatrix} 0.0 \\ 0.2 \\ 0.4 \\ 0.4 \\ 0.0 \end{pmatrix}$$

And OWA score can be calculated as

$$\begin{aligned} &OWA(0.97 \ 0.89 \ 0.76 \ 0.83 \ 0.71) \\ &= 0.0 * 0.97 + 0.2 * 0.89 + 0.4 * 0.83 + 0.4 * 0.76 + 0.0 * 0.71 = 0.746 \end{aligned}$$

A very important aspect of OWA is reordering and sorting of criteria into descending order. It shows that criteria c is not associated with any weight, but weights are associated with specific order of criteria.

Similarly, for parameters $a = 0$ and $b = 0.5$, the corresponding weights vector and OWA score will be

$$W = \begin{pmatrix} 0.2 \\ 0.4 \\ 0.4 \\ 0.0 \\ 0.0 \end{pmatrix}$$

$$\begin{aligned} & \text{OWA}(0.97 \ 0.89 \ 0.76 \ 0.83 \ 0.71) \\ & = 0.2 * 0.97 + 0.4 * 0.89 + 0.4 * 0.83 + 0.0 * 0.76 + 0.0 * 0.71 = 0.81 \end{aligned}$$

And for $a = 0.5$, $b = 1$ the weights vector and OWA score will be

$$W = \begin{pmatrix} 0.0 \\ 0.0 \\ 0.2 \\ 0.4 \\ 0.4 \end{pmatrix}$$

$$\begin{aligned} & \text{OWA}(0.97 \ 0.89 \ 0.76 \ 0.83 \ 0.71) \\ & = 0.0 * 0.97 + 0.0 * 0.89 + 0.2 * 0.83 + 0.4 * 0.76 + 0.4 * 0.71 = 0.91 \end{aligned}$$

3 Proposed Approach

Figure 1 represents our proposed approach of ranking players for assessing their batting performance. The proposed approach consist of various steps like (i) data collection, (ii) data preprocessing, (iii) feature selection, (iv) player score calculation and ranking, and (v) performance evaluation by considering expert opinion about different rank list. In data collection phase, we have extracted the data of active ODI batsman from World Wide Web and international cricket council (ICC) Web site. In preprocessing phase, we converted the collected data to csv file format cleaned it from noise and normalized it to fit into the model. There are various features which can be considered for ranking of batsman in ODI cricket. We selected five important feature by studying their effect on batting performance. A comprehensive explanation of feature selection approach and reason for inclusion and exclusion of a feature in discussed in Sect 4. The selected features were considered for player score

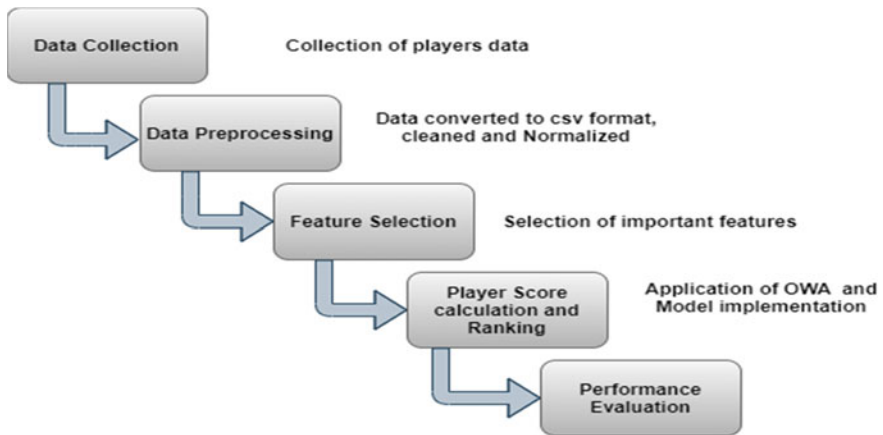


Fig. 1 Proposed model of ranking for assessing batting performance in cricket

calculation using OWA and ranking of the batsman. In the final step, we evaluated our proposed model by taking experts opinion about different rank lists and found out that our approach is most favored one.

4 Data Collection and Feature Selection

To assess the performance of players, we extracted the list of top 100 International Cricket Council (ICC) ODI batsman from ICC Web site www.icc-cricket.com on February 12, 2021. Then, we extracted the data of these 100 players from www.espncricinfo.com. We converted the data file to csv format, cleaned, and normalized the data using the Google Colab Notebook. The extracted data has various features depicting the statistics of batting performance and value of the players. The feature available for accessing the batting performance and value of a player is number of matches played, batting average, batting strike rate, number of runs scored in career, number of hundred scored, number of fifties scored, number of sixes and fours hit, highest score, numbers of inning in which the batsman was not out till the end of inning, current points in ICC ranking, etc. The features discussed above have variable correlations with the batting performance of the player and considering all them for performance measuring and ranking will increase the computational complexity of the system. We have analyzed the impact of every batting feature on the performance of the batsman and then choose the most feasible features for calculating the player score. The literature review and data analysis suggest that the batting average represented by equation no. (ii) is the most basic feature which suggest the quality and rank of a batsman. But as there are fixed number of overs bowled in ODI, so fluency of scoring runs is also an import aspect in this. So, some people have considered these two features for assessing the batting performance. The

Table 1 Batting features selected to calculate player score

S. no.	F1	F2	F3	F4	F5
Features	No. of matches	Batting average	Strike rate	Total career runs	Current ICC rating points
Data type	Int	Float	Float	Int	Int

fluency of scoring runs is known as batting strike rate which is represented using equation no. (iii). We have included both these features for assessing the batting performance and some other.

$$\text{Batting Average} = \frac{\text{Total Career runs}}{\text{No. of innings} - \text{No. of not out innings}} \tag{2}$$

$$\text{Batting Strike Rate} = \frac{\text{Total Career runs}}{\text{Total Number of balls Played}} \times 100 \tag{3}$$

Similar to other domains of life, the impact of experience of batsman cannot be ignored while assigning a rank to him. It was the experience of out of form MS Dhoni in 2011 World Cup final that India came out of tough phase and won the match. We have incorporated experience using two variables, i.e., number of matches a player has played and total runs he has scored in his career. ICC rates the players on the very recent performance of player which is also an important feature of evaluating player performance and rank. Therefore, we have incorporated ICC rating as one of the feature is assessing the player performance. We have ignored other features as either they do not have much significance or they have their indirect impact on player performance through some other variable. The number of fours and sixes has been ignored directly, but their impact have been incorporated through strike rate. Also, it is evident that Michael Bevan of Australia was much better batsman than Shahid Afridi of Pakistan. But Afridi was famous for his six hitting abilities and has highest number of sixes in ODI cricket. We have ignored number of not out innings directly, but it has been included indirectly as it is considered while calculating batting average. The features selected for analyzing the batting performance and calculating the rank are given in Table 1.

5 Results and Discussions

In this paper, we have developed a systematic approach for ranking batsman in ODI cricket for assessing their performance. The supreme body which governs the international cricket is international cricket council (ICC). ICC does rank the batsman, bowlers, and all-rounders in different formats of cricket. But ICC does not disclose the method and variables based on which it rank the players. So it may be a subjective process. Many researchers and academicians have proposed their solutions to reduce

Table 2 Rank list using quantifier as many as possible

Rank	Batsman name	Rank list 2 ($a = 0.3, b = 0.8$)
1	Virat Kohli	0.832912
2	Rohit Sharma	0.713745
3	Ross Taylor	0.68262
4	Eoin Morgan	0.62096
5	Martin Guptill	0.605769

the subjectivity involved in performance evaluation and ranking. But their proposal is either based on few batting features or they have evaluated the performance for a very short span of time like for a series or for a calendar year which may not be the proper indication of a player rank. Our objective was to minimize the subjectivity involved in performance assessment and ranking and to design a compressive technique which should incorporate all the important batting features as variables with minimum complexity. To achieve the goal, we have analyzed the batting features and selected five important features. After feature selection and preprocessing, we have applied a fuzzy-based multi-criteria aggregation operator OWA to calculate player score. The OWA can accept three set of quantifiers (i) at least half ($a = 0.5, b = 1$), (ii) at most half ($a = 0, b = 0.5$), and (iii) as many as possible ($a = 0.3, b = 0.8$). We have considered all three parameters to calculate player score of top 100 batsman in ICC ODI rank list. After calculation of player score, the list have been sorted in descending order based on player score and player with highest score is assigned rank 1.

The ranked list of top five batsman along with their player score is presented in Tables 2, 3, and 4. The advantage of this method is that it has reduced the subjectivity of performance assessment by incorporating well defined set of features, this is

Table 3 Rank list using quantifier at most half

Rank	Batsman name	Rank list 3 ($a = 0, b = 0.5$)
1	Virat Kohli	0.967779
2	Rohit Sharma	0.801585
3	Ross Taylor	0.791601
4	Chris Gayle	0.743935
5	Eoin Morgan	0.724921

Table 4 Rank list using quantifier at least half

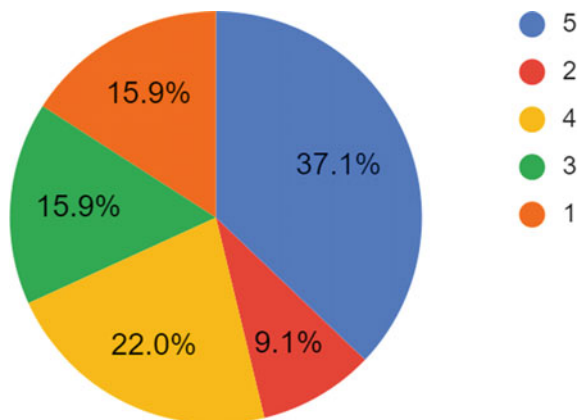
Rank	Name of batsman	Rank list 1 ($a = 0.5, b = 1$)
1	Virat Kohli	1
2	Chris Gayle	0.965291
3	Rohit Sharma	0.895511
4	Ross Taylor	0.888238
5	Babar Azam	0.841419

dynamic in nature as the feature score of features considered in calculation changes with every game. So, rank of player may also change with every match played in international cricket, it has a simple calculation procedure which reduces the complexity of ranking system. The other significance of this work is that besides considering the whole career records, it also considers the recent performance of batsman by incorporating ICC rating as one of the features.

6 Performance Evaluation

Evaluation and comparison are important aspects of scientific research. Any scientific research which is not compared with the existing techniques and do not outperform the existing techniques has no significance. We have compared our ranks with the ICC ODI rank for batsman by taking expert opinion through a Goggle form. We posted four rank lists of top 20 batsman named as rank list 1, rank list 2, rank list 3, and rank list 4, and asked the question “Which of the following ODI Rank list for batsman you consider is best”. We also asked them to show their interest in cricket on scale of 1–5, 1 for minimum and 5 for maximum. Another question was about the profession of the responders. We prorogated the Google form on WhatsApp group of teachers and students of university. A total of 132 people send their responses to us. Most of the responders are university students enrolled in graduation, masters, and Ph.D. programs. Some responders are university teachers and some from other professions also. Most of the people who responded have filled maximum value 5 or 4 for interest in cricket which is shown in Fig. 2. This suggest that these people have a good understanding of cricket. Responder’s choice about various rank lists is shown in Fig. 3 using bar chart. Rank list 1 which is calculated using $a = 0.5$ and $b = 1$ have received maximum votes. Rank list 4 is ICC ODI rank which was obtained in the month of February. This rating by users suggest that our ranking algorithm is better and reliable.

Fig. 2 Responders interest in cricket



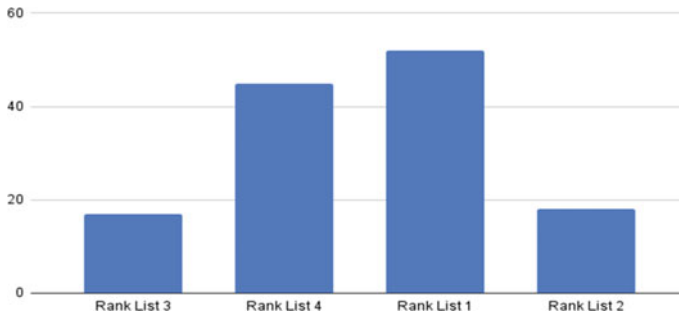


Fig. 3 Experts opinion about rank list

7 Conclusion

In this paper, we have proposed fuzzy aggregation technique to reduce the subjectivity involved in performance assessment and ranking of the batsman in cricket. A systematic approach for feature selection, data preprocessing, player score calculation, and ranking is used to evaluate the performance and rank of active ODI batsman. We have taken public opinion about different rank lists of top 20 ODI batsman and found that our rank list is more favored. Although we have used it to assess the performance and rank of batsman, it can be used to assess the bowling performance and in other multi-criteria decision-making situations. The player score calculated can also be utilized in other decision-making problems in cricket match, for instance, decision making in a match interrupted by rain or bad light or deciding a winner in a tied match.

References

1. N. Rohde, An 'economic' ranking of batters in test cricket*. *Econ. Pap.* **30**(4), 455–465 (2011). <https://doi.org/10.1111/j.1759-3441.2011.00138.x>
2. G.D.I. Barr, B.S. Kantor, A criterion for comparing and selecting batsmen in limited overs cricket. *J. Oper. Res. Soc.* **55**(12), 1266–1274 (2004). <https://doi.org/10.1057/palgrave.jors.2601800>
3. G. Singh, N. Bhatia, S. Singh, Fuzzy logic based cricket player performance evaluator. *IJCA Spec. Issue "Artificial Intell. Tech. Approaches Pract. Appl.* 11–16 (2011)
4. F. Ahmed, K. Deb, A. Jindal, Multi-objective optimization and decision making approaches to cricket team selection. *Appl. Soft Comput. J.* **13**(1), 402–414 (2013). <https://doi.org/10.1016/j.asoc.2012.07.031>
5. G.R. Amin, S.K. Sharma, Measuring batting parameters in cricket: a two-stage regression-OWA method. *Meas. J. Int. Meas. Confed.* **53**, 56–61 (2014). <https://doi.org/10.1016/j.measurement.2014.03.029>
6. S. Akhtar, P. Scarf, Z. Rasool, Rating players in test match cricket. *J. Oper. Res. Soc.* **66**(4), 684–695 (2015). <https://doi.org/10.1057/jors.2014.30>

7. S.B. Jayanth, A. Anthony, G. Abhilasha, N. Shaik, G. Srinivasa, A team recommendation system and outcome prediction for the game of cricket. *J. Sport. Anal.* **4**(4), 263–273 (2018). <https://doi.org/10.3233/jsa-170196>
8. P. Premkumar, J.B. Chakrabarty, S. Chowdhury, Key performance indicators for factor score based ranking in One Day International cricket. *IIMB Manag. Rev.* **32**(1), 85–95 (2020). <https://doi.org/10.1016/j.iimb.2019.07.008>
9. R.R. Yager, On ordered weighted averaging aggregation operators in multicriteria decision-making. *IEEE Trans. Syst. Man Cybern.* **18**(1), 183–190 (1988). <https://doi.org/10.1109/21.87068>
10. S. Saquib, S. Jamshed, S. Rashid, Feature-based opinion mining approach (FOMA) for improved book recommendation. *Arab. J. Sci. Eng.* (2018). <https://doi.org/10.1007/s13369-018-3282-3>
11. S. Saquib, J. Siddiqui, R. Ali, S.S. Sohail, J. Siddiqui, R. Ali, OWA based book recommendation technique. *Procedia Comput. Sci.* **62**(Scse), 126–133 (2015). <https://doi.org/10.1016/j.procs.2015.08.425>
12. J. Malczewski, Ordered weighted averaging with fuzzy quantifiers: GIS-based multicriteria evaluation for land-use suitability analysis. *Int. J. Appl. Earth Obs. Geoinf.* **8**(4), 270–277 (2006). <https://doi.org/10.1016/j.jag.2006.01.003>
13. N. Ghasemkhani, S.S. Vayghan, A. Abdollahi, B. Pradhan, A. Alamri, Urban development modeling using integrated fuzzy systems, ordered weighted averaging (OWA), and geospatial techniques. *Sustain.* **12**(3) (2020). <https://doi.org/10.3390/su12030809>
14. X. Mi, T. Lv, Y. Tian, B. Kang, Multi-sensor data fusion based on soft likelihood functions and OWA aggregation and its application in target recognition system. *ISA Trans.* **112**(xxxx), 137–149 (2021). <https://doi.org/10.1016/j.isatra.2020.12.009>
15. A. Oukil, S.M. Govindaluri, A systematic approach for ranking football players within an integrated DEA-OWA framework. *Manag. Decis. Econ.* **38**(8), 1125–1136 (2017). <https://doi.org/10.1002/mde.2851>
16. S.S. Sohail, J. Siddiqui, R. Ali, An OWA-based ranking approach for university books recommendation. *Int. J. Intell. Syst.* **33**(2), 396–416 (2018). <https://doi.org/10.1002/int.21937>
17. K. Anwar, J. Siddiqui, S.S. Sohail, Machine learning-based book recommender system : a survey and new perspectives. **13**, 231–248 (2020)
18. S. Ahmad, K. Alnowibet, L. Alqasem, J.M. Merigo, M. Zaindin, Generalized OWA operators for uncertain queuing modeling with application in healthcare. *Soft. Comput.* **25**(6), 4951–4962 (2021). <https://doi.org/10.1007/s00500-020-05507-1>

Cardiac Problem Risk Detection Using Fuzzy Logic



T. Sai Vyshnavi, Shruti Prakash, Vyomikaa Basani, and K. Uma Rao

Abstract Currently, the use of computer technology is paving the way for a revolution in the medical field for diagnosis and treatment. With approximately 17 million people dying every year due to cardiac health issues, the need for application of diagnostic tools to assess the risk of cardiac problems has been a huge area of interest. The objective of this paper is to build an assessment tool using fuzzy logic and to assess the risk of a person to a cardiac problem. This is a unique tool that is used to assess the risk of any general cardiac problem. While there are fuzzy logic applications for specific cardiac issues, this tool gives a generalized approach. The input parameters considered are blood pressure, cholesterol, blood sugar level, BMI, heart rate, and smoking. Based on the output of fuzzy logic tool, the risk of cardiac problem is assessed. The proposed system is tested for different values of input parameters to evaluate its performance and the system exhibited satisfactory results. This developed system is simple and efficient to use and can also be used for self-diagnosis.

Keywords Heart diseases · Fuzzification · Rule base · Defuzzification

1 Introduction

In recent times, technology in medical field has been influential on many processes and practices of healthcare professionals. The implementation of deep learning-based algorithms has been advanced in the health sector. There is a lot of imprecision and uncertainty involved in the diagnosis of diseases. The nature and impact of a single disease may vary depending on the patient because different diseases can have common and similar symptoms. Hence, to deal with such imprecision and uncertainty, fuzzy logic is used.

T. Sai Vyshnavi (✉) · S. Prakash · V. Basani · K. Uma Rao
RV College of Engineering, Bangalore, Karnataka, India

K. Uma Rao
e-mail: umaraok@rvce.edu.in

Fuzzy logic resembles the human decision-making methodology, and it deals with vague and imprecise information. It introduces the concept of partial truth according to which the truth values are between “completely true” and “completely false,” i.e., it recognizes more than simple true and false values. It is a logic that is used to designate fuzziness. Fuzzy logic is often misinterpreted as probability; probability is about uncertainty, whereas fuzziness deals with the lack of distinction of an event.

According to the WHO, 17 million people on an average are dying due to cardiac-related diseases every year. A number of risk factors such as high blood pressure, high blood cholesterol, and diabetes increase the chances of developing cardiac problems. Hence, people must take necessary steps to reduce their risk for cardiac problems with simple lifestyle changes and regular cardiac health check-ups to ensure early diagnosis.

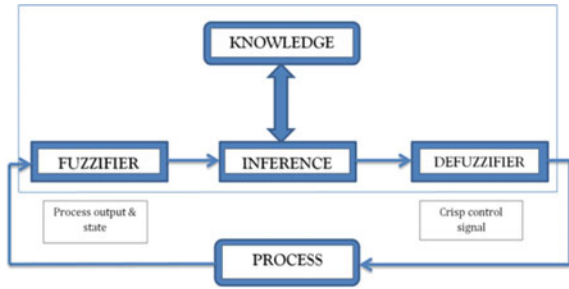
2 Diagnostic Tools for Cardiac Issues

Several research studies are available in the domain of cardiac disease risk prediction using artificial neural network and fuzzy logic system. Kasbe and Pippal [1] developed a fuzzy expert system for the diagnosis of cardiac-related diseases with ten input parameters and one output indicating the increasing heart disease risk. Kowsigan et al. [2] presented diagnosis of heart disease using fuzzy logic and MATLAB with five input parameters, and output is classified into three categories. The paper by Padmavati Kora et al. [3] focuses on the implementation of fuzzy logic-based clinical detection model for coronary risk prevention that consists of seven input parameters and one output classified into three classes. The paper by Kumar and Kaur [4] consists of six input fields, and output field contains integer values from 0 to 1. The paper by Mahdi [5] presents a fuzzy logic-based computer-aided diagnostic (CAD) system that consists of five input variables, and the output was classified in the range from 0 to 4 indicating the danger level. In the study by Chitra and Seenivasagam [6], unsupervised classification system was adopted for heart attack prediction with a total of 13 input attributes. The classifier reports 92% accuracy with the records collected from 270 patients. The study by Senthil Kumar [7] was conducted to diagnose heart patients. He has taken 13 input attributes and categorized output into three classes.

3 Proposed Fuzzy Expert System

The system comprises of seven input parameters and an output denoting the risk level of heart disease. The input parameters are systolic BP, diastolic BP, LDL cholesterol, blood sugar level, BMI, smoking, and heart rate (Fig. 1).

Fig. 1 Interlinking of the components of fuzzy logic system



1. **Fuzzifier:** This module performs fuzzification which is the process of converting crisp input values into fuzzy sets in order to make it compatible with the fuzzy set representation of the input parameters.
2. **Inference engine:** Inference engine performs the computation of the overall value of the output variable based on the individual contribution of each rule in the rule base.
3. **Knowledge base:** This module is very important in the design of a fuzzy logic system. It consists of a data base and a rule base. The database provides information about the input fuzzy sets, the control output, and their membership functions. The rule base consists of a set of rules mapping the input parameters to the output. This module can be constructed either by experts or self-learning algorithm.
4. **Defuzzifier:** This module performs defuzzification which converts the output value in the form of a fuzzy quantity to a crisp value.

Membership function specifies the degree to which a particular input belongs to a fuzzy set. The membership value always lies between 0 and 1. They are used to map the non-fuzzy input values to fuzzy variables and vice versa.

3.1 Fuzzy Expert System Design

The implementation of fuzzy logic-based system begins with fuzzification. Fuzzification converts crisp input values into linguistic fuzzy sets.

Input Variables: Table 1 consists of seven input parameters along with their specific ranges. These parameters are considered as inputs to detect the risk factor for a cardiac disease.

1. **Cholesterol:** LDL cholesterol is a major cause of heart disease. It leads to fatty deposits within arteries which reduces the flow of blood and oxygen to the heart. This can lead to chest pain and heart attack (Fig. 2).
2. **Blood pressure level:** High blood pressure can damage the arteries by making them less elastic, which results in less flow of blood and oxygen to the heart and leads to heart disease (Figs. 3 and 4).

Table 1 Input parameters along with their ranges

Sl. no.	Input variable	Range	Description
1	LDL cholesterol	40–100 100–129 130–159 160–189 190 and above	Optimal Near/above optimal Borderline high High Very high
2	Systolic BP	90–120 120–129 130–139 140–179 Higher than 180	Normal Elevated High BP stage 1 High BP stage 2 Hypertension
3	Diastolic BP	60–80 80–89 90–120 Higher than 120	Normal High BP stage 1 High BP stage 2 Hypertension
4	BMI	19–24.9 25–29.9 30–39.9	Normal Overweight Obese
5	Heart Rate	40–60 60–100 More than 100	Low Normal High
6	Smoking	0–9 10–19 20 and more	Non/light smoker Moderate smoker Heavy smoker
7	Blood sugar level	80–100 100–125 126+	Normal Impaired glucose Diabetic

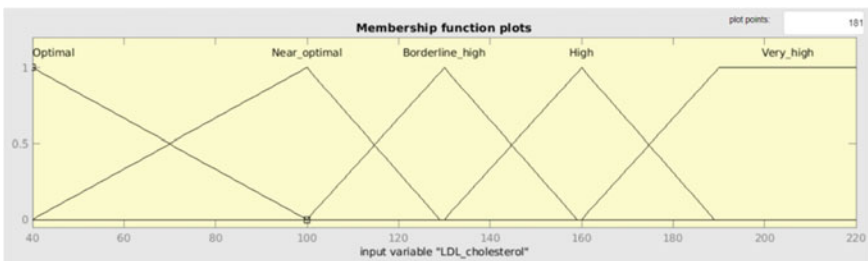


Fig. 2 Membership function for LDL cholesterol

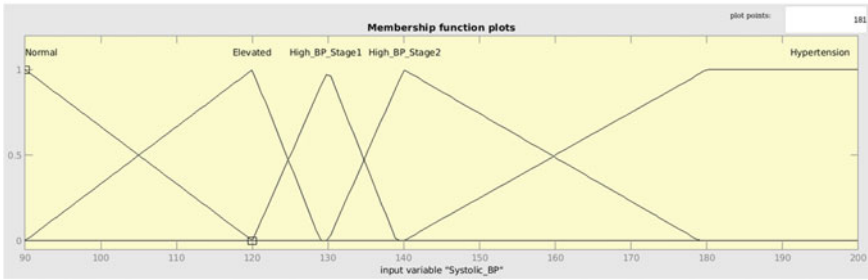


Fig. 3 Membership function for systolic BP

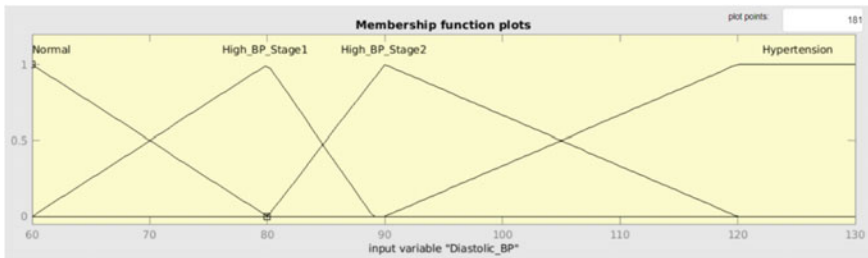


Fig. 4 Membership function for diastolic BP

3. *Body Mass Index (BMI)*: There is a close correlation between heart failure and obesity. In general, the rise of BMI by 1 kg/m² increases the risk of heart failure by 5% in case of men and 7% in case of women (Fig. 5).
4. *Blood sugar level*: A person with diabetes is more likely to develop heart disease because high blood glucose level can damage blood vessels and the nerves that control heart and blood vessels (Fig. 6).
5. *Smoking*: Smoking raises the risk of developing a cardiac disease. Due to cigarette, the blood pumped by the heart to the rest of the body gets contaminated. This can damage the heart and blood vessels and might lead to development of cardiovascular disease (Fig. 7).

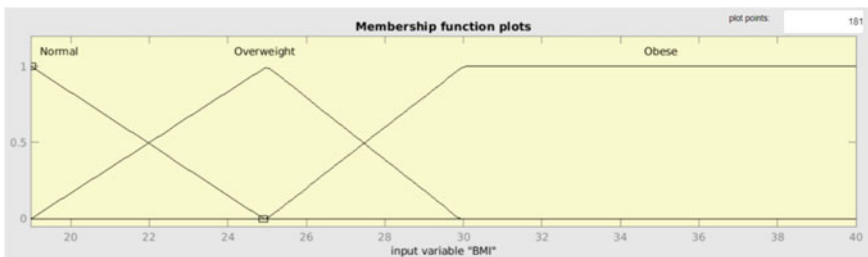


Fig. 5 Membership function for BMI

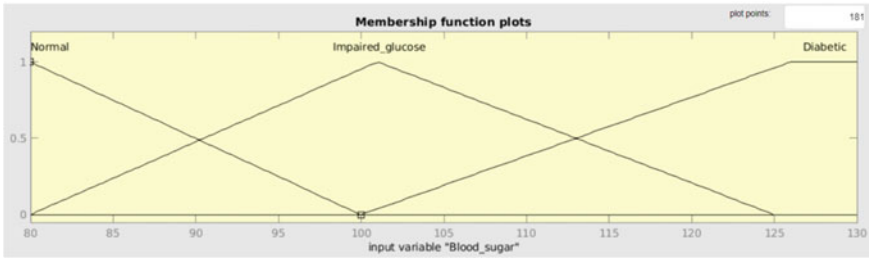


Fig. 6 Membership function for blood sugar level

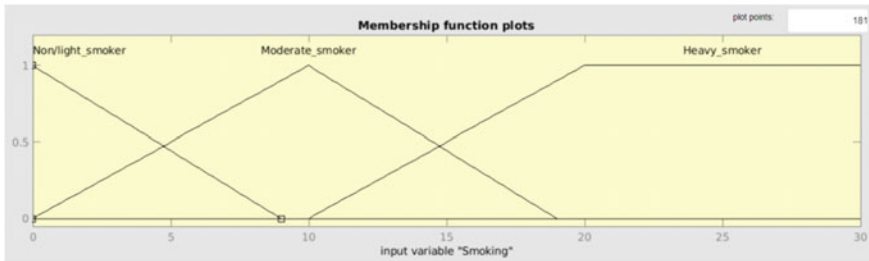


Fig. 7 Membership function for smoking

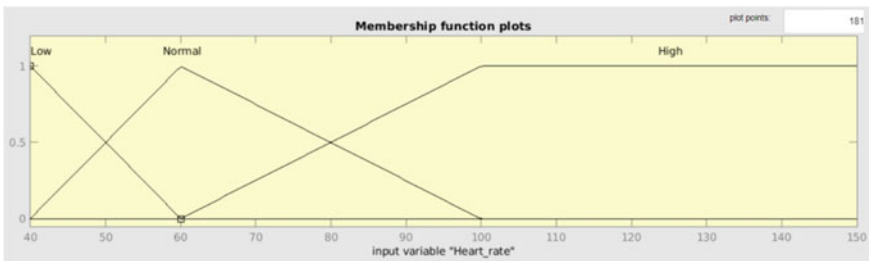


Fig. 8 Membership function for heart rate

- 6. *Heart Rate*: Heart rate is the number of times the heart beats per minute. Rapid or irregular heartbeat is one of the symptoms of heart failure (Fig. 8).

3.2 Rule Base

Rule formation is the main part of the fuzzy inference system since the output depends on the rules that are generated. MATLAB fuzzy logic toolbox consists of rule editor using which rules are entered with the help of logical operators (and/or) provided in

Table 2 Output parameter with its ranges

Output	Range
Healthy	0–0.25
Low risk	0.25–0.5
Medium risk	0.5–0.75
High risk	0.75–1

the toolbox. For the rule base of this project, 232 rules have been generated. The list of some rules is as shown:

1. If BMI is normal, blood sugar is normal, heart rate is normal, LDL cholesterol is optimal, smoking is non/light smoker, systolic BP is normal, and diastolic BP is normal, then output is healthy.
2. If BMI is normal, blood sugar is normal, heart rate is high, LDL cholesterol is optimal, smoking is non/light smoker, systolic BP is normal and diastolic BP is normal, then output is low.
3. If BMI is obese, blood sugar is normal, LDL cholesterol is optimal, smoking is moderate smoker, systolic BP is normal, and diastolic BP is normal, then output is medium.
4. If BMI is overweight, blood sugar is impaired glucose, heart rate is low, LDL cholesterol is optimal, smoking is heavy smoker, systolic BP is normal, and diastolic BP is normal, then output is high.

3.3 Defuzzification

Defuzzification is performed to convert the set of modified control output values into a crisp value. In this project, only one output denoting the risk level of cardiac disease in patients is considered. The membership function of output is divided into four ranges as given in Table 2. The membership function plot of the output is shown in Fig. 9.

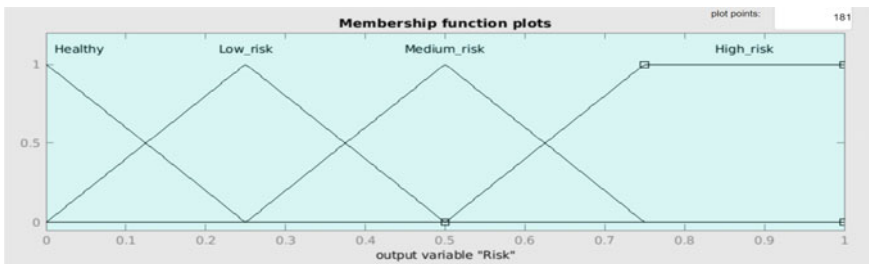


Fig. 9 Membership function for risk (output)

4 Results

This system has seven input parameters and one output parameter. The range of values for each class in the output gives an idea about the risk factor.

There is no standard dataset available for these input parameters in the repository. Hence, the values for input parameters have been taken from 15 laboratory test reports, and some sample results have been displayed below.

- *CASE STUDY 1*: If the output lies in the range 0–0.25, the person is said to be healthy. All the parameters of this patient lie in the normal range. Hence, this person is healthy and is at no risk for cardiac problems (Fig. 10).
- *CASE STUDY 2*: If the output lies in the range 0.25–0.5, the person is said to be at a low risk of having a cardiac issue (Fig. 11).
- *CASE STUDY 3*: If the output lies in the range 0.5–0.75, the person is said to be at medium risk of having a cardiac issue. This patient has BMI which is close to

Fig. 10 Case study with output as “Healthy”

<pre> COMMAND WINDOW Enter BMI: 19 Enter Blood sugar level: 80 Enter Heart rate: 58 Enter LDL Cholesterol: 67 Enter Smoking: 0 Enter Systolic BP: 90 Enter Diastolic BP: 60 </pre>	<p>The output is: 0.21081</p> <p>Healthy</p>
--	--

Fig. 11 Case study with output as “Low risk”

<pre> COMMAND WINDOW Enter BMI: 20 Enter Blood sugar level: 85 Enter Heart rate: 60 Enter LDL Cholesterol: 70 Enter Smoking: 0 Enter Systolic BP: 90 Enter Diastolic BP: 60 >> </pre>	<p>The output is: 0.29773</p> <p>Low risk</p>
---	---

the overweight range, smokes three cigarettes per day, blood sugar level is also close to impaired glucose range, LDL cholesterol is above optimal, and the BP is slightly elevated. Hence, this person is at a medium risk of facing a cardiac issue (Fig. 12).

- *CASE STUDY 4:* If the output lies in the range 0.75–1, the person is said to be at a high risk of having a cardiac issue. This patient falls in the overweight category due to high BMI; LDL cholesterol is borderline high, smokes six cigarettes per day, and has high BP stage 2. Hence, this person is at a high risk of facing a cardiac issue (Fig. 13).

Figure 14 shows the MATLAB rule viewer and the graphical result.

Fig. 12 Case study with output as “Medium risk”

```
COMMAND WINDOW
Enter BMI:
23
Enter Blood sugar level:
96
Enter Heart rate:
72
Enter LDL Cholesterol:
109
Enter Smoking:
3
Enter Systolic BP:
124
Enter Diastolic BP:
80
>> | _____ The output is:
0.67999
Medium risk
```

Fig. 13 Case study with output as “High risk”

```
COMMAND WINDOW
Enter BMI:
33
Enter Blood sugar level:
110
Enter Heart rate:
69
Enter LDL Cholesterol:
142
Enter Smoking:
6
Enter Systolic BP:
141
Enter Diastolic BP:
94
>> | _____ The output is:
0.8337
High risk
```

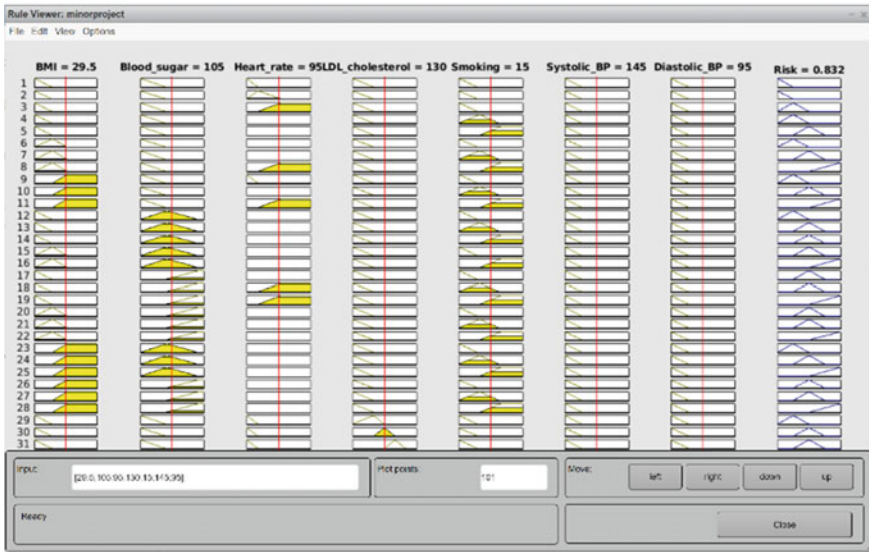


Fig. 14 MATLAB rule viewer

5 Conclusion

Fuzzy expert system designed to detect the risk of cardiac issues is appreciated for its simplicity as it can handle multiple input parameters that may be imprecise. In this project, a fuzzy logic-based assessment tool is developed to assess the cardiac health of a person. This is only a risk assessment tool which can be used to alert the patients and help them make any lifestyle changes based on the risk factor. The tool has been extensively tested for various conditions. The designed tool works for a wide range of input parameters and efficiently assesses the risk.

The fuzzy expert system developed in this project can be extended for additional scope of further diagnosis of cardiac issues and can also be implemented using the neural network approach. It can be developed to detect complicated cardiac issues by taking symptoms as the input parameters. This system can also be implemented for the diagnosis of other diseases and medical conditions such as tuberculosis and cancer.

Acknowledgements We are indebted to Karnataka Council for Technological Upgradation and R.V. College of Engineering® under the R&D Project of “Design and Development of therapeutic exoskeleton for muscular dystrophy and comatose.”

References

1. T. Kasbe, R.S. Pippal, Design of heart disease diagnosis system using fuzzy logic system, in *International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, India* (2017), pp. 3183–3187
2. M. Kowsigan, A. Christy Jebamalar, S. Shobika, R. Roshni, A. Saravanan, Heart disease prediction by analysing various parameters using fuzzy logic. *Pakistan J. Biotechnol.* **14**(2), 157–161 (2017)
3. P. Kora, K. Meenakshi, K. Swaraja, A. Rajani, M. Kafiul Islam, *Detection of Cardiac Arrhythmia Using Fuzzy Logic*, vol.17 (Elsevier, 2019), pp. 1–6
4. S. Kumar, G. Kaur, Detection of heart diseases using fuzzy logic. *Int. J. Eng. Trends Technol.* **4**(6), 2694–2699 (2013)
5. M.S. Mahdi, M.F. Ibrahim, S.M. Mahdi, P. Singam, A.B. Huddin, Fuzzy logic system for diagnosing coronary heart disease. *Int. J. Eng. Technol.* **8**(1.7), 119–125 (2019)
6. R. Chitra, Dr. V. Seenivasagam, Heart attack prediction system using fuzzy C means classifier. *IOSR J. Comput. Eng.* **14**(2), 23–31 (2013)
7. Dr. A.V Senthil Kumar, Diagnosis of heart disease using advanced fuzzy resolution mechanism. *Int. J. Sci. Appl. Inf. Technol.* **2**(2), 22–30 (2013)

Deep Learning Approach for Motor-Imagery Brain States Discrimination Problem



Saptarshi Mazumdar and Rajdeep Chatterjee

Abstract Brain signals can be used to control robotic limbs for partial or fully paralyzed persons. Electroencephalography (EEG) is a widely used noninvasive brain signal recording technique. It is essential to process and understand the hidden patterns associated with a specific cognitive or motor task. Here, the focus is on motor-imagery (MI) EEG signal classification. There is a significant difference between machine learning and deep learning algorithms at the feature extraction phase. In this paper, a discrete wavelet transform-based feature selection followed by one-dimensional (1D) convolutional neural network (ConvNet) approach has been proposed to interpret motor-imagery left-hand and right-hand movements. The proposed model has been compared with the existing SOTA techniques on the same BCI competition II dataset III. It outperforms the traditional machine learning models and achieves 91.43% classification accuracy.

Keywords BCI · ConvNet · Deep learning · DWT · EEG · Motor imagery

1 Introduction

Brain–computer interface (BCI) is an interrelated process of extracting brain signals through electroencephalogram (EEG) and produces a set of data that can be processed using an external interface. EEG is a popular brain signal recording technique due to its portability, cost effectiveness, and high discrete measurement resolution w.r.t time [1]. EEG brain rhythms are microvolt electrical signals generated in the brain due to the activities of neurons. EEG can be used for understanding different types of cognitive or motor activities. Motor imagery (MI) indicates that a subject imagines its limb movement while, in reality, it is not moved physically. Accurate classification of such motor-imagery movements is a challenging task [2]. BCI, along with MI understanding, can help us develop an substitute pathway for communication

S. Mazumdar (✉) · R. Chatterjee
School of Computer Engineering, KIIT Deemed to be University,
Bhubaneswar, 751024 Odisha, India

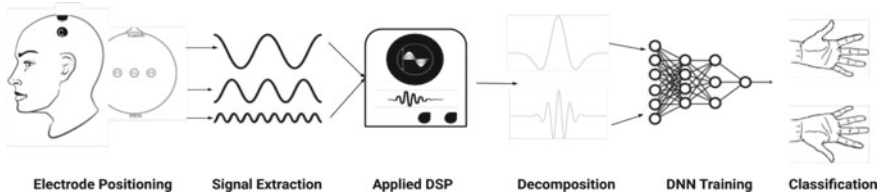


Fig. 1 Complete workflow of hand classification

between the different limbs and the brain for paralyzed persons. For this experiment, BCI competition II dataset III, Graz dataset, has been used. Author had used three bipolar (anterior ‘+’, posterior ‘-’) EEG channels, measured over $C3$, Cz , and $C4$. BCI based on EEG motor imagery [3] is an emerging field of biomedical applications. Different thinking or imagining activities can also be determined and discriminated using motor-imagery classification problem [4].

1.1 Contribution

This paper proposes a new model based on deep learning to classify different brain states using MI EEG signal. The raw EEG input signal has been preprocessed before directly feeding into the discrimination model. The novelty of this work is to develop a lightweight (with less trainable parameters) model for more accurate classification of the left-hand and the right-hand motor-imagery movements. The overall objective of this paper has been described in Fig. 1.

1.2 Organization

The paper has been distributed into six sections. The relevant research works have been discussed in Sect. 2. It is followed by Sect. 3 containing the background concepts used in this paper. The proposed method has been discussed in Sect. 4. The results and analysis are given in Sect. 5. Finally, the paper has been concluded in Sect. 6.

2 Related Study

A good number of relevant and related research articles have been studied. A few widely used methodologies for the statistical analysis of BCI datasets are fast-independent component analysis (Fast ICA) and principal component analysis (PCA). The Fisher discriminant analysis (FDA) and singular value decomposi-

tion (SVD) are also being used for signal decomposition. Authors have used different feature extraction and classifier combinations to achieve high accuracy [5, 6]. In [7], authors have implemented rough set theory for feature selection from the extracted EEG data by computing approximated upper and lower envelop and positive regions [8, 9]. Equivalence relation-based discernibility matrix is an alternative to γ -based reduct calculation. In [10], authors have implemented a feature reduction process based on discernibility matrix in EEG data. In [11], the classical discernibility matrix has been extended to a fuzzified version so that it can deal with real-valued data directly, unlike its predecessors, in [10], which works only with a discretized dataset. Authors have proposed a Morlet wavelet with quadratic Bayesian learner for better EEG signal classification [12]. In [13], authors have used multi-variate empirical mode decomposition (MEMD) and short-time Fourier transform (STFT) for feature extraction and KNN classifier. The obtained results corroborate that the proposed scheme achieves the state-of-the-art performance. Besides the different machine learning approaches, authors have introduced deep learning techniques such as ConvNet and stacked auto-encoder (SAE) to interpret motor-imagery EEG signal classification [14]. However, a comprehensive review is made available for the research community in [15].

2.1 Feature Selection and Extraction

Ignoring the C_z electrode signal helped to increase the variance of two different output classes [16], so while creating classification data, signals from electrode C_3 and C_4 are only taken into consideration. For feature extraction and elimination of unwanted frequencies (noise), elliptic filtration [17, 18] is the best option for motor-imagery signals. Energy entropy and adaptive autoregression-based wavelet transformations help to eliminate the redundancy of a signal, and it is a widely used procedure for their consistent performance [19]. In the comparison of different approaches of feature selection, 4th label Daubechies decomposition helped to achieve the best output [6, 20].

2.2 Classifier Selection

For classification of brain state discrimination problem, among several dimension reduction techniques like singular value decomposition, independent component analysis, Fisher discriminant analysis, etc., principle component analysis successfully stands out [5]. Reducing interference of different components leads to an effective solution. Among different supervised learning algorithms, used for motor imagery like support vector machine [5], ensemble classification [21], Bayes quadratic with spatio-spectral filtration [12], and cosine distance-based KNN [13] helped to achieve a proper differentiation. However, deep learning approaches like

deep, shallow, and hybrid ConvNet architecture with filter bank common spatial patterns (FBCSP) [22], concatenated ConvNet with stacked auto-encoder [14] helped to gain significant improvements in classification. Fast compression residual convolutional neural networks (FCRes-CNNs) [23] have shown a great improvement in the classification process.

3 Background

3.1 Signal Portioning

According to dataset author's description, in each trial, first 3 s are used for instruction, beep, and cue display. Each provided signal in the dataset had a duration of 9 s, from which the first 3 s are ignored for an effective classification process (see Fig. 2).

3.2 Elliptic Filtration

An elliptic filter is an independently adjustable filter with an equal ripple factor in both passband and stopband. It has a faster transition gain for a specific ripple. The gain (G) of an elliptic filter can be represented using Eq. 1,

$$G_n(f) = \frac{1}{\sqrt{\varepsilon^2 * R_n^2 * (\xi, f/f_0)}} \quad (1)$$

where f_0 is the cutoff frequency, ε is denoted as the ripple factor of the system, and ξ is used as a selective factor. R_n is the n th order Chebyshev rational function. If the ripple factor of the stopband can be reduced to zero, the Chebyshev type I IIR filter can be achieved. If the passband ripple factor is reduced to zero, then the Chebyshev type IIR filter can be achieved.

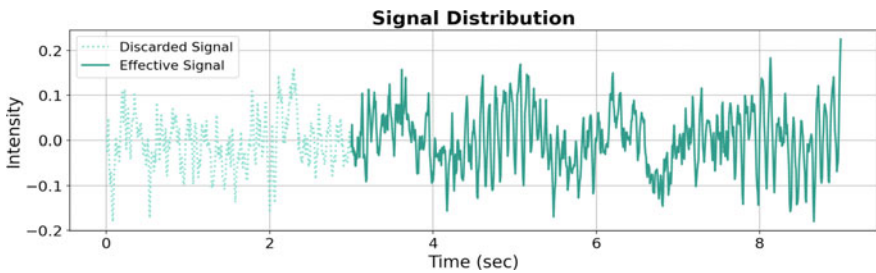


Fig. 2 Signal distribution and effective signal measuring

For reducing the SNR in motor-imagery signals, both hardware setup [24] and algorithmic approaches [25] are widely used for elliptic filtration.

3.3 Discrete Wavelet Transform

Wavelet transform is used to analyze stationary and non-stationary signals using localized functions in Fourier and natural spaces. Discrete wavelet transform (DWT) decomposes a complete signal into a set of differentiable wavelets, orthogonal to its scaling, and translation to remove the redundant set of information [17, 20]. The scaling properties of the wavelet imply mathematical models to its discrete translation like dilation equation (two-scale relation) Eq. 2.

$$\phi(x) = \sum_{n=-\infty}^{\infty} a_n \phi(2x - n) \tag{2}$$

where $\phi(x)$ signifies the scaling function and a_n is the finite set of coefficients.

In this experiment, the fourth-order Daubechies wavelet decomposition has been used. Being discrete, each step of decomposition divided the input into equal half sets of approximated and detailed signals. The fourth-order decomposition prepared an output set of size 96 (i.e., 384/4).

3.4 Alpha Beta Filtration

The main frequency labels of human EEG signal are distributed in the table below:

Band	Freq (Hz)
Delta, δ	$f < 4$
Theta, θ	$4 < f < 7$
Alpha, α	$8 < f < 15$
Beta, β	$16 < f < 31$
Gamma, γ	$31 < f$
Mu, μ	$8 < f < 12$

Since the α and β segments of this dataset are leading to greater classification accuracy, these two signal bands have been used for this purpose. Chebyshev type II (inverse) infinite impulse response filter of order six is used for this purpose for having the most negligible ripple factor in the passband (see Fig. 3). The gain of a Cheby2 filter can be calculated as given in Eq. 3,

$$G_n(f, f_0) = \frac{1}{\sqrt{1 + \frac{1}{\varepsilon^2 T_n^2(f_0/f)}}} \tag{3}$$

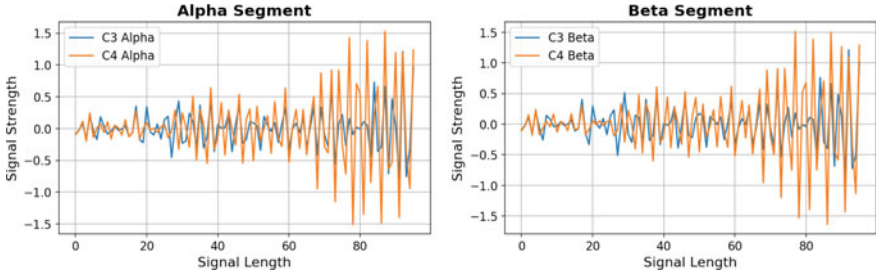


Fig. 3 Sample electrode-specific EEG frequency band distribution for a signal

where f_0 is the cutoff frequency, f is the angular frequency, and T_k^2 is Chebyshev polynomial.

The trigonometric representation of the first kind (T_k) is given in Eq. 4

$$T_k(z) = \begin{cases} \cos(k \arccos(z)), & \text{for } |k| \leq 1 \\ \cosh(k \cosh^{-1}(z)), & \text{for } k \geq 1 \\ (-1)^k \cosh(k \cosh^{-1}(-z)), & \text{for } k \leq -1 \end{cases} \quad (4)$$

Since the Chebyshev polynomial oscillates between -1 and 1 for the stopband, the gain variation range of stopband is given in Eq. 5

$$0 \leq \text{Gain}_n \leq \frac{1}{\sqrt{1 + \frac{1}{\epsilon^2}}} \quad (5)$$

Having less gain in the stopband, the signal–noise ratio is higher, which led to a better clean band-specific signal. Two bands are extracted from each electrode signal.

3.5 Selection and Processing of Electrode Signals

Since $C3$ and Cz electrode signals are almost overlapping, the correlation between these two electrode-specific signals is more significant, which could lead to biased recognition; the Cz signal has been ignored.

Now they are re-arranged. The α segments of both electrodes are kept together, and the same is done for β segments.

4 Proposed Approach

4.1 Convolutional Neural Network

Convolutional neural network (ConvNet) is a deep learning approach to assign biases depending on the differentiable fundamentals of an input. A ConvNet is prepared with fully connected layers to treat nonlinear functionalities along with feature differentiation. The ConvNet structure used in this experiment is shown in Fig. 4.

Network Architecture The proposed ConvNet contains six 1-dimensional convolutional layers, 2 maxpooling layers with pool size 7 each, and 2 dropout layers with 0.2 and 0.1 dropout capability, respectively.

4.2 Proposed Architecture

The proposed architecture uses elliptic filtration for signal-to-noise ration (SNR) reduction, fourth-level Daubechies wavelet decomposition for feature extraction, and deep ConvNet with local maximum pooling, and fully connected hidden layers are used for state classification. Uses of exponential linear unit (ELU) activation function helped to overcome the dying ReLU problem. Hard sigmoid activation is used in the output layer as it is a binary classification. A bird’s eye view of the proposed pipeline has been shown in Fig. 5.

This approach helps to recreate the signal using the significant portions, which helps to create the different subjects more distinguishable. The use of ConvNet generates more features which lead to more accurate classification.

Used Dataset Institute for Biomedical Engineering, University of Technology (Graz), Department of Medical Informatics, has given the BCI competition II dataset III [26]. There is a total of 280 trials (instances) in the used dataset, which has binary class labels (left hand/right hand). Again, it is a balanced dataset containing the same number of total instances in each class. It is recommended to consider the first 140

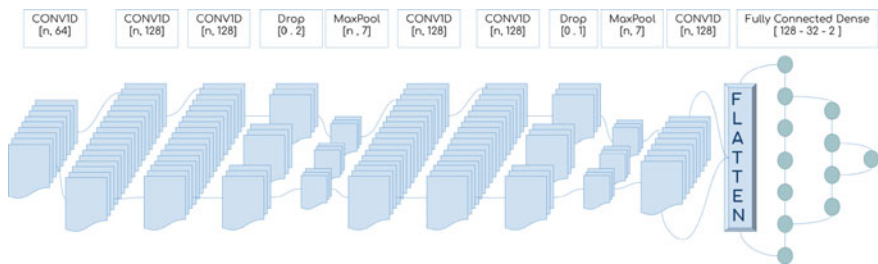


Fig. 4 Architecture of the proposed 1D-ConvNet

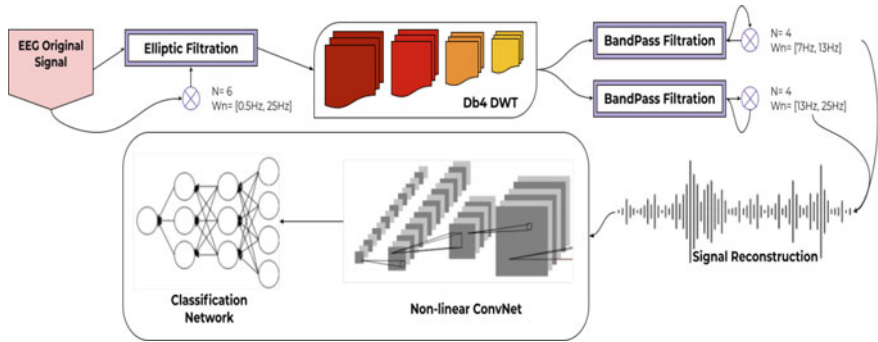


Fig. 5 Proposed system architecture

Table 1 Dataset description: BCI competition II (dataset III)

Electrodes	Sample size	Class label	Training/testing	Cutoff frequency
C3 and C4	768 samples (6 s × 128 Hz)	1—left-hand, 2—right-hand movement	140/140	0.5–50 Hz

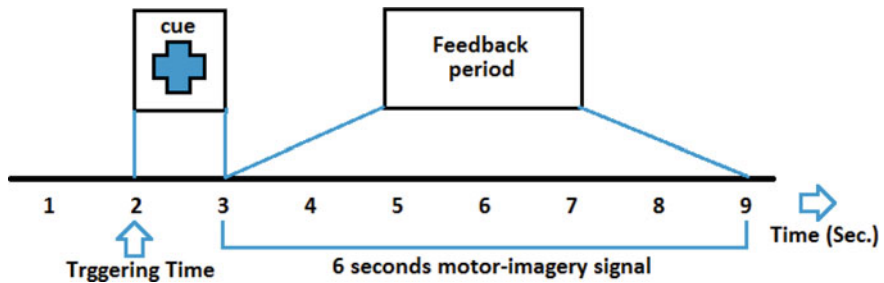


Fig. 6 6 s motor-imagery EEG signal

instances as a training set and the remaining 140 instances as a test set. The decision classes are numeric 1 and 2, which suggest left-hand and right-hand movements, respectively. The dataset contains raw EEG signals for three EEG electrodes C3, Cz, and C4. The aim is to work with the motor-imagery EEG signal in our thesis. The left-hand and right-hand movements are dominantly related to the C3 and C4 electrodes regions of the human scalp (see Table 1). Our interest is given actual motor-imagery EEG signal which starts at 3 s and lasts for another 6 s (total 9 s signal input with 6 s motor-imagery feedback). The IEEE 10–20 electrode placement and the signal description are shown in Fig. 6.

This dataset does not have any significant artifacts. However, the raw EEG signal is filtered with an elliptic bandpass filter using the cutoff frequencies of 0.5 and 50 Hz at a sampling rate of 128 Hz.

5 Result and Analysis

Two experiments have been performed to achieve a robust approach to classification. One follows the guidelines provided by dataset authors, and one manages k-fold cross-validation to avoid any biases [27, 28].

Experiment 1: Hold Out Validation In this experiment, the complete data has been divided into half for the classifier’s training, and the other half is used for validation of the model. Graze dataset had 140 random trials and their levels for training. The training and validation sets are decomposed using the fourth-level Daubechies wavelet transformation and two different bandpass filters to extract alpha and beta-level signals, after which four signal streams could be extracted, each of length 96. The reconstruction of the signal is done using 4 methods: *serial_segments_together*, *serial_electrodes_together*, *parallel_segments_together*, and *strict_parallel_segments_together*.

Among which, *serial_segment_together* reconstruction leads to better classification. After reconstruction, the signal data is fed to the proposed neural net. The defined ConvNet model successfully achieved 100% training accuracy and 91.43% validation accuracy. The performance can be seen in Figs. 7 and 8.

The few best-performing classification techniques used on the BCI competition II dataset III are compared with our proposed methods. The comparative analysis is given in Table 2 and shown in Fig. 9.

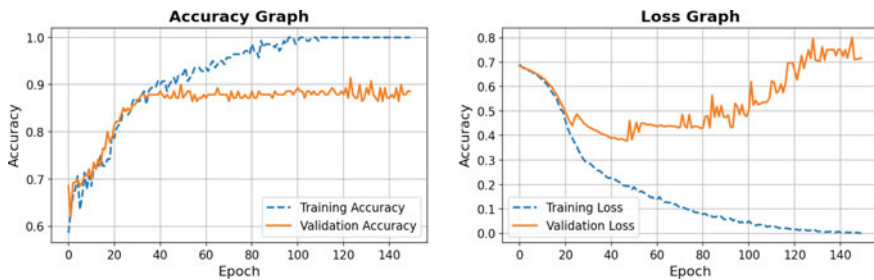


Fig. 7 Accuracy and loss plot for DNN training

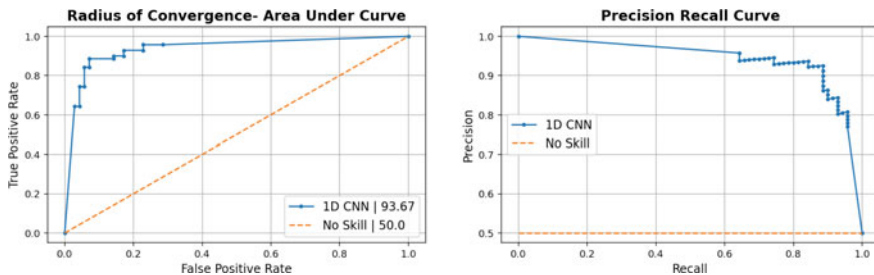
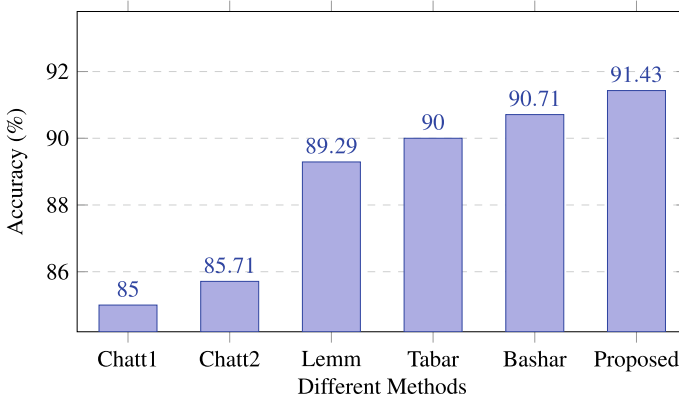


Fig. 8 ROC area under curve and precision recall plot

Table 2 Comparative analysis of a few best-performing techniques on BCI competition II dataset III

References	Methods used	Classifiers	Accuracy (%)
Chatterjee et al. [6] (Chatt1)	Wavelet energy–entropy	SVM (kernel: linear/polynomial)	85.00
Chatterjee et al. [5] (Chatt2)	Average power + band power + wavelet energy–entropy + RMS + statistical features (Table V)	MLP	85.71
Lemm et al. [12]	Morlet wavelet	Bayes quadratic	89.29
Tabar et al. [12]	STFT images	CNN-SAE	90.00
Bashar et al. [13]	MEMD + STFT	KNN (cosine distance)	90.71
Proposed approach	DWT third-level coefficients	1D-ConvNet	91.43

**Fig. 9** Accuracies of a few best-performing techniques on BCI competition II dataset III

Experiment 2: Fivefold Cross-Validation In this experiment, the complete data has been shuffled, and fivefold cross-validation is performed to verify the robustness of the prepared model. For this experiment, the same operations are carried out. The ConvNet model gained $90.35 \pm 4\%$ average accuracy. The results of fivefold can be visualized in Fig. 10.

6 Conclusion

BCI is now a critical research and development area. It has different components. Here, the aim is to develop a deep learning model that can capture the brain signals' information. As the motor-imagery BCI deals with life-critical decisions, the predic-

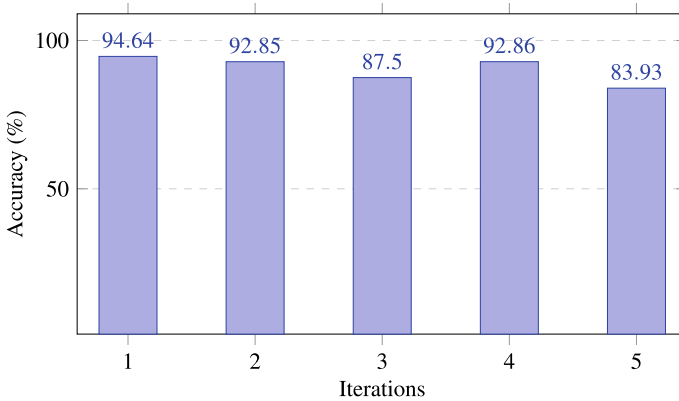


Fig. 10 Fivefold cross-validation accuracy stream

tion accuracy needs to be very high. The proposed 1D-ConvNet deep learning model performs very well on the used dataset to discriminate between left-hand and right-hand movements. Earlier, different machine learning approaches have been explored on the same dataset. However, the proposed 1D-ConvNet models in their first avatar outsmart all traditional machine learning approaches in this paper. The best result obtained from the study is 91.43% which is the best-reported performance for the BCI competition II dataset II to date. Therefore, it can be concluded that the deep learning approaches are suitable to discriminate the EEG brain signals.

The study can be further extended with other BCI datasets and deep learning models. Finally, an end-to-end application needs to be deployed to examine the robustness and scalability of the deep learning models.

References

1. D.J. McFarland, J.R. Wolpaw, EEG-based brain-computer interfaces. *Curr. Opin. Biomed. Eng.* **4**, 194–200 (2017)
2. J. Kalcher, C.D. Flotzinger, S. Göllly, N.G. Pfurtscheller, Graz brain-computer interface II: towards communication between humans and computers based on online classification of three different EEG patterns. *Med. Biol. Eng. Comput.* **34**(5), 382–388 (1996)
3. C.S. Nam, A. Nijholt, F. Lotte, *Brain-Computer Interfaces Handbook: Technological and Theoretical advances* (CRC Press, Oxford, UK, 2018)
4. K.J. Miller, G. Schalk, E.E. Fetz, M. den Nijs, J.G. Ojemann, R.P.N. Rao, Cortical activity during motor execution, motor imagery, and imagery-based online feedback. *Proc. Natl. Acad. Sci.* **107**(9), 4430–4435 (2010)
5. R. Chatterjee, T. Bandyopadhyay, EEG based Motor Imagery Classification using SVM and MLP, in *2016 2nd International Conference on Computational Intelligence and Networks (CINE)* (IEEE, 2016), pp. 84–89

6. R. Chatterjee, T. Bandyopadhyay, D.K. Sanyal, D. Guha. Comparative analysis of feature extraction techniques in motor imagery EEG signal classification, in *Proceedings of First International Conference on Smart System, Innovations and Computing* (Springer, 2018), pp. 73–83
7. P. Jahankhani, K. Revett, V. Kodogiannis, Data mining an EEG dataset with an emphasis on dimensionality reduction, in *2007 IEEE Symposium on Computational Intelligence and Data Mining* (IEEE, 2007), pp. 405–412
8. Z. Pawlak, Rough sets. *Int. J. Comput. Inf. Sci.* **11**(5), 341–356 (1982)
9. J. Komorowski, Z. Pawlak, L. Polkowski, A. Skowron, Rough sets: a tutorial, in *Rough Fuzzy Hybridization: A New Trend in Decision-Making* (1999), pp. 3–98
10. R. Chatterjee, D. Guha, D.K. Sanyal, S.N. Mohanty, Discernibility matrix based dimensionality reduction for EEG signal, in *2016 IEEE Region 10 Conference (TENCON)* (IEEE, 2016), pp. 2703–2706
11. R. Chatterjee, T. Bandyopadhyay, D.K. Sanyal, D. Guha, Dimensionality reduction of EEG signal using Fuzzy Discernibility Matrix, in *2017 10th International Conference on Human System Interactions (HSI)* (IEEE, 2017), pp. 131–136
12. S. Lemm, C. Schafer, G. Curio, BCI competition 2003–data set III: probabilistic modeling of sensorimotor/spl mu/rhythms for classification of imaginary hand movements. *IEEE Trans. Biomed. Eng.* **51**(6), 1077–1080 (2004)
13. S.K. Bashar, M.I.H. Bhuiyan, Classification of motor imagery movements using multivariate empirical mode decomposition and short time Fourier transform based hybrid method. *Eng. Sci. Technol. Int. J.* **19**(3), 1457–1464 (2016)
14. Y.R. Tabar, U. Halici, A novel deep learning approach for classification of EEG motor imagery signals. *J. Neural Eng.* **14**(1), 016003 (2016)
15. A. Craik, Y. He, J.L. Contreras-Vidal, Deep learning for electroencephalogram (EEG) classification tasks: a review. *J. Neural Eng.* **16**(3), 031001 (2019)
16. R. Chatterjee, A. Chatterjee, Orthogonal matching pursuit-based feature selection for motor-imagery EEG signal classification. *Int. J. Comput. Appl. Technol.* **64**(4), 403–414 (2020)
17. R. Chatterjee, T. Bandyopadhyay, D.K. Sanyal, Effects of wavelets on quality of features in motor-imagery EEG signal classification, in *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)* (IEEE, 2016), pp. 1346–1350
18. R. Chatterjee, D.K. Sanyal, Study of different filter bank approaches in motor-imagery EEG signal classification, in *Smart Healthcare Analytics in IoT Enabled Environment* (Springer, 2020), pp. 173–190
19. R. Chatterjee, N.B.J. Nashkar, D.K. Sanyal, Cellular automata-based pattern classifier for brain-state discrimination problem. *ICIC Express Lett.* **14**(7) (2020, in press)
20. M. Mohamed, M. Deriche, An approach for ECG feature extraction using Daubechies 4 (DB4) wavelet. *Int. J. Comput. Appl.* **96**(12), 36–41 (2014)
21. R. Chatterjee, A. Datta, D.K. Sanyal, Ensemble learning approach to motor imagery EEG signal classification, in *Machine Learning in Bio-signal Analysis and Diagnostic Imaging* (Elsevier, 2019), pp. 183–208
22. R.T. Schirrmester, J.T. Springenberg, L.D.J. Fiederer, M. Glasstetter, K. Eggersperger, M. Tangermann, F. Hutter, W. Burgard, T. Ball, Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp.* **38**(11), 5391–5420 (2017)
23. J.-S. Huang, Y. Li, B.-Q. Chen, C. Lin, B. Yao, An intelligent EEG classification methodology based on sparse representation enhanced deep learning networks. *Front. Neurosci.* **14** (2020)
24. M.S. Diab, S.A. Mahmoud, A 1.7 nW 24 Hz variable gain elliptic low pass filter in 90-nm CMOS for biosignal detection, in *2019 IEEE International Symposium on Circuits and Systems (ISCAS)* (IEEE, 2019), pp. 1–5
25. R. Widadi, I. Soesanti, O. Wahyunggoro, EEG classification using elliptic filter and multilayer perceptron based on gamma activity features, in *2018 4th International Conference on Science and Technology (ICST)* (IEEE, 2018), pp. 1–5

26. BCI-Competition-II, *Dataset III* (Department of Medical Informatics, Institute for Biomedical Engineering, University of Technology Graz, 2004). Accessed 6 June 2015
27. R.J. Roiger, *Data Mining: A Tutorial-Based Primer* (CRC Press, Boca Raton, 2017)
28. M. Mohammadpour, M.K. Ghorbanian, S. Mozaffari, Comparison of EEG signal features and ensemble learning methods for motor imagery classification, in *2016 Eighth International Conference on Information and Knowledge Technology (IKT)* (IEEE, 2016), pp. 288–292

Automated Diagnosis of Breast Cancer: An Ensemble Approach



Surbhi Gupta

Abstract Breast cancer is the most dominant cancer among women and has caused millions of deaths in the world. Automated learning techniques make a significant contribution to cancer prediction studies. In this study, we focused on ensemble learning techniques for breast cancer diagnosis prediction. The dataset used in this work is publicly available at the University of California, Irvine repository. To achieve more consistent results, our study advocates the use of ensemble approaches. The proposed methodology increased prediction accuracy from 76% with the “RBF” kernel to 81% with ensemble learning. The simulation results prove that the proposed model can serve as a cancer prediction model. This paper presents an ensemble approach to integrate the simulation results of multiple classification models.

Keywords Breast cancer diagnosis · Machine learning · Prediction modeling · Support vector machines · Ensemble learning

1 Introduction

Breast cancer, being the most recurrent malignancy among women, affects 2.09 million cases every year [1]. Breast cancer occurs when cells in the breast grow out of control [2]. Malignancy can originate in any part of the breast. Breast cancer can metastasize to other body parts as well [3]. Research studies affirm that breast cancer arises due to many aspects. Most breast cancers are found in older women. Inadequate availability of resources with pitiful medical structures is the prime cause that the majority of women are detected in late stages [4]. As per the report by World Health Organization [1] in 2018, it is estimated that 627,000 women died from breast cancer. Also it contributes a major share in cancer deaths among women. Breast cancer outcomes and survival rates can be readily improved with

S. Gupta (✉)

Model Institute of Engineering and Technology, Kot bhalwal, Jammu and Kashmir, India

Shri Mata Vaishno Devi University, Katra, Jammu and Kashmir, India

timely recognition of malignancy [5]. Early diagnosis approaches ought to be prioritized is to increase the proportion of breast cancers identified at an early stage, reducing the risks of breast cancer deaths [6, 7]. Automated cancer diagnosis can be done using machine learning (ML) approaches [8]. Such decision making systems can help in providing more efficient diagnosis verdicts [9–11].

The electronic record used in the research work is available in UC Irvine (UCI) machine learning repository. Machine learning approach used for predicting breast cancer in the research study is support vector machines (SVMs) [12]. The final model is built combining SVMs (constructed using different kernels) using ensemble learning techniques [13]. The current study compares the results acquired by different SVMs and establishes the best performance of “RBF” kernel. Further majority voting and weighted majority voting [14] is used as final models. Both the ensemble models achieve great prediction results but the model built using the weighted majority voting performed the best of all the learners. Ensemble-based model can be used for decision making in medical area.

2 Related Work

A brief description of the research studies that investigated the performance of ML models for breast cancer prediction is presented in this section.

The research study [15] employed semi-supervised learning on gene expression-based outcome prediction for cancer patients. Studies like [16] employed ensemble classifier for predicting the diagnosis of breast cancer. A prominent work [7] compared multiple ensemble techniques and concluded the superior performance of stacking multiple classifiers. Further, [17] established the efficiency of weighted voting ensemble method for prediction modeling. Another study used Wisconsin breast cancer dataset [18] and investigated the performance of decision trees [19], K-nearest neighbors (KNN) (N.S. Altman) and neural network [20] for predicting breast cancer. A recent work employed multiple data balancing techniques and multiple ensemble architectures of decision trees (DT), KNN, Naïve Bayes, random forest (RF), XGB and GBC for the classification of breast cancer. An analysis of research articles reviewed is presented in Table 1.

3 Dataset Analysis

UCI Breast Cancer Coimbra dataset was prepared by Peng et al. [29]. Records of females detected with breast malignancy were collected during 2009 and 2013 and is released in 2018. This dataset contains record of 64 malignant cases and 52 healthy persons. A total record of 116 instances each has 10 clinical factors. The attributes used for building the model are age, BMI, glucose, insulin, HOMA, leptin,

Table 1 Analysis table

Study	ML methods	Best performance	Results
Li and Zhou [11]	RF, SVM, AdaBoost, Coforest	Coforest	Avg Err = 27%
Seera and Lim [21]	Fuzzy Min–Max, Cart, RF, hybrid system	Hybrid system	Acc = 98.8%
Sumbaly [6]	Decision tree	J48	Acc = 94%
Purwar and Singh [22]	Combination of MLP and K-means clustering	Hybrid prediction model	Acc = 99%
Nilashi and Ibrahim [23]	EM-PCA-fuzzy, PCA-SVM, PCA-KNN	EM-PCA-fuzzy	Auc = 93%
Quinlan [24]	C4.5, modified C4.5	C4.5	Auc = 94.7%
Andre (1999)	Fuzzy, GA, Evolutionary GA	Fuzzy GA	Auc = 97.4%
Nauck and Kruse [25]	Neuro fuzzy classifier, pruning rule-based	NEFCLASS	Auc = 95%
Setiono [16]	Fuzzy GA, Neuro-rule	Neuro-rule	Auc = 98%
Abonyi and Szeifert [26]	Bayes algorithms, GA-tuned Fuzzy	Fuzzy clustering	Auc = 95.6%
Electrical and Engineering (2004)	NN with RBF kernel, general regression neural network (GRNN) and PNN	Statistical neural network	Auc = 98.8%
Polat and Güne [27]	Least square-support vector machine (LS-SVM) classifier algorithm	LS-SVM	Auc = 98.5%
2007 (Engineering, 2007)	MLP, PNN, RNN, SVM	SVM	Auc = 99.5%
Akay [28]	F-Score, linear SVM and nonlinear SVM	RBF-SVM	Auc = 99.5%
Peng et al. [29]	Sequential forward floating search (SFFS) and SVM	Filter and wrapper method	Auc = 99.5%
Salama et al. [30]	DT (J48), MLP, NB, sequential minimal optimization (SMO), and KNN	SMO + J48-MLP + IBk	Auc = 77.3%
Kumar et al. [31]	SVM-Naive Bayes-J48	Ensemble model	Acc = 97%
Diagnosis [32]	ANN, DTs, KNNs, SVM	Genetically optimized NN	Acc = 99%
Parthiban [23]	WPSO-SSVM, K-means, fuzzy	WPSO-SSVM	Acc = 95%
Salman [33]	ANN-GA, ANN-PSO, ANN-FWA	ANN-FWA	Acc = 98%
Saygili [34]	SVM, KNN, NB, J48, RF and MLP	Random forest	Acc = 98.7%

(continued)

Table 1 (continued)

Study	ML methods	Best performance	Results
Gupta [35]	DT, KNN, Naïve Bayes, RF, XGB, GBC, Stacking	Boosted stacking	Acc = 99%

adiponectin, resistin and MCP-1 and classification. The target variable is “classification” which is categorical having two categories namely, malignant (1) and healthy (2).

4 Proposed Framework

The proposed model in the study is grounded on the ensemble learning techniques that combine weak learners for construction of superior classification model. Four-fifth of total data is used for training while one-fifth of data makes the testing set [36]. Ensemble approaches have marked their significance in most of the earlier studies [37–40]. Hence, we resorted to take advantage of ensemble learners. In the first stage, three SVM models (SVM_1, SVM_2 and SVM_3) are constructed using kernel poly, linear and RBF, respectively. At the final stage, these three base classifiers (SVMs) are combined using vote and weighted majority vote ensemble technique that results in constructing improved single model.

The output of the resultant prediction model is either of the two labels, i.e., healthy or malignant which is the prime objective of the proposed ML architecture.

4.1 Classification Models

This section provided a detailed description of the classification models used in the study for construction of the machine learning prediction model.

4.1.1 Support Vector Machine (SVM) [12]

The superiority of SVM models can be distinguished from other machine learners on the basis its consideration of decision boundary. An optimal decision boundary is such carefully chosen that the proximity from the closest data instances of all the classes can be maximized. The distance between the decision boundary and the data points are termed as support vectors. Figure 3 shows the hyper planes used to classify data points.

- *SVM_1*: This classifier is used with polynomial kernel; a value for the degree parameter of the SVC class is passed with poly kernel. The working of poly

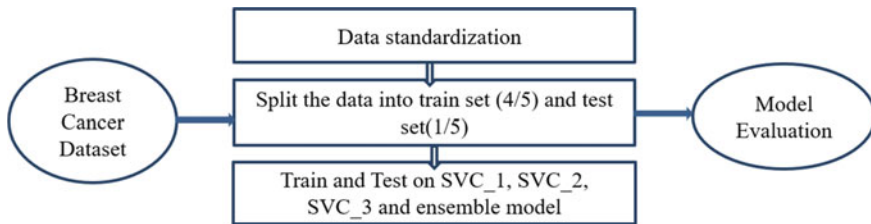


Fig. 1 Flowchart of proposed methodology

kernel is shown in Eq. (1).

$$k(x_i, w_j) = (x_i \cdot w_j + c)^d \quad (1)$$

- *SVM_2*: *SVM_2* is built using the parameter “linear” as the value for the kernel parameter. The working of linear kernel is shown in Eq. (2).

$$k(x_i, w_j) = x_i \cdot w_j \quad (2)$$

- *SVM_3*: This classifier is used in its default settings, i.e., kernel type radial basis function (RBF) is used. The working of RBF kernel is shown in Eq. (3).

$$k(x_i, w_j) = \exp(-\lambda \|x_i - w_j\|^2) \quad (3)$$

The proposed methodology is described in Fig. 1.

4.1.2 Combination Method

In order to combine heterogeneous classification approaches (frequently termed “weak learners”) for improving the robustness of the prediction model, we have used the following two approaches:

4.1.3 Voting [13]

All the three SVMs are combined using voting classifier. The priority of each of the SVM learners (base level) is same irrespective of their performance. The vote of each of the base learner is considered for the final output. The learner receiving more than half of the votes makes the final prediction. The ensemble voting method consolidates the expectation results from a few prepared models. An impediment of this methodology is that each model contributes a proportional add up to the group expectation, regardless of how well the model performed. The working of majority voting classifier is shown in the Eq. (4).

$$\varphi = \text{mode}\{S_1(x), S_2(x), S_3(x)\} \tag{4}$$

$S_1(x), S_2(x), S_3(x)$ represent the classification by SVM_1, SVM_2 and SVM_3

4.1.4 Weighted Voting [13]

The method of combining the base level classifiers using the majority voting combination method was put forward. Contrasting the above-mentioned method, i.e., majority voting that gives equal priority to each of the classifiers. In weighted voting, priority of the meta level learners can be increased based on their performance. At base level (level-1) weighted vote combination method is pragmatic on the outputs generated from every meta level prediction model. The prediction model thus created will produce the final weighted voting-based classifier. The term ‘‘weight’’ used is in context to the accurateness or accuracy score of the base level model. A weighted voting ensemble, weights the contribution of every ensemble member based on the performances of models. This empowers more contribution by better performing models. On contrasting the two group strategies, the weighted ensemble technique can be considered more reliable over the voting ensemble. The working of weighted voting classifier is shown in the Eq. (5).

$$\varphi = \arg \max \sum_{j=1}^3 w_j \{S_j(x)\} \tag{5}$$

The final model used in the study is represented in the Fig. 2.

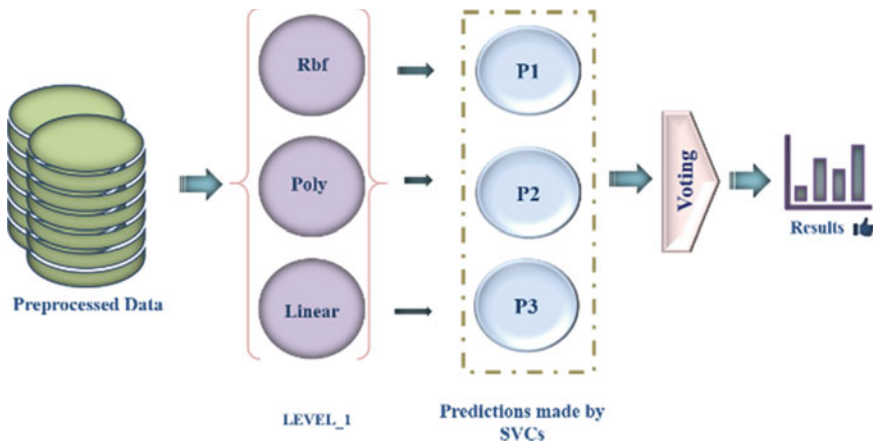
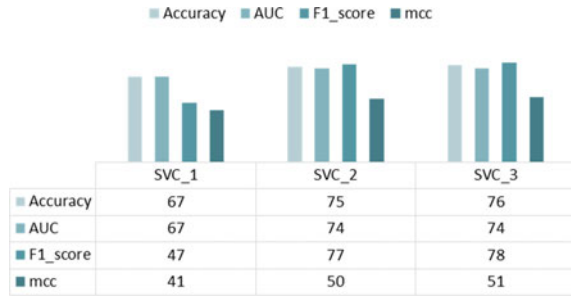


Fig. 2 Ensemble model

Fig. 3 Comparison of SVMs



The ML architecture proposed in the study ensembles the three SVM models, i.e., SVM_1 (SVM_poly), SVM_2 (SVM_linear) and SVM_3 (SVM_RBF) with the majority voting and weighted majority voting technique.

5 Simulation Results

The results gathered by assessment of the approaches are compared on the basis of accuracy, area under the curve (AUC), F1_score and Matthew’s correlation coefficient (MCC).

5.1 Comparison of SVMs

All the three SVM models are evaluated on different parameters. The results are shown in the Fig. 3.

Figure 3 infers that SVM_3, i.e., SVM model using “RBF” kernel performed best and SVM_2, i.e., SVM using linear kernel also performed well. The worst performance results were achieved by SVM using “poly” kernel, i.e., SVM_1 on unbalanced data.

5.2 Results Achieved by the Proposed Approach

All the learning approaches are assessed using different assessment constraints. The results thus obtained are tabulated in Table 2. To highlight the best prediction performers, we have bold-faced such techniques in each of the column. The bold-faced last row of the table indicates the best performance of the weighted vote combination technique. Other than weighted vote, majority voting also performed quite satisfactorily.

This research article is an exploratory study aimed to build predictive models for breast cancer. The classification models used in different breast cancer datasets

Table 2 Evaluation of prediction outcomes

ML techniques	Accuracy	AUC	F1_Score	MCC
SVC_1	67	67	47	41
SVC_2	75	74	77	50
SVC_3	76	74	78	51
Voting	80	79	80	59
Weighted vote	81	81	80	62

have marked the prominence of automated learning. In this study, we have accessed the performance of kernel-based learners, i.e., SVMs as these are efficient learning models [17, 41, 42]. The proposed model is an ensemble model due to their considerable accomplishments in previous research studies [11, 22, 35, 43]. The projected approach is an ensemble of SVMs that has proved its noteworthy performance in in former studies. The prediction outcomes obtained by using the proposed system highlight the efficient role of ensemble strategies for prediction modeling. An accuracy of 76% was achieved by the best base level learner (SVM_3 or SVM_RBF), whereas the projected approach results in achieving 81% accurateness. In future, we aim to explore the performance of deep learning approaches on healthcare datasets as many studies [44–47]. Also, multiple techniques have been incorporated in healthcare [48–50] that have achieved high prediction accurateness using advanced recent approaches and transfer learning models [51].

6 Conclusion

The objective of this investigative study is to build and evaluate an automated diagnosis prediction model that can be employed as a biomarker of breast cancer, grounded on anthropometric records and factors that can be assembled with examination of blood samples. Proposed model has shown magnificent performance. The prediction accurateness of 81% has been achieved using weighted majority voting. Our exploratory study authenticates the effectiveness of ensemble-based machine learning models as efficient predictive models.

Prime limitation of the study is that the technique is fewer amounts of data available. Our focus in the future will be to access the performance of the model on other cancer or disease datasets to establish the generalization of model as best prediction model.

Data Availability Statement

The dataset used in this study is available on URL:

<https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra>

Conflict of Interest The authors declare that they have no conflict of interests to disclose.

References

1. J. Ferlay, Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods 1–13 (2018). <https://doi.org/10.1002/ijc.31937>
2. T.J. Key, P.K. Verkasalo, E. Banks, Rev. Epidemiol. Breast Cancer **44**, 133–140 (1865)
3. H. Kennecke, R. Yerushalmi, R. Woods, M.C.U. Cheang, D. Voduc, C.H. Speers, ... Gelmon, K. J. Clin. Oncol. Metastatic Behav. Breast Cancer Subtypes **28**(20), 3271–3277. <https://doi.org/10.1200/JCO.2009.25.9820>
4. B.O. Anderson, S. Braun, S. Lim, R.A. Smith, S. Taplin, D.B. Thomas, ... D. Panel, *Early Detection of Breast Cancer in Countries with Limited Resources* (2003)
5. K. Jp, G. Pc, Regular self-examination or clinical examination for early detection of breast cancer (2) (2008)
6. R. Sumbaly, Diagnosis of breast cancer using decision tree data mining technique **98**(10), 16–24 (2014)
7. Y. Xiao, J. Wu, Z. Lin, X. Zhao, A deep learning-based multi-model ensemble method for cancer prediction. Comput. Methods Prog. Biomed. **153**, 1–9 (2020). <https://doi.org/10.1016/j.cmpb.2017.09.005>
8. K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction. CSBJ **13**, 8–17 (2015). <https://doi.org/10.1016/j.csbj.2014.11.005>
9. Y. Chen, W. Ke, H. Chiu, Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. Comput. Biol. Med. **48**, 1–7 (2014). <https://doi.org/10.1016/j.compbimed.2014.02.006>
10. A.B. Levine, C. Schlosser, J. Grewal, R. Coope, S.J.M. Jones, S. Yip, Rise of the machines: advances in deep learning for cancer diagnosis. *TRENDS in CANCER*, 1–13 (2019). <https://doi.org/10.1016/j.trecan.2019.02.002>
11. M. Li, Z.H. Zhou, Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **37**(6), 1088–1098 (2007). <https://doi.org/10.1109/TSMCA.2007.904745>
12. C.S. Ong, A.J. Smola, R.C. Williamson, Learning the kernel with hyperkernels. J. Mach. Learn. Res. **6** (2005)
13. S. Gupta, M.K. Gupta, Computational prediction of cervical cancer diagnosis using ensemble-based classification algorithm. Comput. J. (2021)
14. L. Parthiban, Abnormality detection using weighed particle swarm optimization and smooth support vector machine **28**(11), 4749–4751 (2017)
15. N. Shukla, M. Hagenbuchner, K.T. Win, J. Yang, PT US CR. Comput. Methods Prog. Biomed. (2017). <https://doi.org/10.1016/j.cmpb.2017.12.011>
16. Setiono, R, Generating concise and accurate classification rules for breast cancer diagnosis **65** (n.d.)
17. S. Bashir, U. Qamar, F. Hassan, Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble (2014). <https://doi.org/10.1007/s11135-014-0090-z>
18. W.H. Wolberg, Multisurface method of pattern separation for medical diagnosis applied to breast cytology **87**(December), 9193–9196 (1990)
19. J.R. Quinlan, Improved use of continuous attributes in C4 **5**(4)(1996), 77–90 (2006)
20. T. Masters, Probabilistic neural networks. practical neural network recipies in C++ **3**, 201–222 (1993). <https://doi.org/10.1016/b978-0-08-051433-8.50017-3>
21. M. Seera, C.P. Lim, A hybrid intelligent system for medical data classification. Expert Syst. Appl. (2013). <https://doi.org/10.1016/j.eswa.2013.09.022>
22. A. Purwar, S.K. Singh, Expert systems with applications hybrid prediction model with missing value imputation for medical data. Expert Syst. Appl. **42**(13), 5621–5631 (2015). <https://doi.org/10.1016/j.eswa.2015.02.050>
23. A.M. Nilashi, O. Ibrahim, An analytical method for diseases prediction using machine learning techniques. Comput. Chem. Eng. (2017). <https://doi.org/10.1016/j.compchemeng.2017.06.011>

24. J.R. Quinlan, Simplifying decision trees. *Int. J. Hum. Comput. Stud.* **51**(2), 497–510 (1999). <https://doi.org/10.1006/ijhc.1987.0321>
25. D. Nauck, R. Kruse, Obtaining interpretable fuzzy classification rules from medical data **16**, 149–169 (1999)
26. J. Abonyi, F. Szeifert, Supervised fuzzy clustering for the identification of fuzzy classifiers **24**, 2195–2207 (2003). [https://doi.org/10.1016/S0167-8655\(03\)00047-3](https://doi.org/10.1016/S0167-8655(03)00047-3)
27. K. Polat, S. Güneş, Breast cancer diagnosis using least square support vector machine. *Dig. Sig. Process.* **17**(4), 694–701 (2007)
28. M.F. Akay, Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Syst. Appl.* **36**(2), 3240–3247 (2009). <https://doi.org/10.1016/j.eswa.2008.01.009>
29. Y. Peng, Z. Wu, J. Jiang, A novel feature selection approach for biomedical data classification. *J. Biomed. Inform.* **43**(1), 15–23 (2010). <https://doi.org/10.1016/j.jbi.2009.07.008>
30. G.I. Salama, M.B. Abdelhalim, M.A. Zeid, Using multi-classifiers (2012)
31. U.K. Kumar, M.B.S. Nikhil, K. Sumangali, Prediction of breast cancer using voting classifier technique 108–114 (2017)
32. C. Diagnosis, Machine learning with applications in breast cancer diagnosis and prognosis. 1–17 (2018). <https://doi.org/10.3390/designs2020013>
33. I. Salman, Impact of metaheuristic iteration on artificial neural (2018). <https://doi.org/10.3390/pr6050057>
34. Saygili, A.: Classification and diagnostic prediction of breast cancers via different classification and diagnostic prediction of breast cancers via different classifiers (December 2018) (2019)
35. S. Gupta, M.K. Gupta, A comprehensive data-level investigation of cancer diagnosis on imbalanced data. *Comput. Intell.* (2021)
36. A. Celisse, A survey of cross-validation procedures for model selection *. **4**, 40–79 (2010). <https://doi.org/10.1214/09-SS054>
37. S. Gupta, M.K. Gupta, R. Kumar, A Novel Multi-Neural Ensemble Approach for Cancer Diagnosis. *Appl. Artif. Intell.* 1–36 (2021). <https://doi.org/10.1080/08839514.2021.2018182>
38. S. Gupta, M.K. Gupta, Computational model for prediction of malignant mesothelioma diagnosis. *The Comput. J.* (2021). <https://doi.org/10.1093/comjnl/bxab146>
39. S. Gupta, M. Kumar, Prostate cancer prognosis using multi-layer perceptron and class balancing techniques. In 2021 Thirteenth Int. Conf. Contemp. Comput. (IC3-2021), 1–6 (2021). <https://doi.org/10.1145/3474124.3474125>
40. S. Gupta and M. Gupta, Deep learning for brain tumor segmentation using magnetic resonance Images. *IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, 1–6 (2021). <https://doi.org/10.1109/CIBCB49929.2021.9562890>
41. S.-B. Cho, H.-H. Won, Machine learning in DNA microarray analysis for cancer classification. in *Proceedings of the First Asia-Pacific Bioinformatics Conference on Bioinformatics 2003*, vol. 19 (2003), pp. 189–198
42. H.O. İlhan, E. Celik, The mesothelioma disease diagnosis with artificial intelligence methods. in *Application of Information and Communication Technologies, AICT 2016—Conference Proceedings* (2017). <https://doi.org/10.1109/ICAICT.2016.7991825>
43. P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, 3–42 (2006). <https://doi.org/10.1007/s10994-006-6226-1>
44. S. Gupta, M.K. Gupta, A comparative analysis of deep learning approaches for predicting breast cancer survivability. *Arch. Comput. Methods Eng.* 1–17 (2021). <https://doi.org/10.1007/s11831-021-09679-3>
45. S. Gupta, A. Gupta, Y. Kumar, Artificial intelligence techniques in Cancer research: Opportunities and challenges. In 2021 Int. Conf. Technol. Advancements and Innovations (ICTAI). 411–416. (2021). IEEE. <https://doi.org/10.1109/ICTAI53825.2021.9673174>
46. S. Gupta, Y. Kumar, Cancer prognosis using artificial intelligence-based techniques. *SN Comput. Sci.* **3**(1), 1–8 (2022). <https://doi.org/10.1007/s42979-021-00964-3>
47. Y. Kumar, K. Sood, S. Kaul, R. Vasuja, R., Big data analytics and its benefits in healthcare. In *Big Data Analytics in Healthcare* (pp. 3–21). (2021) Springer, Cham

48. Y. Kumar, Recent advancement of machine learning and deep learning in the field of healthcare system. In *Comput. Intell. Mach. Learn. Healthcare Inform.* 7–98 (2021)
49. Y. Kumar, R. Singla, Federated learning systems for healthcare: Perspective and recent progress. In: Rehman M.H., Gaber M.M. (eds) *Federated Learning Systems. Stud. Comput. Intell.* **965**. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-70604-3_6
50. Y. Kumar, S. Gupta, R. Singla, Y.C. Hu, A systematic review of artificial intelligence techniques in cancer prediction and diagnosis. *Arch. Comput. Methods Eng.* 1–28 (2021)
51. Y. Kumar, S. Gupta, W. Singh, A novel deep transfer learning models for recognition of birds sounds in different environment. *Soft. Comput.* (2022). <https://doi.org/10.1007/s00500-021-06640-1>

Survey on Formation Verification for Ensembling Collective Adaptive System



Muhammad Hamizan Johari, Siti Nuraishah Agos Jawaddi, and Azlan Ismail

Abstract The increasing discoveries in the autonomous system had caught researchers attentions. They aimed to find the suitable way to automate the formation of system components in reacting towards the dynamic environments. Among the challenges in designing adaptive systems are to verifying a system's formation with the consideration of challenges such as uncertainty or scalability. Verified formation indicates the correctness of the formation in handling the changes in the environments. The outcome of the process is the verification of the formation in satisfying the specification of the system. This paper surveys the state-of-the-art formation verification in addressing the formation of collective adaptive systems (CAS) components that applying ensemble concepts. The paper also includes verification techniques used in verifying CAS formation and the tools used for formation verification.

Keywords Collective adaptive system · Ensemble frameworks · Formation verification

1 Introduction

Designing an autonomous system is a challenging task [1]. For the system to be adapting autonomously towards changes in the system's environments, a robust and flexible design must be adapted [2]. However, such design is hard to obtain due to the sheer number of elements that needed to be considered [3]. A collective adaptive systems (CAS) is a branch of self-adaptive systems (SAS) where it offers adaptation as a collective action taken by a group of system's components [4, 5]. An example

M. H. Johari (✉) · S. N. A. Jawaddi
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

A. Ismail
Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
e-mail: azlanismail@uitm.edu.my

of CAS is the ensemble system, where by implementing the concepts of ensembles, the collective software formation of the system can be modelled into. However, problems such as uncertainty [6] and scalability [7] limit the effectiveness of the collective software formation of the system. These have led into inaccurate exploitation of the collective action of the system in adapting towards the changes of system environments [8].

Previous survey papers have done reviews on CAS engineering which touch the aspect of architecture and analysis. De Nicola et al. [9] and Krupitzer et al. [10] discussed the topics of the engineering methods of CAS throughout the years. While Krupitzer et al. [10] introduces self-adaptation and its traits on engineering CAS, De Nicola et al. [9] focuses on the methods and techniques used in engineering CAS. In [11], the survey discussed on numbers of formal methods used for specifying and verifying the system's behaviours of CAS. In comparison to existing review works, this review focus on the challenges, frameworks, and verification approaches for realizing CAS. The reason we emphasize challenges in CAS formation to be reviewed in this paper is to identify the effects on CAS formation if the challenges are not considered during development. In highlighting the frameworks for CAS formation, we focus on the frameworks that implement the ensemble concept to form CAS formation. The importance of ensemble frameworks is for identifying frameworks that able to produce CAS formation under the consideration of the challenges faced by the formation. In terms of verification approaches, we want to identify the approaches implemented for verifying the CAS formation in the ensemble frameworks.

This paper is organized as follow. In Sect. 2, we provide the overview of CAS. Section 3 explains the fundamental concept of CAS formation. Section 4 elaborates the ensemble frameworks in terms of the purpose, and its relation to CAS formation. In Sect. 5, we discuss several verification tools that can be used to verify CAS formation. Lastly, in Sect. 6, we highlight the existing works that have implemented the verification approaches in the CAS frameworks.

2 Collective Adaptive Systems (CAS)

Conceptually, collective adaptive systems (CAS) is an extension of SAS. Building from SAS as an atomic component, CAS contains a collection of SAS. The role of an atomic component is to handle its own tasks adaptively. As a collection, each atomic component is supposed to achieve the common goal(s) adaptively [12]. Each atomic component interacts with each other and they adjust their own behaviour according to the current needs. They are also can join or leave the group whenever needed, depending on the context and rules of the defined group [13].

The main purpose of having CAS is to deal with complex and difficult tasks autonomously in many domains [14] such as *transportation and logistic* [15], *smart infrastructure* [16] and *manufacturing industry*. Hence, in a practical setting, a CAS

may include a group of heterogeneous components with different kind of architectures, sensors, computing powers, communication abilities, and capability to operate at different timescales.

For an example, we use a scenario of decentralized E-Mobility case study [17], there is two CAS formations responsible on the parking allocation of Ecars in the parking lots. In this formation, a group of Ecars are inquiring parking spaces from the nearest parking lot. By considering that there are only three available parking spaces, some of the Ecars may not able to be allocated in one of the formations. In terms of the formation that its parking spaces is filled, the formation will send the Ecar to the other formation considering there is an available parking space. If the other formation has an available parking space it will accept the Ecar and if there is no available parking spaces the formation will pass the Ecar to the next formation. In terms of the individual Ecar, it will considering the distance and battery level of the car before deciding to enter the other formation or not. Hence, in the decentralized approach for the E-Mobility case study, multiple formations can be engineered with the consideration of both formation and the individual components involved in the formation. References published for the case study included with ensemble concepts and verification of CAS formation can be seen in Hoch et al. [17] and Pavic et al. [18].

3 Formation for CAS

The categories of CAS formation can be divided into four types of formation [19]. The categories are as follows: Type 1 is a formation that able to anticipated changes and reactions during design-time, Type 2 is a formation that contains alternative strategies in reacting towards changes, Type 3 is a formation that is aware of its own objective and operates with uncertain knowledge on the environment, and Type 4 is a formation that adapts its functionalities by following a biological examples on responding towards changes in environment. However in building such formations, developers need to consider some challenges such as uncertainty and scalability.

Uncertainty is a state of unknown information where it is impossible to exactly explain the current condition, a future result, or more than one viable result [20]. Uncertainty in software formation is a challenging aspect for CAS due to the various possibilities it can bring. Hence, they have to be observed or even estimated to ensure the successful formation of CAS. Uncertainty in CAS can be categorized into external and internal uncertainty [21]. External uncertainty comes from the environment where the software is deployed, whilst internal uncertainty comes from the difficulty in determining the effect of adaptation on the system quality objectives. Furthermore, the sources of uncertainty can be varied as mentioned in [6], among others, due to simplifying assumptions, changes of parameter in future operation, decentralization, and complexity in expressing user's requirements. Without the consideration of uncertainty in CAS formation, task execution of the formation could be halted or delayed due to collaboration problems caused by uncertainty.

In a CAS formation, challenges of scalability are tested through how well do the formation performs in adapting towards the resources provided in the system and how well the formation works with the addition or reduction of said resources [22]. However, constructing a CAS formation with this condition is tricky due to the massive information exchanges between the CAS components and for the formation to act autonomously [23]. The types of scalability are many, as stated in Bondi [24], namely, load scalability, space scalability, space–time scalability, structural scalability, distance scalability, and speed-distance scalability. In Amorim et al. [22], the authors explained the types of scalability according to the condition which fits the scalability types. In relation to CAS, the concern of scalability is preferable in CAS as CAS prioritize dynamicity and working in a decentralized manner. This is due to the decentralized system functions in controlling the system local information scale follows the system size and the performance for the information collection and control implementation are local by each component.

4 Ensemble Frameworks for CAS Formation

Ensemble framework is a framework of a dynamic system with the capability of synthesis a formation of system components known as ensembles. Each ensemble is a dynamic group of system components where the components work by mutually cooperating together in achieving a common goal [25]. In this section, the elements that we discuss are the purpose of each ensemble frameworks, the process of CAS formation and the verification approach implemented for each ensemble framework. ASCENS framework comprises of three components namely, state of the affairs (SOTA), general ensemble model (GEM), and service component ensemble language (SCEL) as stated in Wirsin et al. [26]. In forming a CAS formation, SOTA is used for describing the specification of the domain system and the requirements of the CAS formation in the system. Then, the specific component identifies the intended behaviour of the system by referring to the state space. State space is the set of all possible actions of the CAS formation at a single point of time and the trajectory explains the various actions distributed over time. GEM is engineered to act as a semantic foundation for numerous types of calculi and formal methods which often have a specific associated logic in verifying the CAS formation. Stochastic models are also implemented for helping in capturing the information of probability for specifically trajectories to evaluate the quality of the CAS formation formed.

For the concept of ensemble structures and specification for CAS formation in Helena framework, there are three phases included consisting of *components specification*, *ensemble structures specification*, and *ensemble specification* [27]. Components specification is where the specification of the CAS components' attributes and role in the CAS operation take place. For the ensemble structure specification in forming CAS formation, the role of the components is taken for the specification of the CAS formation such as which components belong to a specific formation and the limitation of a number of components allowed to be in the formation. Then, the CAS

formation's behaviour is specified for the collaboration amongst the components in the ensemble specification phase. The semantics of ensemble specification is where the modelling of the CAS formation take place before reacting towards the changes in the environment. Here, the CAS formation in the structure of an ensemble is defined as states where the specification for the action of the component roles is needed for the collaboration task. Then the modelling of the CAS formation using the automata theory takes place before relaying the solution to the components in runtime.

The development of dependable emergent ensembles of components or DEECO framework focuses on two first-class concepts, namely, component and ensemble [25]. The concept of ensemble and components in ensemble system is coined by Keznikl et al. [28], and further developed for fulfilling the needs of dynamicity in modern software architecture in the form of a framework built on Java annotations [25]. In the ensemble factory, the current knowledge of the CAS and the environment is collected from the knowledge container and combined with the requirement of the CAS formation. Then, the factory will generate a problem description of the current state of the CAS. If CAS formation formed by the factory able to satisfy the specification needed to solved the problem description. The ensemble factory will send a CAS formation instance to the ensemble class for reformation of the CAS formation.

The trait-based component ensemble framework or TCOEF coined by Bures et al. [29], address the needs in developing a complex system where the CAS formation built using ensemble concept may overlap, be nested, and dynamically formed and dismantled in a distributed environment built on numerous network platform. TCOEF executes the individual steps of the component's MAPE-K loop and supervising resolving ensembles while also provides a basic library of reusable traits. TCOEF is presented to handle ensembles formation specifically catering for situation-dependent formation by specifying ensembles in a more complex hierarchical situation [30].

5 Verification Tool for CAS Formation

There exists several verification tools for CAS formation. First, the Simple PROMELA Interpreter (SPIN) model checker [31] is a specifically designed model checker for model checking of concurrent systems. Verification of the model checker starts when the model checker accepted temporal logic representation of correctness claim of the component formation using the PROMELA language. Then, the correctness claim of the formation is modelled into a Buchi automaton and computes the synchronous product of the claim along with the automaton representing the global state space. In the Helena framework, SPIN model checker was integrated [32]. The verification of CAS formation here is by translating the formation model into PROMELA language acceptable in the SPIN model checker.

An SMT solver [33] is an approach for the usage of satisfiability decision of formulas in satisfiability modulo theories (SMT) by the addition equality reasoning,

fixed-size bit-vectors, arrays, quantifiers, and other first-order theories. The approach also enables usage of extended static checking, predicate abstractions, and bounded model checking among others. For formation verification, Z3 SMT solver is used to verify the CAS formation of DEECo framework [23]. Verification of the CAS formation is by satisfying the constraint of CAS formation to prove whether the formation is able to reach the objective under the conditions of the constraints.

Choco solver [34] is a constraint programming solver used for solving satisfaction problems or optimization problems. The approach is able to operate varieties of variable types such as integer, set variables, real variables, and expression while accepting more than 70 constraints for system modelling. In solving the problems of the system, the approach searching capability is parameterized through a set of predefined variable and value selection heuristics. For the conversion of the system model into a problem, the approach entered into a preprocessing mode where it will perform automatic improvements for the system model. In TCOEF framework [30] verifies CAS formation using Choco solver by satisfying the constraints for reaching the objective of the formation.

Alloy analyzer [35] is created due to the fact that coding languages leaves imperfections during design explorations where expressions using such languages are often verbose and indirect. One of the feature of the tool is relational logic, where the approach offers a combination of for-all and exists-same quantifiers of first-order logic with the operators of set theory along with relational calculus. Then, the small scope analysis for running smaller tests in helping designer specifying a scope for bounding the typing in the system specification and translation to satisfiability (SAT) problem helps in translating the system design problem into an SAT problem, where the variables are in the form of simple bits instead of relations. In Cámara et al. [36], alloy analyzer is used in finding all relational model that satisfies the constraints imposed by the CAS formation of Tele-Assistance System following behavioural changes in the system.

UPPAAL model checker [37] is an approach used in verifying real-time systems. The model checker designed for verifying systems by model the systems into a network of timed automata with an extension on integer variables, structured data types, user defined functions and channel synchronisation. Timed automata is a finite-state machine with a variable of time. In formation verification, the model checker is used to verify CAS formation of Gamma Statechart in Graics et al. [38]. The research integrated UPPAAL in the Gamma framework for formation verification by mapping the Gamma composite system model to a network of timed automata which is the input formalism of UPPAAL.

Process analysis toolkit (PAT) model checker is an approach that applies model checking properties against system functioning under fairness [39]. The model used for modelling in PAT is called labelled transition systems and the language is named PAT language. Verification for a model under fairness works by examining fair executions of a specific system and deciding whether certain property is true. The system is workable if and only if there is at least one execution that satisfies the fairness constraints (e.g. goals obtained, cost used in performing tasks). In Mahmood et al. [40], the problem addressed by the study is to formulate a suitable conceptual model

allowing representation of the static structure and dynamic behaviour of a complex system. The objective of PAT in the framework is to analyse the composability of the extended finite-state machine (EFSM) model by evaluating goal reachability and constraint satisfaction.

6 Related Works

The previous works related to this review focus on the verification of CAS formation built using ensemble frameworks. In research done by Hennicker et al. [32], the research applies SPIN model checker to verify formation in HELENA framework. The research also promotes automation of formal verification by translating CAS formation into SPIN model checker language through translation language of HELENA light. Here, the role of the SPIN model checker is to verify CAS formation model with different behaviours on different iteration. Meanwhile in verifying CAS formation using constraint satisfaction technique, Krijt et al. [23] applied SMT solver's constraint satisfaction for model checking CAS formation. The verification of the CAS formation is by satisfying a set of conditions required by the DEECo framework. In research done by Hnetyuka et al. [30], the application of constraint satisfaction technique using Choco solver is used to verify CAS formation in TCOEF framework. The formation verification in this paper is by considering the possibility that component in a formation may belong in two separate formations. The application of Choco solver also applied in Al Ali et al. [41] to verify CAS formation in TCOOF-trust framework. For the modelling of Tele-Assistance System (TAS), Cámara et al. [36] applied alloy analyzer for formation verification. Usage of alloy analyzer is to analysing the design space for families of a software system in finding suitable design decisions by considering formal guarantees of solutions and tradeoffs among the qualities of solutions.

For the related work that applied probabilistic model checking, Mahmood et al. [40] applied labelled transition system for modelling the CAS formation of extended finite-state machine. In Graics et al. [38] UPPAAL model checker is used for verifying CAS formation in Gamma framework which followed the statechart structure by modelling the CAS formation into a timed automata. Usage of probabilistic model checker in Gamma framework and EFSM-based framework utilizes a model that can promote probability into the CAS formation execution during runtime. However, the usage of probabilistic model checker is scarce for the formation verification of ensemble-based CAS formation. Therefore, consideration of applying probabilistic model checking in ensemble-based CAS formation is needed especially considering the challenges that affect the performance of CAS formation during runtime. This is because by applying probabilistic model checking, analysis of CAS formation could be done with the consideration of possible changes in decision-making and also with the consideration of multiple objective which is one of CAS main behaviours.

7 Conclusion

To conclude our review on formation verification in CAS formation built using ensemble concept, we reviewed several elements that fulfilling the context of ensemble-based CAS formation. First, we discussed the overview of CAS and give an example of a case study that represent the CAS formation of components in adapting towards changes in the environment. Second, we discussed the categories of CAS formation and the challenges faced by CAS formation and emphasizes on uncertainty and scalability. Third, we discussed the frameworks that applied the ensemble concept for the formation of CAS components and focus on how the related tool used for formation verification is implemented into the framework. Fourth, we discussed the verification tools used in the framework and how the tools help in the formation verification of ensemble-based CAS formation. Lastly, we discussed the related work for formation verification of ensemble-based CAS formation. Here, we also discussed the verification techniques applied to non-ensemble CAS formation. We found that in ensemble-based CAS formation, probabilistic model checking has yet to be implemented for formation verification. Therefore, in the future, we hope that probabilistic model checking can be applied for ensemble-based CAS formation for modelling the CAS formation with the probabilities of the decision-making process.

Acknowledgements Azlan Ismail acknowledges the support of the Fundamental Research Grant Scheme, FRGS/1/2018/ICT01/UITM/02/3, funded by Ministry of Education Malaysia.

References

1. G. Holzmann, Design and validation of computer protocols (1991). <https://doi.org/10.1145/122419.1024051>
2. W. Wang, Y. Koren, Scalability planning for reconfigurable manufacturing systems. *J. Manuf. Syst.* **31**(2), 83–91 (2012). <https://doi.org/10.1016/j.jmsy.2011.11.001>
3. M. Sayagh, N. Kerzazi, B. Adams, F. Petrillo, Software configuration engineering in practice interviews, survey, and systematic literature review. *IEEE Trans. Softw. Eng.* (8), 1 (2018). <https://doi.org/10.1109/TSE.2018.2867847>
4. M.D. Angelo, S. Gerasimou, S. Ghahremani, J. Grohmann, I. Nunes, On learning in collective self-adaptive systems: state of practice and a 3D framework. in *14th International Symposium on Software Engineering for Adaptive and Self-Managing Systems* (2019)
5. P. Clements, D. Garlan, R. Little, R. Nord, J. Stafford, Documenting software architectures: views and beyond. *Proc. Int. Conf. Softw. Eng.* **6**, 740–741 (2003). <https://doi.org/10.1109/icse.2003.1201264>
6. N. Esfahani, S. Malek, Uncertainty in self-adaptive software systems. *Softw. Eng. Self-Adapt. Syst. LNCS* **7475**, 242–251 (2013)
7. R. De Lemos, H. Giese, H.A. Müller, M. Shaw, J. Andersson, M. Litoiu, B. Schmerl, G. Tamura, N.M. Villegas, T. Vogel, D. Weyns, L. Baresi, B. Becker, N. Bencomo, Y. Brun, B. Cukic, R. Desmarais, S. Dustdar, G. Engels, K. Geihs, K.M. Göschka, A. Gorla, V. Grassi, P. Inverardi, G. Karsai, J. Kramer, A. Lopes, J. Magee, S. Malek, S. Mankovskii, R. Mirandola, J. Mylopoulos, O. Nierstrasz, M. Pezzè, C. Prehofer, W. Schäfer, R. Schlichtning, D.B. Smith, J.P. Sousa, L. Tahvildari, K. Wong, J. Wuttke, Software engineering for self-adaptive systems: a second

- research roadmap. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **7475 LNCS**, 1–32 (2013). <https://doi.org/10.1007/978-3-642-35813-5>
8. R. Frei, G.D.M. Serugendo, Self-organizing assembly systems. *IEEE Trans. Syst. Man Cybern. Part C-Appl. Rev.* **41**(6), 885–897 (2011). <https://doi.org/10.1109/TSMCC.2010.2098027>
 9. R. De Nicola, S. Jähnichen, M. Wirsing, Rigorous engineering of collective adaptive systems: special section. *Int. J. Softw. Tools Technol. Transfer* **22**, 389–397 (2020). <https://doi.org/10.1007/s10009-020-00565-0>
 10. C. Krupitzer, F.M. Roth, S. Vansyckel, G. Schiele, C. Becker, A survey on engineering approaches for self-adaptive systems. *Pervasive Mobile Comput.* **17**(PB), 184–206 (2015). <https://doi.org/10.1016/j.pmcj.2014.09.009>. <https://doi.org/10.1016/j.pmcj.2014.09.009>
 11. D. Weyns, M.U. Iftikhar, D.G. De La Iglesia, T. Ahmad, A survey of formal methods in self-adaptive systems. in *ACM International Conference Proceeding Series* (2012), pp. 67–79. <https://doi.org/10.1145/2347583.2347592>
 12. M. Salehie, L. Tahvildari, Self-adaptive software: landscape and research challenges. *ACM Trans. Auton. Adap. Syst.* **4**(2) (2009). <https://doi.org/10.1145/1516533.1516538>
 13. S. Anderson, N. Bredeche, A.E. Eiben, G. Kampis, M. van Steen, *Adapt. Collect. Syst.* **72** (2013). <http://focas.eu/adaptive-collective-systems/>
 14. D. Sanderson, N. Antzoulatos, J.C. Chaplin, J. Pitt, C. German, A. Norbury, E. Kelly, S. Ratchev, Advanced manufacturing: an industrial application for collective adaptive systems (2015). <https://doi.org/10.1109/SASOW.2015.15>
 15. R.R.S.V. Lon, T. Holvoet, RinSim: a simulator for collective adaptive systems in transportation and logistics. in *2012 IEEE Sixth International Conference on Self-Adaptive and Self-Organizing Systems* (2012), pp. 1–2
 16. E.J. Oughton, W. Usher, P. Tyler, J.W. Hall, Infrastructure as a complex adaptive system. *Complexity* **2018**, 1–11 (2018). <https://doi.org/10.1155/2018/3427826>
 17. N. Hoch, H.P. Bensler, D. Abeywickrama, The E-mobility case study. *Softw. Eng. Collect. Auton. Syst.* **257414**, 513–533 (2015)
 18. I. Pavić, H. Pandžić, T. Capuder, Electric vehicle based smart e-mobility system—definition and comparison to the existing concept. *Appl. Energy* **272**(February), 115153 (2020). <https://doi.org/10.1016/j.apenergy.2020.115153>. <https://doi.org/10.1016/j.apenergy.2020.115153>
 19. N.A. Qureshi, A. Perini, N.A. Ernst, J. Mylopoulos, Towards a continuous requirements engineering framework for self-adaptive systems (2010), pp. 9–16. <https://doi.org/10.1109/rerunt.ime.2010.5628552>
 20. N. Esfahani, S. Malek, Taming uncertainty in self-adaptive software. in *ESEC/FSE '11: Proceedings of the 19th ACM SIGSOFT Symposium and the 13th European Conference on Foundations of Software Engineering* (2011), pp. 242–251
 21. R. Calinescu, D. Perez-palacin, D. Weyns, Understanding uncertainty in self-adaptive systems. in *1st IEEE International Conference on Autonomic Computing and Self-Organizing Systems* (2020)
 22. S. Amorim, E.S. De, J.D. McGregor, Scalability of ecosystem architectures. *2014 IEEE/IFIP Conf. Softw. Architect.* **7–11 April**, 49–52 (2014). <https://doi.org/10.1109/WICSA.2014.36>
 23. F. Krijt, Z. Jiracek, T. Bures, P. Hnetyнка, F. Plasil, Automated dynamic formation of component ensembles taking advantage of component cooperation locality. in *5th International Conference on Model-Driven Engineering and Software Development* (2016)
 24. A.B. Bondi, Characteristics of scalability and their impact on performance. *Proceedings Second International Workshop on Software and Performance WOSP 2000*, 195–203 (2000). <https://doi.org/10.1145/350391.350432>
 25. T. Bures, I. Gerostathopoulos, P. Hnetyнка, J. Keznikl, M. Kit, F. Plasil, Deeco—an ensemble-based component system. in *16th International ACM Sigsoft symposium on Component-based software engineering (CBSE'13)* **June**, 81 (2013). <https://doi.org/10.1145/2465449.2465462>
 26. M. Wirsing, M. Hözl, M. Tribastone, F. Zambonelli, ASCENS: engineering autonomic service-component ensembles. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **7542 LNCS**, 1–24 (2013). <https://doi.org/10.1007/978-3-642-35887-6-1>

27. R. Hennicker, A. Klarl, Foundations for ensemble modeling—the Helena approach. *Specification, Algebra, Softw.* 359–381 (2014). <https://doi.org/10.1007/978-3-642-54624-28>
28. J. Keznikl, T. Bureš, F. Plášil, M. Kit, Towards dependable emergent ensembles of components: the DEECo component model. in *Proceedings of the 2012 Joint Working Conference on Software Architecture and 6th European Conference on Software Architecture, WICSA/ECSA 2012* (2012), pp. 249–252. <https://doi.org/10.1109/WICSA-ECSA.212.39>
29. T. Bures, I. Gerostathopoulos, P. Hnetyнка, F. Plasil, F. Krijt, J. Vinarek, J. Kofron, A language and framework for dynamic component ensembles in smart systems. *Int. J. Softw. Tools Technol. Transfer* (2020). <https://doi.org/10.1007/s10009-020-00558-z>
30. P. Hnetyнка, T. Bures, J. Pacovsky, Using component ensembles for modeling autonomic component collaboration in smart farming. in *SEAMS '20: Proceedings of the IEEE/ACM 15th International Symposium on Software Engineering for Adaptive and Self-Managing Systems* (2020), pp. 156–162
31. G. Holzmann, The model checker SPIN. *IEEE Trans. Softw. Eng.* **23**(5), 279–295 (1997). <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=588521>
32. R. Hennicker, A. Klarl, M. Wirsing, Model-checking HELENA ensembles with spin. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **9200**, 331–360 (2015). <https://doi.org/10.1007/978-3-319-23165-516>
33. L. De Moura, N. Bjørner, Z3: an efficient SMT solver. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **4963 LNCS**, 337–340 (2008). <https://doi.org/10.1007/978-3-540-78800-324>
34. N. Jussien, G. Rochart, X. Lorca, N. Jussien, G. Rochart, X. Lorca, O. Source, J. Constraint, Choco: an open source java constraint programming library. in *CPAIOR'08 Workshop on Open-Source Software for Integer and Constraint Programming (OSSICP'08)* (2008)
35. D. Jackson, Alloy: a language and tool for exploring software designs. *Commun. ACM* **62**(9), 66–76 (2019). <https://doi.org/10.1145/3338843>
36. J. Cámara, D. Garlan, B. Schmerl, Synthesizing tradeoff spaces with quantitative guarantees for families of software systems. *J. Syst. Softw.* **152**, 33–49 (2019). <https://doi.org/10.1016/j.jss.2019.02.055>
37. G. Behrmann, A. David, K.G. Larsen, A tutorial on uppaal 4.0 (2006)
38. B. Graics, V. Molnár, A. Vörös, I. Majzik, D. Varró, Mixed-semantics composition of statecharts for the component-based design of reactive systems, vol. 19 (2020). <https://doi.org/10.1007/s10270-020-00806-5>
39. Y. Liu, J. Sun, J.S. Dong, An analyzer for extended compositional process algebras. in *Proceedings—International Conference on Software Engineering* (2008), pp. 919–920. <https://doi.org/10.1145/1370175.1370187>
40. I. Mahmood, T. Kausar, H.S. Sarjoughian, A.W. Malik, N. Riaz, An integrated modeling, simulation and analysis framework for engineering complex systems. *IEEE Access* **7**, 67497–67514 (2019). <https://doi.org/10.1109/ACCESS.2019.2917652>
41. R. Al Ali, T. Bures, P. Hnetyнка, J. Matejek, F. Plasil, J. Vinarek, Toward autonomically composable and context-dependent access control specification through ensembles. *Int. J. Softw. Tools Technol. Transf.* (2020). <https://doi.org/10.1007/s10009-020-00556-1>
42. J. Hillston, J. Pitt, M. Wirsing, F. Zambonelli, Collective adaptive systems: qualitative and quantitative modelling and analysis. in *Dagstuhl Seminar* 4(14512), 68–113 (2014)
43. J. Petri, Software structure evolution and relation to system defectiveness (2014)

An Inquisitive Prospect on the Shift Toward Online Media, Before, During, and After the COVID-19 Pandemic: A Technological Analysis



Anshul Gupta , Sunil Kr. Singh , Muskaan Chopra ,
and Shabeg Singh Gill 

Abstract Online media plays a vital role in defining the future of tomorrow. Almost every field in the present-day is dependent on technology and online media either for procuring better outputs or for the satisfaction of end clients. In this paper, the authors have tried to bring out the various factors that led to the shift of the majority of individuals to online media, briefly discussing its impact on the economy and social factors. It leverages the facts of how the offline media got impacted not only during the days of the novel coronavirus but before as well. It has been plausibly displayed that fake word gets out quicker and more considerably than authentic news utilizing online media. A boost toward the use of E-platforms has predominantly been taken a closer look at in the middle sections of the paper. Further, the exploratory analysis fore- casts the number of online media users by 2031 and presents inquisitive visualizations on the study of various websites during, before, and after the pandemic in the later sections of the paper.

Keywords Online media · Digital media · Coronavirus · COVID-19 · Work from home · Internet · E-Learning · Technology · Data visualization · Data analysis

1 Introduction

The fact that as the audio and video conferencing utility increments essentially, to become a daily need, associations have also increased their innovative foundations to serve the respective sectors they work in. This leads to a prompt and an expanded

A. Gupta (✉) · S. Kr. Singh · M. Chopra
Department of Computer Science and Engineering, Chandigarh College Of Engineering and
Technology, Sector 26, Chandigarh 160019, India

S. Kr. Singh
e-mail: sksingh@ccet.ac.in

S. S. Gill
CSSS, Indraprastha Institute of Information Technology (IIIT), New Delhi, India
e-mail: shabeg19388@iiitd.ac.in

interest in transmitting the data at a faster speed, network hardware, and programming administrations using cloud-based services as well. The recent times have been a great example of the statement that ‘as the representatives get accustomed to telecommuting, meeting, and executing the necessary requirements of the clients on the web, firms will move to work from home as a standard as opposed to as an exemption’. This is being adopted and received by numerous organizations [1, 2] which have the advanced framework set up to deal with the necessary burden and transfer speed. Schooling is another area where there an emotional move to the online method of executing has too much been justified. Since the start of the lockdown, lectures and other various classes in schools, colleges, and elsewhere have moved their classes to video conferencing stages like Google Meet and Zoom.

Alongside these, coordinated methods of instructing and offbeat stages of development like Udemy, Coursera, and edX have likewise seen an increment in enrollments as well as daily number of visitors [3]. A few organizations are currently moving totally to the online mode for the approaching scholarly year, except for meetings that require an actual presence, for example, the University of Cambridge in the United Kingdom, the framework, and the administration of California State in the United States [4]. The said transformation or the shift toward the online media is not only because of certain restrictions imposed by the offline mode of working, but also due to the advances in the technological sector of the society which help the world cope up with the change in exertion for a better tomorrow. Growing start-ups in artificial intelligence, Internet of things [5], blockchain, machine learning, green computing have made it possible for the establishment of a majority of what is being embraced by associations as a component of their day to day need. Innovations in technology lend a chance to build upon a secure, trusted, and believed data control components [6]. As training and medical care administrations observe a move to these advanced areas, the said technologies empower an approach to get and verify authentications, well-being records, clinical records, and report necessary solutions [7] in a shorter time-frame. It is sure that examination of the plan of such frameworks, alongside keeping up their convenience and value will acquire significance.

In addition to the above talk, a distinguished and well-cited fact about this surge is the enormous increase in the digital devices across the globe as in present times every house has a phone to get the necessary information from the Internet, so why not prefer using the resource than wasting it unnecessarily. The downfall in the Internet rates with the coming of big telecom firms, is yet another undeniable fact in apprehension to the previous statement. One must understand that the gig economy is driven by online stages may it be any sector, education, food, delivery, healthcare, travel, news, etc. which truly complies that ‘Everything is just a click away’. Besides, there are platforms that employ laborers on an impromptu, short-contract, and generally casual premise. Notable instances of these incorporate Ola Cabs, Uber, Airbnb, Swiggy, and Zomato.

With the upcoming of novel coronavirus (COVID-19), the movement of the world confined from one room to another comparable to a user shifting from ‘one online-web app to another’ due to the increased number of resources then. Various industries got a boost, while the others faced serious consequences. The exploration also

focuses in comparing the rise of such platforms due to COVID-19 pandemic through visualizations in later sections.

On the contrary, one central point of contention includes the designation of work as well as coordinated effort, within the groups, or with the individuals working on different projects. The said concern, confronts an ascent in the grade and significance in the world even after pandemic, as the representatives of work from home culture and laborers, increment. The examination may zero in on parts of the procedure of how task standards are set, agreements, developing the trust, and group working, among others. There also exists a possibility that online media may not be feasible to be trusted upon in every scenario as no one knows when about the shutdown of Internet services which solely is the base of this online format. The brief description of the sections included in the paper is as follows: Sect. 1 introduces the paper stating various causes of the online surge. Section 2 ponders upon the past findings of researchers. Section 3 briefly discuss the role of COVID-19 in uplifting various platforms, applications and industries. Section 4 provides a comparative analysis of various online platforms and their usage before, during, after, and in present besides providing a forecast on the expected number of online media users by 2031. Section 5 concludes the paper.

2 Literary Work

Numerous reports have showed the effect of Internet on Indian economy and there is no denying reality that Internet has turned into the favored decision for Indian with regards to look, read news or get in contact with companions/family members [8].

Innovations in information, and especially the web, will stay vital to the world after corona, whereby developments will counter the irrespective misuse. The crucial role of such developments will definitely be played by the administration, and strict guidelines of the actual web for restricting certain unidentified activities. Despite the fact that the web is a worldwide asset and not a single nation can handle its conventions and highlights, its neighborhood access and accessibility, stay an internal issue of the nation facing it. It was certainly due to the pandemic there were certain nations that had confined admittance to the web [9], for specific reasons. The virus has carried the globe to a situation where the individuals not associated with web are confronting absolute prohibition. Due to stern distancing measures set up, certain customized schedules prefer getting logged in to the web in major administrations. Consequently, the persons which are on some unacceptable extremity of the advanced separation are totally forgotten about. Purposes behind the separation are many: unreasonably expensive gadget access, exorbitant access to the Internet, content importance, or administrative Internet closures [10]. In non-industrial nations, the condition is more genuine. Accordingly, it turns out to be critical to investigate the prospects of guaranteed availability. Albeit these issues have been investigated and examined before [11], coronavirus has led to such as situation that web access appears to be important for endurance. Even a couple of investigations have proposed

that admittance or no-admittance into information and communication technologies may build up cultural disparities [12], where the post-pandemic circumstance may upgrade this further.

With a critical computerized partition in social orders, the upstream in the requirement of information from the web has resuscitated the conversation of ZR (zero-rating) plans. ZR plans permit the organizations in allowing its clients' to get information from their destinations and administrations, without bearing information charges. Generally, this is not carefully allowed as it disregards the essential standards of Internet fairness, where web traffic must have a similar need and cost. For example, India had excellence in the record of directing ZR plans. Albeit the public authority did not allow the usage of such plans, in the consequence of the pandemic, TRAI or the telecom administrative authority of India chose to permit forgoing rates for audio and video data for specific sites [13]. Since zero-rating plans can be helpful in noticeable conditions which is obvious in the case of India, exploration on the certain limitations imposed on different boundaries whereby permitting zero-rating plans may build socio-government assistance has colossal down-to-earth suggestions, both for the organizations as well as the controllers of such plans. The current writing on unhindered Internet guidelines and zero-rating plans [14, 15] structures the premise to upgrade survey regarding this viewpoint. Certain concerns to be contemplated comprise: growing telecom foundation, giving sponsored web gadgets, free additional information, or postponing off clients' membership expenses [16].

Advanced cash digital installments and computerized monetary standards are probably going to play a vital part in the world after pandemic. Since these advanced installments are 'hands-free', they are likely to be energized by administrative agencies, and will probably go for a rapid increase. The two foremost gains of advanced cash are identified as the fight against the virus. To begin with, physical currency like notes and coins were speculated to transmit the viral infection and computerized installment was like the 'filthy cash' [17, 18]. Online conveyance administrations were urging clients to make installments through advanced installment frameworks like a credit/charge card or portable installments, with orders by the public authority in a few pieces of India [19]. This is probably going to bring about a flood in computerized installment utilization, which will prompt work on the dispersion of advanced installment innovation. Secondly, in the course of the lockout, there was a cut in the occupations, the administrations gave help by providing the users with installment applications and making them aware of computerized installment techniques. These are advantageous methods of asset move from givers to beneficiaries, as seen in past emergency alleviation cases also [20]. While currently in a state of transition from the offline system to the online one, the COVID-19 pandemic made this transition more rapid due to a halt in conventional ways of living. Therefore, the authors visually analyzed and forecasted the data regarding the growth, pros, and cons of online media.

3 Role of Online (Digital) Mediums in COVID-19 Pandemic

The virus changed the way people used online media. Enormous precautions have been adopted to react to the pandemic, with advanced media leading a crucial role, customarily in the utilization of perceptible mediums to circulate the data, portable prosperity to facilitate more clinical resources, web-based media to advance general mankind efforts, and computerized devices to help populace the board and infection following. Computerized media additionally deals with difficulties such as hoaxes, misguidance, and data leaks. The authors support the expanded utilization of advanced media with an attention on improving trust, building social fortitude and decreasing the clinical weight in office-based destinations [21]. With the increase in social distancing norms, better approaches are being searched out to associate and communicate, virtually, for the most part through video calls and meetings. A major lift has been given to applications which customarily waited in respective obscurities, similar to a video-chat visiting app provided by Google, namely, Duo, and Houseparty that permits gatherings of companions through a solitary visual communication and mess around simultaneously, and the business stages like Google Meet, Zoom, and Microsoft Teams for schooling and work from home [22], which has been visualized in the Fig. 1.

As well as being a worldwide danger, coronavirus is alluded to as an infodemic. The immediate admittance to unending content with access through these stages, for example, social media networks like Twitter, Instagram, and video streaming application like YouTube leave clients vulnerable to hoaxes and problematic data. The said data may emphatically impact singular practices, restricting gathering union and subsequently the adequacy of government countermeasures to the infection. It has been feasibly exhibited that bogus news spreads faster and more substantially than certifiable news via online media. Different investigations have exhibited that fake news have more reach by means of online-based stages as looked at to ordinary print and TV sources [24].

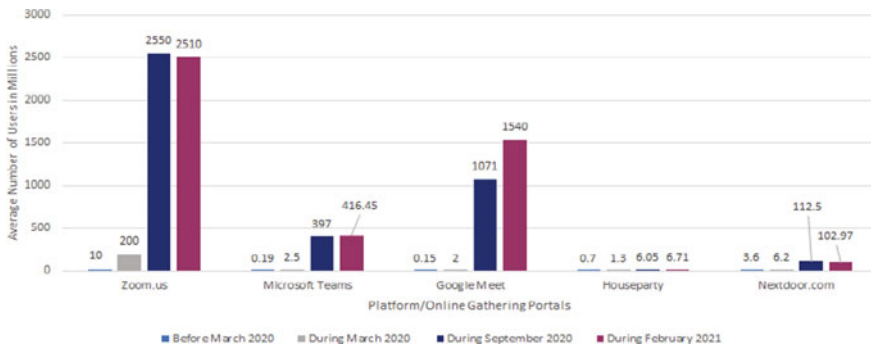


Fig. 1 Online platforms and their average number of users in respective months [23]

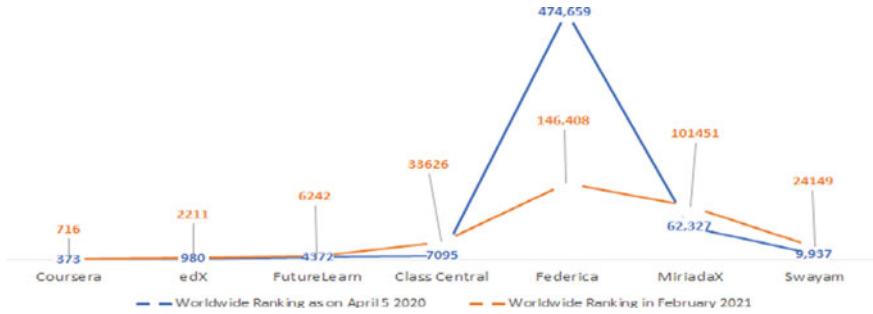


Fig. 2 Worldwide ranking of E-learning websites

4 Comparative Analysis and Visualization of Data

Currently, there is a large movement in media, payment, education platforms from offline to online. Several solutions, as well as problems, arise with this shift. The following subsections provides a broader insight to these statistics and provides their future aspects through visual analysis.

4.1 Based on Worldwide Ranking of E-Learning Platforms

Worldwide rankings, refers to the ranking of each of the websites, namely, Coursera (coursera.org), edX (edx.org), FutureLearn (futurelearn.com), ClassCentral (classcentral.com), Federica (federica.edu), MiriadaX (miriadax.net), and Swayam (swayam.gov.in). The ranking of the said website is based on the data collected and formulated using the one from SimilarWeb [25] that measures the traffic rank of the particular website in comparison to all of the other websites working live on the globe. These are as shown in the Fig. 2.

4.2 Monthly Visitors at E-Learning Platforms

Subsequent months of the lockdown followed an increase in the number of visitors at E-learning portals as shown in the Fig. 3. The orange horizontal bar shows that users frequently visited these portals, which can also be apprehended that as those were the initial days of the COVID-19 pandemic and imposed restrictions, the users were in a constant search of gaining new skills and learning something new. Data for the same was also gathered from similar web [25] and converted to a useful one for the following visualization.

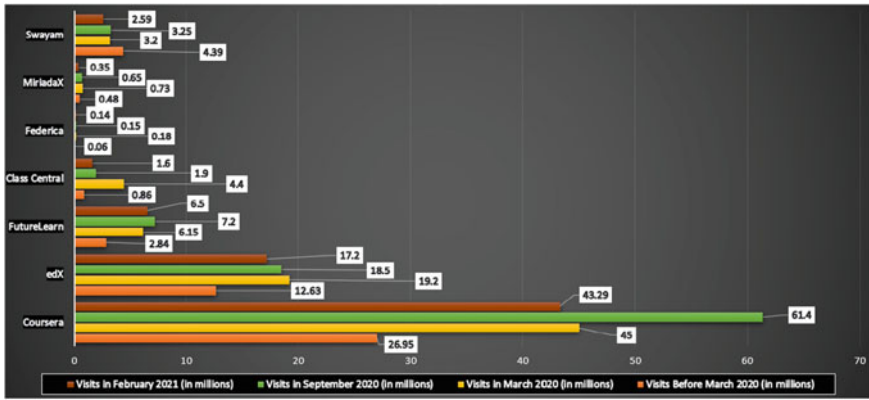


Fig. 3 Trends, showcasing the rise in the number of visitors at E-learning Websites during various months

4.3 Expected Number of Online Media Users by 2031

The expected number of users tells the impact and the forecast of the data gathered from Statista [26] over the years 2011 to 2020. With the information accumulated, it tends to be examined that before the finish of 2020 there were around 376.1 million clients using the online form of media which when determined over a scope of 10 years, the numbers are expected to surge to 796.45 million by 2031 as depicted through the perception in the underneath given Fig. 4.

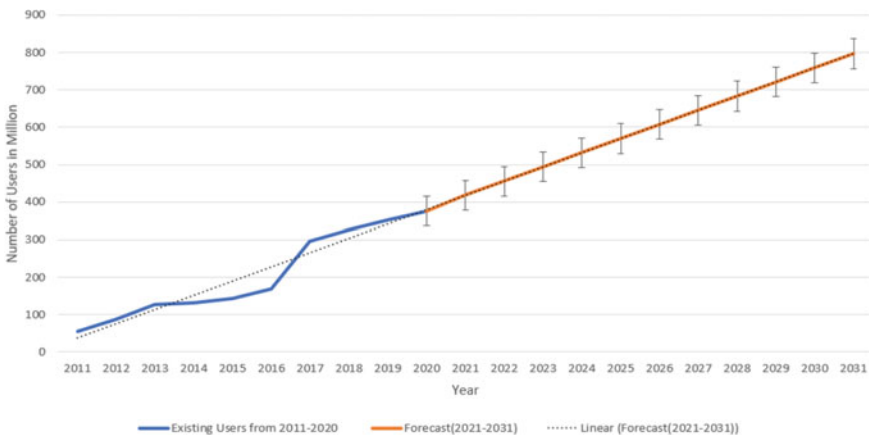


Fig. 4 Expected number of users of online media by 2031 with the existing 2011–20 data-set

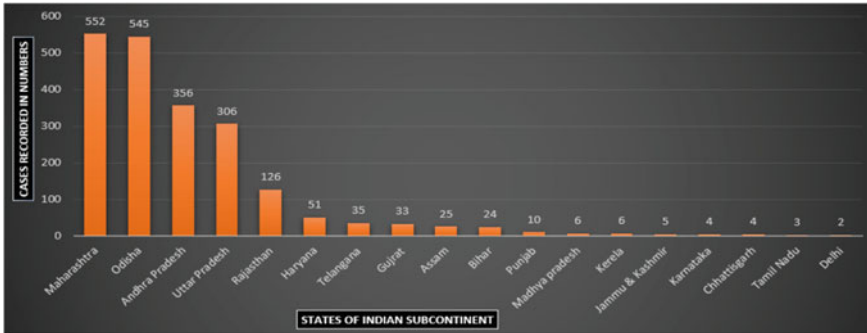


Fig. 5 List of online frauds, registered in various Indian states (2019)

4.4 Shortcomings of Using the Online Form of Media

Just as not every scenario is worth breath-taking, the same is the case here. With the increase in the number of resources, the potential risks of getting trapped also increases. Figure 5 also alludes to the same, whereby it mentions the number of registered frauds through an online medium in the Indian Subcontinent with the data gathered from [27] while the authors believe that the actual number (including the unregistered ones) is yet unknown. As authors, one cannot overlook the negative aspect just to conclude and it becomes effectively important to portray every side and leave everything up to the reader's understanding, which was the primary reason for visualization in Fig. 5.

5 Conclusion

Numerous reports have shown the effect of the Internet on Indian Economy and there is no denying reality that the Internet has turned into the favored decision for people with regards to look, read news or get in contact with companions/family members. COVID-19 is also alluded to as an infodemic since a lot of fake news left people of society vulnerable. Along with the upstream utilization of computerized innovations, the exploration focuses in comparing the rise of platforms such as virtual meeting applications and online education due to COVID-19 pandemic and also throw light on the demerit and online fraud statistics through visualizations in later sections. Also, the expected number of increase in number of users on online media/network is forecasted by authors. The advances in technology and advanced cash digital installments and computerized monetary standards will stay crucial and necessary in post-pandemic situation as well. Further situations as that of the pandemic has carried the society to a circumstance where those minimally associated with the web are confronting absolute negligence.

References

1. A. Akala, More big employers are talking about permanent work-from-home positions. *CNBC* (2020), <https://www.cnbc.com/2020/05/01/major-companies-talking-about-permanent-work-from-home-positions.html>
2. S. Khetarpal, Post-covid, 75% of 4.5 lakh TCS employees to permanently work from home by '25; from 20%. *Business Today* (2020). www.businesstoday.in/current/corporate/post-coronavirus-75-percent-of-3-5-lakhtcs-employees-permanently-work-from-home-up-from-20-percent/story/401981.html
3. D. Shah, Moocwatch 23: pandemic brings moocs back in the spotlight—class. *The Report by class central* (2020). <https://www.classcentral.com/report/moocwatch-23-moocs-back-in-the-spotlight/>
4. A. Jacobs, et al., Virus forces cambridge to hold most classes online next year—the New York times. *New York Times* (2019). <https://www.nytimes.com/2020/05/19/world/corona-virus-news.html>
5. M. Gupta, S.K. Singh, The internet of things: an overview of the awareness, architecture and application. *Int. J. Latest Trends Eng. Technol.* **12**, 19–24 (2019). <https://doi.org/10.21172/1.124.05>
6. N. Upadhyay, Demystifying blockchain: a critical analysis of challenges, applications and opportunities. *Int. J. Inf. Manage.* **54**, 102120 (2020). <https://doi.org/10.1016/j.ijinfomgt.2020.102120>. <https://www.sciencedirect.com/science/article/pii/S0268401219303688>
7. R. De', N. Pandey, A. Pal, Impact of digital surge during covid-19 pandemic: a viewpoint on research and practice. *Int. J. Inf. Manage.* **55**, 102171–102171 (2020). <https://doi.org/10.1016/j.ijinfomgt.2020.102171>
8. M. Tarique, Importance of online media in today's changing trend. tariquenyaz.wordpress.com/ (2014). <https://tariquenyaz.wordpress.com/2014/04/21/importance-of-online-media-in-todays-changing-trend/>
9. M. Chhibber, Militancy in kashmir peaked without 4g, but modi govt keeps forgetting this in court. *The Print* (2020). <https://theprint.in/opinion/militancy-in-kashmir-peaked-without-4g-but-modi-govt-keeps-forgetting-this-in-court/415072/>
10. A. Armbrecht, 4 reasons 4 billion people are still offline. *World Economic Forum* (02 2016), <https://www.weforum.org/agenda/2016/02/4-reasons-4-billion-people-are-still-offline/>
11. M. Warschauer, *Technology and Social Inclusion: Rethinking the Digital Divide* (MIT press, 2004)
12. M. Ragnedda, The third digital divide: a weberian approach to digital in-equalities. *Routledge* (2017). <https://doi.org/10.4324/9781315606002>
13. R.S. Mathews, Request-for-non-charging-of-data. *COAI* (2020). <https://www.medianama.com/wp-content/uploads/Request-for-Non-Charging-of-Data.pdf.pdf>
14. L. Belli, Net neutrality, zero rating and the minitelisation of the internet. *J. Cyber Policy* **2**(1), 96–122 (2017). <https://doi.org/10.1080/23738871.2016.1238954>
15. S. Cho, L. Qiu, S. Bandyopadhyay, Less than zero? the economic impact of zero rating on content competition. *NET Institute Working Paper* **16**(16–04), 46 (2016)
16. J. Shruti, K. Shashidhar, Net neutrality in the time of covid-19. *ORF* (2020). <https://www.orfonline.org/expert-speak/net-neutrality-in-the-time-of-covid-19-65290/>
17. B. Gardner, Dirty banknotes may be spreading the coronavirus, who suggest. *The Telegraph* (2020). <https://www.telegraph.co.uk/news/2020/03/02/exclusive-dirty-banknotes-may-spread-ing-coronavirus-world-health/>
18. M.K. Samantha, Dirty money: the case against using cash during the coronavirus. *CNN* (2020). <https://www.cnn.com/2020/03/07/tech/mobile-payments-coronavirus/index.html>
19. N. Kapoor, Ahmedabad says no to cash on delivery to stop spread of covid-19. *Indi—aTV News* (2020). <https://www.indiatvnews.com/news/india/ahmedabad-digital-payments-mandat-ory-no-cash-on-delivery-to-stop-covid19-616239>

20. I. Pollach, H. Treiblmaier, A. Floh, Online fundraising for environmental non-profit organizations. in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*. IEEE (2005), pp. 178b–178b
21. H. Bao, B. Cao, W. Tang, Digital media’s role in covid-19 pandemic (preprint). *JMIR mHealth and uHealth* **8** (2020). <https://doi.org/10.2196/20156>
22. E. Koeze, N. Popper, The virus changed the way we internet. *The New York Times* (2020). <https://www.nytimes.com/interactive/2020/04/07/technology/coronavirus-internet-use.html>
23. S. Perez, Videoconferencing apps saw a record 62m downloads during one week in march. *Tech Crunch* (2020). <https://techcrunch.com/2020/03/30/video-conferencing-apps-saw-a-record-62m-downloads-during-one-week-in-march/>
24. A. Gupta, A. Bansal, K. Mamgain, A. Gupta, An exploratory analysis on the unfold of fake news during covid-19 pandemic. in: Somani A.K., Mundra A., Doss R., Bhattacharya S. (eds) *Smart Systems: Innovations in Computing*. Smart Innovation, Systems and Technologies, vol 235. Springer, Singapore. https://doi.org/10.1007/978-981-16-2877-1_24
25. SimilarWeb: Data gathering website for analytics. SimilarWeb. <https://www.similarweb.com/>
26. S. Keelery, Number of social network users in India from 2015 to 2018. *Statista* (2020). <https://www.statista.com/statistics/278407/number-of-social-network-users-in-india/>
27. S. Keelery, Number of online banking frauds reported across India in 2019. *Statista* (2021). <https://www.statista.com/statistics/1097957/india-number-of-online-banking-frauds-by-leading-state/>

A Comparative Study of Learning Methods for Diabetic Retinopathy Classification



Qazi Mohammad Areeb and Mohammad Nadeem

Abstract Diabetic retinopathy refers to a state of the human eye that affects retina blood vessels, causing vision impairment or even complete vision loss. In this paper, we classify DR fundus images into the presence and absence of the disease by using a combination of deep learning layers and machine learning algorithms. Machine learning (e.g. random forest and support vector machine (SVM)) and deep learning (e.g. convolutional neural networks (CNNs)) are the most well-known approaches for small and big data, respectively, in image classification tasks. The results are compromised when there is a lack of data. Furthermore, a machine learning algorithm takes less time to train than a deep learning method. As a result, we attempted to develop models that combined machine learning and deep learning approaches. So, three models are proposed in this paper to comparatively study their results. The first model utilises dense layers to classify the data, while the second and third models use SVM and random forest classifiers, respectively. Hence, the model employs CNN and machine learning algorithms to increase the accuracy and efficacy of limited data sets. Usually, it is considered that deep learning algorithms perform better on images, while our results show that random forest outperformed the other approaches. We discovered that combining two learning approaches allows us to get superior outcomes on a short data set without sacrificing accuracy or other measures.

Keywords Diabetic retinopathy · Deep learning · Convolutional neural network · Machine learning

Q. M. Areeb (✉) · M. Nadeem
Department of Computer Science, Aligarh Muslim University, Aligarh, India
e-mail: gj3209@myamu.ac.in

M. Nadeem
e-mail: mnadeem.cs@amu.ac.in

1 Introduction

Deep learning outperforms traditional machine learning with large amounts of training data. However, when there is a limited amount of training data, traditional machine learning (e.g. random forest or SVM) may outperform deep learning. To improve accuracy in image processing applications, features must be extracted/engineered. Alternatively, features can be extracted from convolutional filters used in convolutional neural networks. This is through the process of extracting features with convolutional filters and feeding them into a traditional SVM or random forest classifier to create an image classification solution.

SVM is a popular and efficient supervised learning technique for attribute selection and data classification. SVMs provide robust approaches to predict the binary classification scenarios with high accuracy. In addition, an SVM has high accuracy with less computation power and smaller data. The idea behind SVM is to first identify two support vector hyperplanes and then to maximise the gap between them in order to separate two classes of data. Prior to the advent of deep learning, SVMs performed better in comparison with ANNs in a variety of real-world problems. Random forest represents a collection of random decision trees and is primarily used for supervised learning-based problems such as classification and regression. It is a randomised non-iterative procedure which trains multitude of individual decision trees to yield the final output either by applying mode operation on the set of classes in the case of classification problem or by taking average of predictions obtained from each tree.

Convolutional neural networks (CNNs) are artificial neural networks that are especially built to analyse pixel input and are utilised in image recognition and processing. CNN has emerged as a potent approach to medical image comprehension. CNNs have found applications in various domains especially in medical fields including cancer prediction such as colon, blood or lung cancer, detection of tumour, medical imaging, Alzheimer's and Parkinson's disease prediction and modelling, skin lesions detection, image processing of optical coherence tomography, abnormalities related to body parts such as breast, heart and so on [1]. Convolutional neural networks (CNNs) have been used to diagnose diabetic retinopathy (DR) by analysing fundus images, and they have demonstrated advantages in tasks requiring detection and classification [2].

Diabetes is well known to be a global 'epidemic'. Diabetes afflicted an estimated 415 million individuals globally in 2015, with that figure predicted to grow to 642 million by 2040 [3]. Diabetic retinopathy is becoming more common around the world. According to the International Diabetes Federation (IDF), the global prevalence of diabetic retinopathy (DR) was around 27% from 2015 to 2019. It is also the most common cause of blindness and vision loss in working-age adults. Diabetic retinopathy is a silent disease that may not be detected until retinal alterations have progressed to the point where treatment is ineffective or even impossible.

2 Related Works

A variety of machine learning and deep learning approaches were employed in the DR categorisation. For building the DR classification model, the authors used various types of machine learning algorithms. Support vector machine (SVM), k-nearest neighbour (kNN), artificial neural networks (ANN), unsupervised classifiers (UC) and ensemble classifiers are some of the most often utilised machine learning methods [4].

Mohsin et al. [5], in their study, presented a multichannel CNN for DR detection. The suggested method was evaluated using a DR data set given by EyePACS, which included 35,126 pictures. According to the findings of the experiments, an accuracy of 97.08% was attained. Wan [6] utilised transfer learning in 2018 and analysed how well models like AlexNet, VggNet, GoogleNet and ResNet perform with DR image categorisation. The best classification accuracy was 95.68%, and the findings show that CNNs and transfer learning perform better on DR image classification.

Acharya et al. developed an automated diagnostic technique based on SVM classifiers to distinguish among the possible classes of DR based on its degree of severity [7]. The suggested technique was tested on 300 individuals at various stages of illness, and the attained accuracy, sensitivity and specificity were 82%, 82% and 88%, respectively [6]. Nijalingappa and Sandeep [8] classified diabetic retinopathy into severity categories using the KNN method. The authors of [9, 10] utilised a single ANN algorithm and discovered it to be a superior classification method in the domain of diabetic retinopathy image classification. Random forest was not broadly adopted by researchers. Xiao and Yu [11] employed an RF classifier to detect haemorrhages in retinal images. They selected 55 images from one data set and 35 images from another. They employed 70% of the total images for training the machine learning network, while the remaining 30% were used for testing and classification with an RF method. The experimental findings demonstrated that the RF algorithm obtained a high level of sensitivity.

The authors used a combination of deep learning and machine learning methods in several experiments. For the identification of red lesions, Orlando and Prokofyeva [12] employed an ensemble of deep learning and machine learning approaches. They retrieved characteristics based on intensity and shape using a transfer learning LeNet architecture [13] with ten layers. They attained 97.21% sensitivity and an AUC of 0.9347 [4].

Xu et al. [14] performed a binary classification of DR. Using CNN, they automatically categorised the images in the Kaggle [15] data set as normal or DR images. They utilised a total of 1000 images from the data set. Eight CONV layers, four max pooling levels and two FC layers comprised the CNN architecture. For classification, the softmax function was used on the last layer of CNN. This technique was 94.5% accurate. In their work, Esfahan et al. [16] utilised a well-known CNN, ResNet34, to categorise DR images from the Kaggle data set [15] into normal or DR images. ResNet34 is a pretrained CNN architecture from the ImageNet database. They used a series of image pre-processing methods to increase image quality. The picture count

was 35,000, and the dimension was 512×512 pixels. They reported an 85% accuracy and an 86% sensitivity. Jiang et al. [17] used three pretrained CNN models to categorise their own data set as referable DR or non-referable DR: Inception V3, Inception-ResNet-V2 [18], and ResNet152 [19]. The task was completed with an accuracy of 88.21%.

A deep learning system often takes a long time to train. Deep learning algorithms also attempt to extract high-level characteristics from data. CNN, for example, will attempt to learn low-level characteristics in the early layers and then high-level representation in the later layers, putting it ahead of classical machine learning. Furthermore, deep learning works best with large amounts of data, but machine learning can only handle a small quantity of data. So, in order to get decent results while consuming less time on a limited data set, we create a model that combines deep learning and machine learning techniques. Furthermore, we propose a novel architecture for classifying DR fundus images by combining deep learning layers with a machine learning algorithm on a limited data set. To extract features, convolutional layers were utilised, and machine learning techniques such as SVM and random forest were used to categorise the data. On 2756 fundus images produced from a public data set [20], the suggested architecture was trained and evaluated. In comparison with prior approaches, our work has the following advantages in terms of experiment training time and classification performance. The models' sensitivity, specificity, accuracy and AUC score are all evaluated. Confusion matrices were also provided to help understand the behaviour and performance of the classifier.

3 Proposed Model

The paper classifies images into two categories: DR images and normal images. The pipeline of the proposed model to detect DR is shown in Fig. 1.

3.1 Data Set Description

The OIA-DDR data set [20] provides open access to the fundus images used in this study. It is a high-quality general-purpose data set for diabetic retinopathy classification, segmentation and detection. This open-access collection contains 13,673

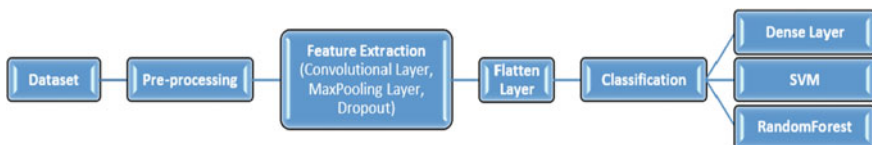


Fig. 1 Pipeline of the model

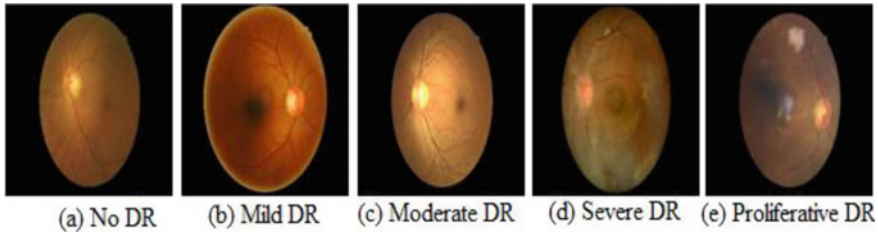


Fig. 2 Stages of diabetic retinopathy (DR) **a** no DR **b** mild DR **c** moderate DR **d** severe DR **e** proliferative DR

fundus images taken at a 45-degree angle and labelled with five DR phases. We chose a random sample of 2756 pictures, of which 1771 were used to train the models, 591 were used for validation and 394 were used for testing. Figure 2 depicts some of the foundation images from the collection.

Images are evaluated on a scale of 0–4. Figure 3 shows the data set’s class labels or scores, as well as the corresponding DR stage and class size. Figure 3a clearly depicts the imbalance in the data set as first class has the maximum number of images, followed by third. To balance the impact of each class, we have divided the data set into two separate non-overlapping groups of images. The presence of the disease is indicated by the labels 1, 2, 3 and 4. The category labels and sizes are shown in Fig. 3b.

3.2 Pre-processing

The DR data set’s input images have a very high resolution, which might make the training process sluggish and end up consuming and running out of memory throughout the training phase. Before being fed into the model, the images are scaled to a resolution of 250×250 pixels. This was done using the OpenCV (<http://opencv.org/>) package.

3.3 Training with Convolutional Layers

CNN is often constructed from a variety of layers, namely fully connected, pooling and convolution. The output of each layer is supplied to the immediate next layer in the architecture, referred as feature map or activation. The pooling layer’s desire is to complicate the characteristics of the specific place. Because certain location features are irrelevant, it just requires other characteristics and their relative position. Pooling layer is placed in between ReLU and convolution layers to reduce the size of image and also to decrease the number of calculated algorithmic parameters. Max pooling

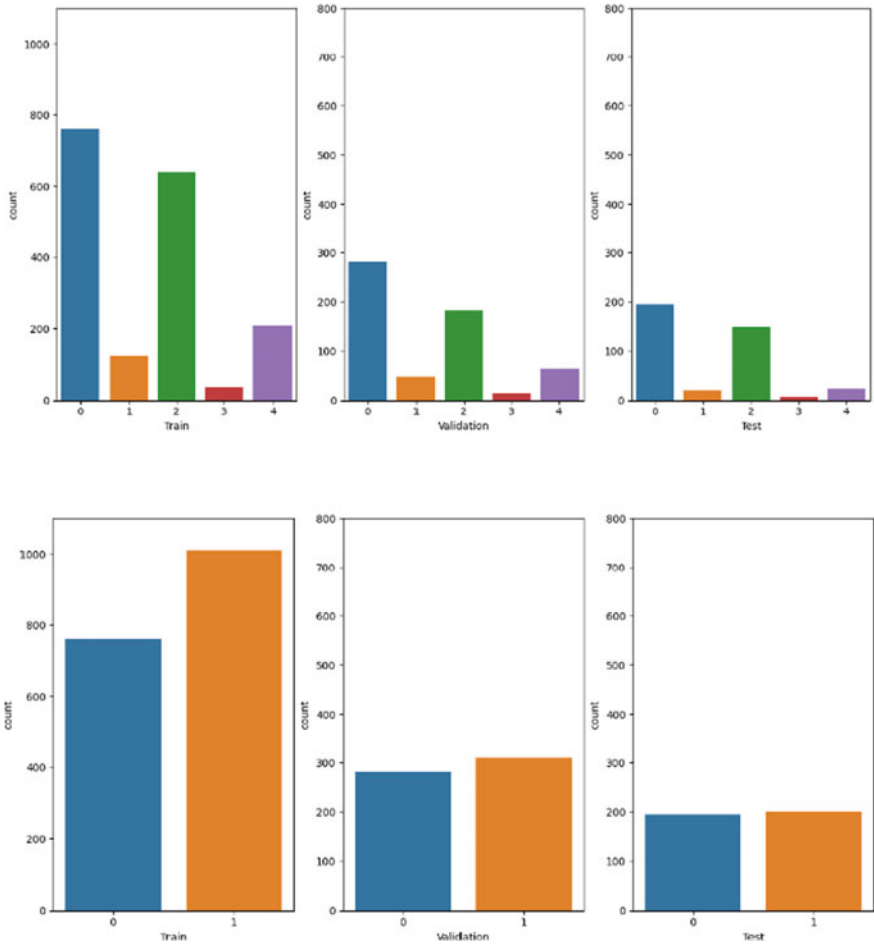


Fig. 3 Data set distribution **a** with five classes **b** with two classes

is the most common. Max pooling summarises the strongest activations throughout a neighbourhood by taking the largest input value within a filter and discarding the remaining values. The logic for this is because the proximity of a highly active feature to another is more important than its exact location. The rectified linear unit (ReLU) layer is a function that converts negative input values to zero. This simplifies and speeds up computation and training while also avoiding the vanishing gradient problem. It is mathematically stated as follows:

$$f(x) = \max(0, x) \tag{1}$$

where x denotes the input of a neuron.

Table 1 Convolutional layers structure

Layer (type)	Output shape	Parameters
conv2d (Conv2D)	(None, 248, 248, 256)	7168
max_pooling2d (MaxPooling2D)	(None, 124, 124, 256)	0
conv2d_1 (Conv2D)	(None, 122, 122, 128)	295,040
max_pooling2d_1 (MaxPooling2D)	(None, 61, 61, 128)	0
dropout (Dropout)	(None, 61, 61, 128)	0
conv2d_2 (Conv2D)	(None, 59, 59, 64)	73,792
max_pooling2d_2 (MaxPooling2D)	(None, 29, 29, 64)	0
dropout_1 (Dropout)	(None, 29, 29, 64)	0
flatten (Flatten)	(None, 53824)	0
Total parameters		376,000

Table 1 depicts the complete architecture of the proposed layers. It was built with the Keras API [21] and the TensorFlow back end [22]. We start with a convolutional layer with 256 kernels and then go on to a max pooling layer. Following that is a convolution layer with 128 kernels, followed by a max pooling layer and a dropout layer of 0.3. Continuing thereafter, convolution layers with 64 kernels are followed by a max pooling layer and a dropout layer of 0.3, followed by a flattening layer at the end to transform the data into one dimension. All convolution layers have a 3×3 kernel with a stride of 1 in all three dimensions, and all pooling levels have a 2×2 pool.

Table 1 depicts the topology of each layer in the proposed network, including input dimensions. Rectified linear unit (ReLU) nonlinearity is used in each convolutional layer for efficient gradient propagation. To reduce overfitting, the dropout layer is employed.

3.4 Developed Models

A CNN model ends with a fully connected layer. Each neuron of this layer is connected with each neuron of the previous layer, making them completely connected with each other. In the first model (Model A), the output from convolutional layers is passed through dense layers of value 32 followed by a dense layer of value 2 with the sigmoid activation function which was used to obtain the class probabilities for final output classes. An adaptive moment estimation (Adam) optimiser was used to calculate the learning rates for different parameters. The loss function computes the difference between the expected and labelled outputs for a given input; binary cross-entropy was utilised as the loss function for this task.

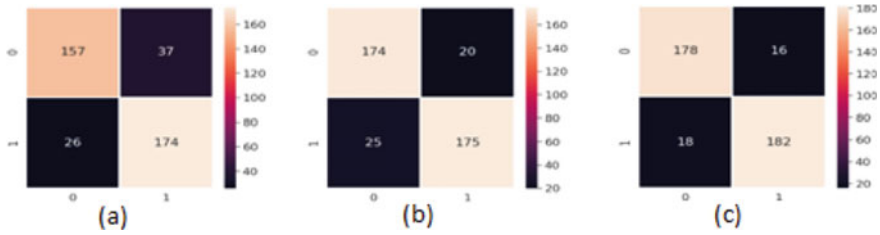


Fig. 4 Confusion matrix on test data set **a** Model A **b** Model B **c** Model C

In the second model (Model B), we use linear SVM as the final layer of the model. The linear SVM was originally formulated for binary classification. The output from convolutional layers is passed through dense layers of value 2 with a linear activation function and a L2 regularisation term. The Adam optimiser was used with a hinge loss function.

Random forest is a classifier that uses a number of decision trees on different subsets of a given data set and averages them to enhance the prediction accuracy of that data set. In the third model (Model C), we use the features from the convolutional layers and pass them through a random forest classifier. Random forest classifier is much faster than CNN and SVM in training.

This section describes the assessment measures that will be used to evaluate and monitor the performance of the proposed network. In this study, 394 images were kept for testing purposes, of which 194 were labelled ‘0’ and the remaining 200 were labelled ‘1’. The models’ sensitivity, specificity, accuracy, AUC score and confusion matrix were all evaluated.

4 Results and Discussion

In this work, three models for binary categorisation of DR into ‘DR’ and ‘non-DR’ are proposed. The network categories were quantitatively defined as 0-No DR and 1-DR. Training accuracies were found to be 91% for Model A, 92% for Model B and 95% for Model C, whereas validation accuracies were 85% and 87% for Model A and Model B, respectively. Confusion matrices on the test data set for each model are shown in Fig. 4. The confusion matrix values are relatively low in the top right and bottom left corners of Fig. 4, indicating that the classification mechanism can distinguish between individuals belonging to classes with significant label differences.

These values in Model C (Fig. 4c) were lower as compared to Model A (Fig. 4a) and Model B (Fig. 4b), which implies that Model C was able to classify images into labelled classes more successfully than Model A and Model B. And accuracy and loss computed during training for Model A and B are shown in Fig. 5. The accuracy steadily improved throughout training, as shown in Fig. 5a, b. After numerous iterations, the performance in terms of accuracy and loss became steady. In model C, the

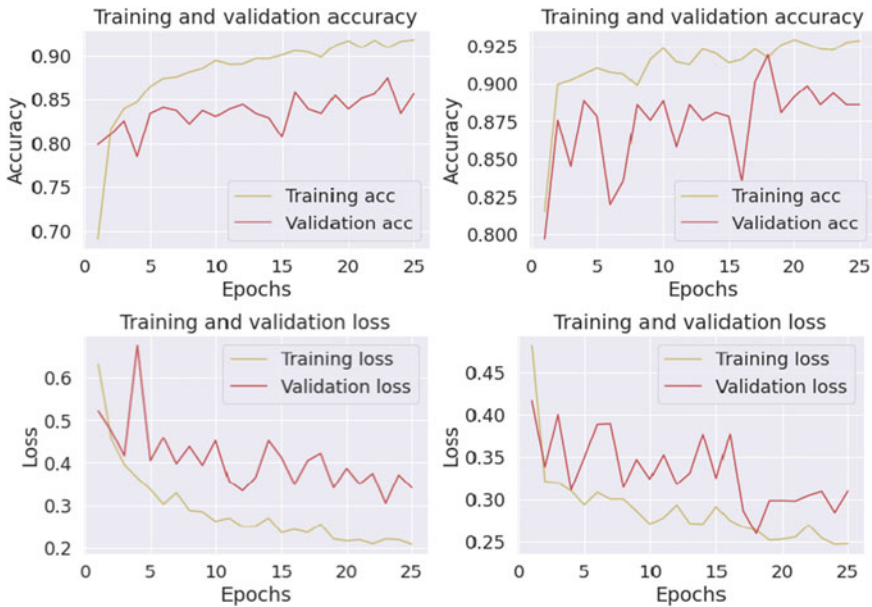


Fig. 5 Accuracy during training of **a** Model A **b** Model B and loss during training of **c** Model A **d** Model B

training took less time as compared to Model A and Model B since the output from the convolution layer was directly fitted to a random forest classifier.

Results from the models are shown in Table 2. We define specificity for this binary classification issue as the number of images is properly recognised as DR out of the actual total without DR and sensitivity as the number of images correctly identified as DR out of the actual total of DR images. The number of images with a valid categorisation is defined as accuracy.

From the table, it can be inferred that both Model A and Model B were equally satisfactory at classifying DR out of total DR images since both achieved a sensitivity of 87%. Furthermore, Model B has higher specificity than Model A, implying that Model B was more successful in classifying non-DR images out of the total non-DR images. The accuracy of Model A and Model B is 84% and 88%, respectively. Moreover, the AUC score of Model B was higher than Model A. Overall, the

Table 2 Results from the models

Model	Sensitivity (%)	Specificity (%)	Test accuracy (%)	Train accuracy (%)	AUC score
Model A	87	80	84	91	0.84
Model B	87	89	88	92	0.88
Model C	91	92	91	95	0.91

SVM classification performs one step higher than the CNN model. This provides that the hinge loss using the L2 norm outperforms the sigmoid function with binary cross-entropy. However, Model C was able to achieve higher sensitivity and specificity as compared to both Model A and Model B. Additionally, accuracy and AUC scores were also noticeably higher than in both the preceding models. This shows that using the machine learning algorithm after feature extraction from the convolutional layer results in better performance than traditional CNN. And moreover, the results have shown that the random forest classifier, in the end, shows better performance than SVM and traditional CNN on diabetic retinopathy problems. For binary classification, an AUC score of 0.91 and an accuracy of 95% were attained, enhancing the outcomes achieved by deep CNN architectures [14, 16, 17] and random forest classifiers [11].

5 Conclusion and Future Scope

Diabetic retinopathy is a serious blinding condition that is one of the consequences of diabetes. Effective and automated identification of diabetic retinopathy lesions has major clinical implications. Early detection enables early treatment, which is critical since early detection can effectively avoid vision damage. In this paper, we are researching using a machine learning algorithm as the last layer of the deep convolutional neural network model in classifying diabetic retinopathy. We were able to successfully classify images in a small data set. While utilising a more general and high-quality data set, our technique was able to achieve decent results without any particular feature detection. Convolutional layers were used to extract features and machine learning algorithms like SVM and random forest were used to classify the data. We trained three models. In all of these models, the input is passed through convolutional layers before being flattened at the end to convert the data into one dimensional. After the flattening layer in the first model, the input is passed through a dense layer. The second and third models use an SVM classifier and a random forest classifier to classify the data, respectively. Random forest was able to achieve high sensitivity as well as high specificity. Usually, it is considered that deep learning algorithms perform better on images, while our results show that random forest outperformed the other approaches. With this approach of using both deep learning and machine learning, we can classify images more easily and save much more time and computational power. Through comparisons of the results of experiments, we can apply this method of binary classification to multi-class classification with a much larger data set.

References

1. D.R. Sarvamangala, V.K. Raghavendra, Convolutional neural networks in medical image understanding: a survey. *Evol. Intell.* <https://doi.org/10.1007/s12065-020-00540-3>
2. G. Litjens, T. Kooi, B. Benjndis, A. Setio, F. Ciompi, M. Ghafoorian et al., A survey on deep learning in medical image analysis, medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017). <https://doi.org/10.1016/j.media.2017.07.005> (PMID: 28778026)
3. K. Ogurtsova, J.D. da Rocha Fernandes, Y. Huang, U. Linnenkamp, L. Guariguata, N.H. Cho et al., IDF diabetes atlas: global estimates for the prevalence of diabetes for 2015 and 2040. *Diabetes Res Clin Pract.* **128**, 40–50 (2017)
4. U. Ishtiaq, S.A. Kareem, E.R.M.F. Abdullah, G. Mujtaba, R. Jahangir, H.Y. Ghafoor, Diabetic retinopathy detection through artificial intelligent techniques: a review and open issues. *Multimedia Tools Appl.* <https://doi.org/10.1007/s11042-018-7044-8>
5. M. Mohsin Butt, G. Latif, D.N.F. Awang Iskandar, J. Alghoza, A.H. Khan, in *Multi-Channel Convolutional Neural Network Based Diabetic Retinopathy Detection from Fundus Images*. 16th International Learning & Technology Conference 2019
6. S. Wan, Y. Liang, Y. Zhang, Deep convolutional neural networks for diabetic retinopathy detection by image classification. *Comput. Electr. Eng.* **72**
7. U. Acharya, C. Chua, E. Ng, W. Yu, C. Chee, Application of higher order spectra for the identification of diabetes retinopathy stages. *J. Med. Syst.* **32**(6), 481–488 (2008). <https://doi.org/10.1007/s10916-008-9154-8> (PMID: 19058652)
8. P. Nijalingappa, B. Sandeep, *Machine Learning Approach for the Identification of Diabetes Retinopathy and Its Stages* (2016)
9. M.A. Al-Jarrah, H. Shatnawi, Non-proliferative diabetic retinopathy symptoms detection and classification using neural network. *J. Med. Eng. Technol.* **41**(6), 498–505 (2017)
10. M.P. Paing, S. Choomchuay, M.D. Rapeporn Yodprom, *Detection of Lesions and Classification of Diabetic Retinopathy Using Fundus Images* (2017)
11. D. Xiao et al., *Retinal Hemorrhage Detection by Rule-Based and Machine Learning Approach* (2017)
12. J.I. Orlando et al., An ensemble deep learning based approach for red lesion detection in fundus images. *Comput. Methods Prog. Biomed.* **153**(C), 115–127 (2018)
13. Y. LeCun et al., Gradient-based learning applied to document recognition. *Proc IEEE* **86**(11), 2278–2324 (1998)
14. K. Xu, D. Feng, H. Mi, Deep convolutional neural network-based early automated detection of diabetic retinopathy using fundus image. *Molecules* **22**(12), 2054 (2017)
15. Kaggle Dataset [Online]. Available <https://kaggle.com/c/diabetic-retinopathy-detection>
16. M.T. Esfahani, M. Ghaderi, R. Kafiyeh, Classification of diabetic and normal fundus images using new deep learning method. *Leonardo Electron. J. Pract. Technol.* **17**(32), 233–248 (2018)
17. H. Jiang, K. Yang, M. Gao, D. Zhang, H. Ma, W. Qian, in *An Interpretable Ensemble Deep Learning Model for Diabetic Retinopathy Disease Classification*. 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2019), pp. 2045–2048
18. C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, in *Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning*. Thirty-First AAAI Conference on Artificial Intelligence (2017), pp. 4278–4284
19. K. He, X. Zhang, S. Ren, J. Sun, in *Deep Residual Learning for Image Recognition*. IEEE Conference on Computer Vision and Pattern Recognition (2016), pp. 770–778
20. OIA-DDR Dataset. <https://doi.org/10.1016/j.ins.2019.06.011>, <http://www.sciencedirect.com/science/article/pii/S0020025519305377>
21. F. C. Keras (2015). <http://keras.io> (io 2017)
22. A. Martn et al., TensorFlow: large-scale machine learning on heterogeneous distributed systems (2016). arXiv preprint arXiv: 1603.04467

CNN for Detection of COVID-19 Using Chest X-Ray Images



Ashish Karhade, Abhishek Yogi, Amit Gupta, Pallavi Landge, and Manisha Galphade

Abstract Coronavirus spread globally in the late 2019, causing the whole world to face an existential health crisis. According to the recent report, animals may also get infected by the virus, so something needs to be done to eliminate this threat named corona. What if we are able to detect the virus at an early stage so that the time it gets to the critical condition, we would be equipped with certain measures. The first thing that gets affected in the body of the infected person are the lungs. So to check out the lung infection, we already have certain traditional techniques, but the automated detection of lung infection using CT Images gives an edge over the traditional healthcare system. Several challenges are faced in the segmentation of infected regions, including high variation in infection characteristics and low-intensity contrast between infections and normal tissues. In this work, we have taken the PA view of the chest X-ray images, which were found unhealthy at the time of screening. After cleaning up all the images or after we are done with data cleaning, we have applied. This work only focuses on the possible methods of classifying COVID-19 infected points, not claiming any medical accuracy. Deep learning to various models evaluates their performances. We have compared CNN, VGG19, Inception V3, Inception-ResNet, ResNet 152, XCEPTION, and we saw that ResNet 152 gave astonishing results.

Keywords Coronavirus · CT images · CNN · VGG19 · Inception V3 · Inception-ResNet · ResNet 152 · XCEPTION

1 Introduction

Coronaviruses are a group of viruses that are spread among humans and animals. This group causes common colds which are of mild to severe nature. SARS and MERS [1] were some of its variants, while the recent outbreak is of type COVID-19

A. Karhade (✉) · A. Yogi · A. Gupta · P. Landge · M. Galphade
Department of Computer Engineering, Sinhgad Institute of Technology, Lonavala, Maharashtra, India

with a total of more than 1200 variants of it, which is why WHO declared it as a public health emergency.

Since December 2019, our world has been facing a health crisis and pandemic situation. According to the global case count reported by the various universities, 3,257,660 identified cases of COVID-19 have been reported so far, including 233,416 deaths and impacting more than 200 countries have been affected by this. The death toll has been increasing and the infection is being spread at an exponential rate in all countries, especially India. The gold standard considered by the global WHO scientist and members is reverse transcription polymerase chain reaction (RT-PCR) [1].

RT-PCR testing is well known testing process but has also reported to suffer from high false negative rate. In addition to RT-PCR X-ray, CT images are also showing good results in both current diagnosis and evaluation. However, according to a study in China, the chest CT analysis acquired 0.96 sensitivity and 0.23 specificity and 0.69 accuracy.

Due to the merit and the three-dimensional view of the lung, the CT -SCAN got an edge over RT-PCR in the analysis. Through this project, we are trying to develop a solution for the faster diagnosing of COVID-19 using images of chest X-rays. We are making the use of artificial intelligence and deep learning preferably to build a robust solution for the problem of slow diagnosing of COVID among people.

2 Related Work

In [1], this author proposed a study based on dataset gathering and applying augmentation. This dataset was released by Mendley with the help of one of China's hospitals; initially, they have a dataset of 300 images and after applying augmentation. They had a dataset of 500 images, i.e., they have created the new dataset based on the previous dataset. They trained their model with the help of the CNN layers but the dataset was too small for this use case and performance can be increased by getting more data.

A bunch of medical students who did not know much of the coding part used the Microsoft AI cognitive [2] class service; they have made their dataset by using the CT scan images of the hospital and as well publicly available dataset. The architecture which MSFT used was not given as it uses the DL platform on its own. They tested their model on their hospital's patients which gave accuracy of 97%. The authors of [3] had taken the advantages of the pre-trained model present. They trained their model with the help of twelve different models and compared their accuracy and other metrics. They used the models like VGG16 [3], VGG19 [3], GOOGLE NET [3], DENSE NET [3], that is they have demonstrated the use of transfer learning to solve their dataset without spending much of their resources. In [4], authors decided to generate more data from the existing dataset, so they generated the fake images using the Generative Adversarial Networks (GAN) and they found that the accuracy is increasing with the size of the dataset, so they used transfer learning to train

the model on their real and fake data, and it turned out the accuracy increased. In [5], authors did not make their custom model for diagnosing the COVID-19 but instead used a readily available model which was previously used for diagnosing pneumonia. The pre-trained model was made by a Chinese researcher, and it was called ChXNet. They used ChXNet and trained this model on new data which was publicly available from radiologists. They had about 1500 samples of training, which significantly improved their accuracy. So, as far as the comparisons are concerned, we have trained our model with much more data and have trained our data on different models and we have tried different methods of preprocessing the data, which in turn has elevated the accuracy and reduced the chances of overfitting.

3 Materials and Methods

The information about the dataset and the methodology used is explained in the subsequent section.

3.1 Dataset

The dataset we have used for this project is gathered from a Kaggle repository but it was not enough to apply deep learning on it, so we have taken more data from the sources publicly provided by the paper. That we have taken into consideration. The images are taken from the padchest, CXR images from a Germany medical school, CXR image from SIRM, Github, Kaggle and Twitter. The dataset contains the frontal view of X-ray images of COVID-19 and non-COVID-19 patients, and the total number of images we had is 211,129. The dataset we used does not only contain the lung images of COVID patients but also for the pneumonia patients and the ones who suffered from lung opacity. The dataset we have gathered is not to show the diagnostic capability of deep learning algorithms but to experience the various ways through which computer vision detects the virus. Since the first part which got damage from the attack of a virus is lungs and since the coronavirus is not alone to attack lungs, there are several diseases like:lung opacity and pneumonia, so this data proved itself useful because the model trained based on this data learned to identify images based on the disease which is given a label in the dataset. COVID-19 images data are collected from publicly available sources and papers. 2473 images were collected from the padchest dataset, 183 images were collected from Germany medical school, 559 images were collected from SIRM, Github, and the rest were collected from various github sources. 10,192 normal images were collected from RSNA and Kaggle sources. Viral pneumonia images were collected from chest X-ray pneumonia dataset from Kaggle. Lung opacity images were collected from the radiological society of North America (RSNA) CXR dataset (Table 1).

Table 1 Number of images used for various categories in train and test dataset

Type	Train	Test
Normal	3255	362
COVID-19	3254	362
Lung opacity	3255	362
Viral pneumonia	1210	135

3.2 Model Formulation

The data we have collected from the Kaggle was already preprocessed since to implement a deep learning algorithm [6] we needed a large amount of data, and as this research is based on the medical background, so not every hospital or organization make their document publicly available for the analysis purpose, so in such cases, data augmentation is the best way, and data augmentation is the way through which we can increase the shape of our data set. Data augmentation generates the synthetic pictures based on the existing pictures in the data set. Hence, we applied data augmentation which also overcomes the possibility of data scarcity, which included rotation, zoom, and shearing of the images. After the dataset is prepared, we used this data to train our proposed model, and to gain a better analysis, we have implemented six different models, and their performance was compared with the help of $F1$ -scores and accuracy scores (Fig. 1).

3.3 Proposed Algorithm

The approach for the proposed algorithm is discussed below.

Step-1: Preprocessing images. We have used ImageDataGenerator by Keras for easy manipulation of all images easily.

1. Re-scaling images (dividing each pixel value by 255 to convert it into [0,1] range)
2. Reshaping image to (224, 224, 3)
3. Shuffle = True

Step-2: Download the pre-trained model

Keras provides in-build support to download the popular pre-trained models that have worked exceptionally well in image classification tests. It also gives the option to download pre-trained weights from the ImageNet competition. We can select pre-trained models like VGG16, VGG19, Inception V3, etc.

Step-3: Fine-tuning the pre-trained model

We did not include the last block of the pre-trained models because our use case is different from the use case that pre-trained models were originally trained

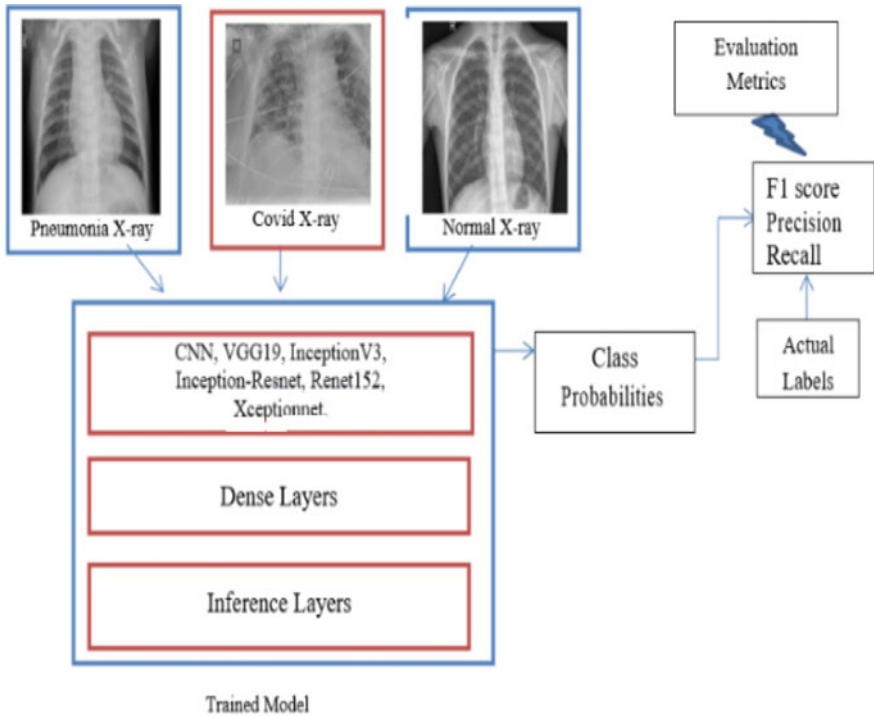


Fig. 1 Proposed model for chest X-ray dataset evaluation

for. We have to fine-tune the models according to our use case. The process of fine-tuning is as below.

1. Put last layer’s output in a variable, say x
2. Add a pooling layer, GlobalAveragePooling() next to x
3. Add a dense layer having activation function as relu [] next to that pooling layer
4. Add a dropout layer with some value after the dense layer
5. Add a dense layer with five neurons at the end of layers having activation function as softmax []
6. We also make the old layers as non-trainable so that our training process does not train the weights already assigned by ImageNet competition

Step-4: Choosing the optimizer and loss function

We have to choose appropriate optimizer and loss function for our model.

1. Optimizer—Adam Optimizer []
2. Loss function—categorical_cross_entropy []

Step-5: Evaluation of all models using $F1$ score and accuracy score.

Step-6: Deploying the model and integrating with web app.

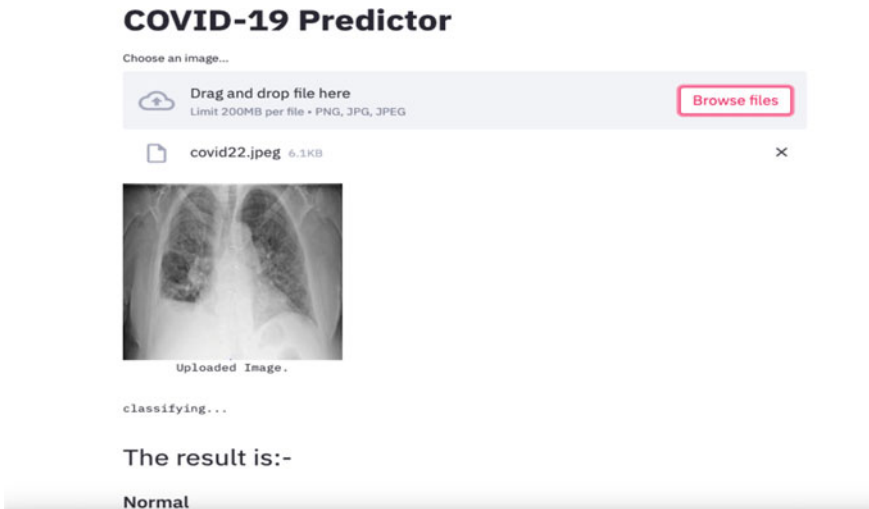


Fig. 2 Web app dashboard for making predictions

We have used, Streamlit to make a web app and integrate our model to web app as shown in Fig. 2.

4 Experimental Results and Discussions

On comparing all the pre-trained and custom CNN models on the training data, we came up with two metrics to decide the superior performance of the model. These metrics are accuracy score and *F1* scores. Tables 2 and 3 show results of accuracy scores and *F1* scores.

As you can see from the above two metrics, accuracy scores, and *F1* scores [7], we can say that the ResNet 152 model performs better than the rest of the models for our use case as shown in Fig. 3. We can use the ResNet 152 model as our final base model and deploy the particular model on the cloud and use it as a production model

Table 2 Accuracy score

No. of epoch	CNN	VGG19	Inception V3	Inception-ResNet	ResNet 152	Xception
Epoch:1	0.8603	0.7447	0.8112	0.7504	0.8589	0.8013
Epoch:2	0.8674	0.8145	0.8226	0.7995	0.8570	0.8311
Epoch:3	0.8636	0.8339	0.8372	0.8123	0.8759	0.8358
Epoch:4	0.8858	0.8367	0.8339	0.8207	0.8697	0.8362
Epoch:5	0.8806	0.8334	0.8523	0.8321	0.8834	0.8485

Table 3 *F1* score

No. of epoch	CNN	VGG19	Inception V3	Inception-ResNet	ResNet 152	Xception
Epoch:1	0.8592	0.7253	0.8067	0.7399	0.8532	0.7969
Epoch:2	0.8651	0.7984	0.8215	0.7941	0.8535	0.8299
Epoch:3	0.8570	0.8308	0.8379	0.8089	0.8764	0.8329
Epoch:4	0.8822	0.8303	0.8320	0.8177	0.8719	0.8380
Epoch:5	0.8725	0.8366	0.8540	0.8283	0.8812	0.8460

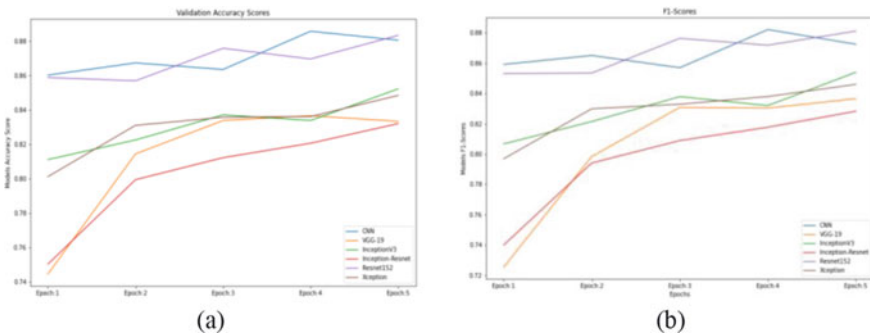


Fig. 3 **a** Validation accuracy scores, **b** *F1* scores

[8]. Figure 4 shows confusion matrix for final selected model based on *F1* score and accuracy.

We can use the ResNet 152 model as our final base model and deploy the particular model on the cloud and use it as a production model [8].

5 Conclusion and Future Scope

We know that current active cases in India are 19M+, and the death rate is increasing, and we do not have that many resources to detect the virus at the early stage. We intend to make a system that can help doctors and pathologists detect infection at the earliest stage, which will help detect COVID faster. As this system helps people detect the symptoms at the earliest stage, this will prevent thousands of people from getting infected daily. We experimented with multiple CNN models, to classify the COVID-19 affected patients, using their chest X-ray scam images. In the future, the large dataset for the chest X-ray will be considered to validate our proposed model on it. We developed this model for the patients who have suffered this COVID-19 disease and also very useful in the many hospitals to detect the disease. It is also very useful to medical professionals for any practical use cases of this project. This

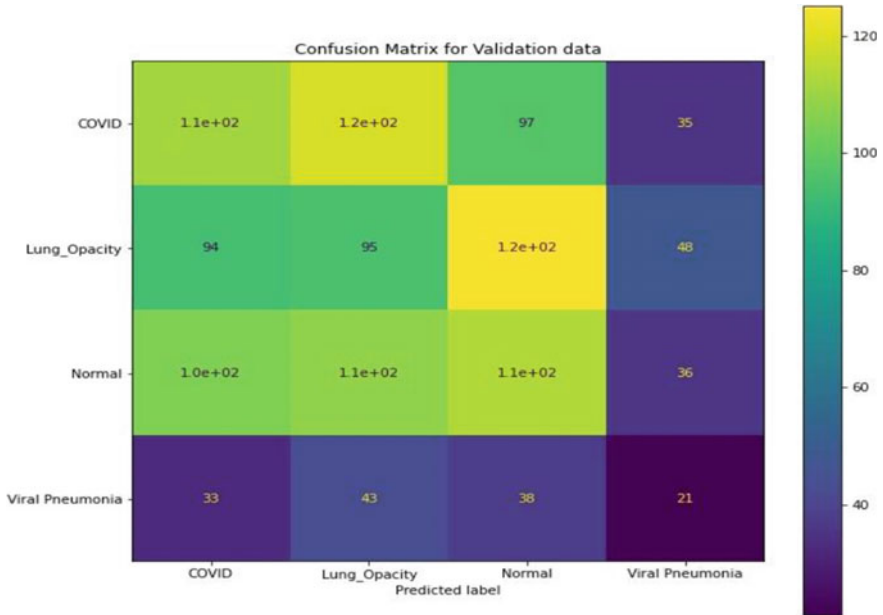


Fig. 4 Confusion matrix for the final model (ResNet 152)

is the solution for pathologists in the form of a web/mobile application with 88.34% accuracy. The work can be extended by adding Chabot functionality that will ask some set of questions about patient's health and the answers will be considered with the model result to get more than 99% accuracy from the model.

References

1. G.C. Ooi et al., *Severe Acute Respiratory Syndrome: Temporal Lung Changes at Thin-Section CT in 30 Patients* (vol. 230, no. 3, 2004), pp. 836–844. <https://doi.org/10.1148/radiol.2303030853>
2. A.A. Borkowski, N.A. Viswanadhan, L.B. Thomas, R.D. Guzman, L.A. Deland, S.M. Mastorides, Using artificial intelligence for COVID-19 chest X-ray diagnosis. *Fed. Pract.* **37**(9), 398 (2020). <https://doi.org/10.12788/FP.0045>
3. T. Majeed, R. Rashid, D. Ali, A. Asaad, *Covid-19 Detection Using CNN Transfer Learning from X-ray Images*, p. 2020.05.12.20098954 (medRxiv, May 2020). <https://doi.org/10.1101/2020.05.12.20098954>
4. M. Loey, F. Smarandache, N.E.M. Khalifa, Within the lack of chest COVID-19 X-ray dataset: a novel detection model based on GAN and deep transfer learning. *Symmetry* **2020** **12**(4), 651 (2020). <https://doi.org/10.3390/SYM12040651>
5. A. Mangal et al., *CovidAID: COVID-19 Detection Using Chest X-Ray* (2020). Accessed 12 Aug 2021 [Online]. Available: <https://arxiv.org/abs/2004.09803v1>
6. S. Wang et al., A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19). *Eur. Radiol.* **31**(8), 6096 (2021). <https://doi.org/10.1007/S00330-021-07715-1>

7. M. Sokolova, N. Japkowicz, S. Szpakowicz, Beyond accuracy, F-Score and ROC: a family of discriminant measures for performance evaluation. AAAI Work. Tech. Rep. **WS-06-06**, 1015–1021 (2006). https://doi.org/10.1007/11941439_114
8. R. Parikh, A. Mathai, S. Parikh, G.C. Sekhar, R. Thomas, Understanding and using sensitivity, specificity and predictive values. *Indian J. Ophthalmol.* **56**(1), 45–50 (2008). <https://doi.org/10.4103/0301-4738.37595>

Crop Yield Prediction Using Weather Data and NDVI Time Series Data



Manisha Galphade, Nilkamal More, Abhishek Wagh, and V. B. Nikam

Abstract Agriculture is the main part of India's economy which provides food security for the country and produces several raw materials for the industries. The development in agriculture is an important aspect in the nearby future. Sugarcane crop is one of the highest producing crops in India, and Maharashtra state is the second highest producer of the sugarcane. In this paper, a novel approach for the yield prediction of the sugarcane crop is proposed based on the weather and soil parameters, normalized difference vegetation index (NDVI), and several machine learning regression techniques. The model is verified using historical data set for the sugarcane crop. The model consists of three stages—(I) Prediction of the weather parameters, (II) prediction of NDVI using weather parameters as input, (III) yield prediction using stage I and II as input. The decision tree regressor gives the highest accuracy of 91.5% for the final model of sugarcane crop yield prediction.

Keywords Machine learning · Crop yield production · Remote sensing data · Normalized difference vegetation index (NDVI) · Time series data · Yield prediction · Weather data

1 Introduction

This work focuses on precision agriculture for farming sustainability on crop yield modeling. The yield modeling at vast uses selection of crop growth parameters, including recognition systems, recommendation engines, data mining and informatics, as well as autonomous control systems.

Agriculture is one of India's key development sectors, as well as the rural sustainability largely depends on the agricultural economy. Because of factors like climate

M. Galphade (✉) · N. More · A. Wagh · V. B. Nikam
Department of Computer Engineering and Information Technology, Veermata Jijabai Technological Institute, Mumbai, Maharashtra, India

V. B. Nikam
e-mail: vbnikam@it.vjti.ac.in

change, water level uncertainty, unpredictable rainfall, improper utilization of pesticides, etc., the agriculture sector has become uncertain even for survival. In relation to this, the paper discusses the detailed conducted analysis of agricultural data in order to understand the production level. The machine learning models presented in this direction contain multiple methods for defining rules as well as patterns of large crop yield data sets. The study suggests that the government is identifying a way to improve small farm household service loans to allow viable conditions to change survivor-ship agriculture into market-oriented agriculture, allowing other smallholder inputs in similar ways that will improve agricultural production at the same time. Results from empirical studies indicate that access to credit has increased crop production as a key driver of smallholder agriculture commercialization.

The major objective of this research is to study and devise methodology so that descriptive analysis can be effectively conducted on crop yield production. In this paper, we propose the novel hybrid approach for the crop yield prediction using the weather parameters and NDVI time series.

This paper contains the following sections. Section 2 explains background. Section 3 introduces data sets used for the final prediction. Problem and prediction model with regression techniques in machine learning is discussed in Sect. 4. Section 5 shows experimental results and analysis for the prediction model. Section 6 concludes our work and gives research direction in the future.

2 Literature Survey

To enhance the quality and crop yield for increasing economic growth and attain profitability, a thought to identify the appropriateness of crops and yield is required [1]. Investigation found that statistical data on agriculture in India was dependent on historical data for climate and production for prediction of crops. Climate changes also impact crop yield with their effect on water quality, soil, and crop in total [2].

2.1 Crop Yield Prediction Using Weather Data

The researchers use three prediction tools, SARIMA and ARMA and ARMAX which are correlated with efficiency and are employed to predict precipitation and temperature which in turn are utilized to forecast crop yield on the basis of the dynamic logical model and random forest (RF) algorithms on Tamilnadu data [3]. The weather data is useful for forecasting crop yields using neural networks, and support vector regression (SVR) [4], ANN, fuzzy logic (FL), and hierarchical methods were utilized in weather prediction that focuses on an assessment of historical rainfall information and data on weather variables with minimal cost and efforts [5].

A regression analysis model was deployed using the fuzzy logic relationship in different degrees to conduct the process of fuzzification using an approach of time

series, based upon regression which covers frequency of data and actual production [6]. Different soil attributes are important factor in crop growth, so by using KNN and support vector regression techniques, the model was developed for yield prediction using comparative analysis [7]. Machine learning has proved to be noteworthy for crop yield prediction, and for this neural network, regression techniques, clustering, generic algorithm are used for sustainable precision farming [8].

2.2 Crop Yield Prediction Using Remote Sensing Data

The NDVI and back scattering coefficients and Polari metric indicators based on Sentinel 1 SAR images show the potential and assimilation value of the SAR and optical remote sensing data in crop models for crop monitoring and yield estimates [9]. Advanced machine learning methods can predict the production of silage maize using averaged and combined NDVI series and RF regression with the overall *R*-value by more than 0.87 each year for forecasting corn yield in 2015 when training in data between 2013 and 2014 [10].

The accuracy result evaluation shows that different thresholds for the recovery of SOS and EOS are enhanced by the model which can be used as a crop yield prediction attribute for various plant species [11]. For the regional yield forecasting, SVM and LASSO regression was used with NDVI from LANDSAT 8 images, and model was developed for spring maize crop [12]. The green spaces area in Mumbai region is calculated using MODIS and LANDSAT 7 multispectral images, and support vector machine analysis shows the 50% reduction in green spaces in the Mumbai region in the last 15 years. Also, comparative approach is given between Geospark and Spatial Hadoop for temporal and unstructured data [13, 14].

The sugarcane crop yield is predicted using the long-term time series of NDVI values of LANDSAT 2; weather attributes and different regression techniques are carried out for better accuracy of the yield prediction model [15]. The spiking neural network (SNN) is used for the prediction of the crop yield from NDVI time series. The MODIS-250m data along with historical crop data is combined for accurate crop yield prediction six weeks before harvesting [16]. The vegetation index is extracted from LANDSAT 8 satellite imagery and combined with the yearly timeline of the crops like wheat and silage maize. Agro-ecological zoning (AEZ) is combined with the spatial data and a model was developed with more than 90% accuracy for the yield prediction [17].

3 Data Set

In this paper, we have focused on the Vidani village which is situated in Phaltan tehsil of Satara district in Maharashtra. The latitude and longitudinal coordinates for this village are 17.9824°N, 74.5113°E. By using these coordinates, we have collected the

Table 1 Weather parameters

Parameter names	Units
Precipitation	In mm
Temperature (2 m above the ground)	In C
Specific humidity (2 m above ground)	In kg kg ⁻¹
Relative humidity (2 m above ground)	In %
Surface pressure	K Pa
Wind speed (10 m)	In m/s
Dew point	In C
Cloud coverage	–
NDVI	–
Soil parameters—phosphorus, nitrogen, potassium	–
Historical production	In Tones
Area	In Hectares

monthly weather data from NASA POWER for the last 20 years. The precipitation data is collected from the CHRS Data Portal. The soil parameters and historical production data since 2000 for the Satara district were collected from the Ministry of Agriculture and Farmers Welfare Department. The NDVI index which ranges from -1 to $+1$ is collected using an earth engine and the satellite used is Sentinel-2. The satellite was launched in 2015 and has a resolution of 10 m. The bands which are useful for the NDVI calculation are B4 and B8. They are the red and near-infrared bands. The normalized difference is calculated from these two bands, and the mean value of the NDVI for each month is calculated using the median() function from the year 2016 to 2019. The parameters used in this data set are listed as in Table 1.

The monthly mean values for each of the above weather parameters such as precipitation, mean-max-average temperature, dew point, humidity, min-max wind speed, cloud coverage is calculated and taken as an input for stage I of the model.

The historical production and area of the last 20 years for the whole district are taken and adjusted for the tehsil and village level on ground basis. The vegetation index is calculated by using the formula:

$$\text{NDVI} = (\text{NIR} - \text{RED}) / (\text{NIR} + \text{RED}) \quad (1)$$

The monthly mean for each value is calculated for the given latitude, longitude coordinates and stored in the database which is used for the stage II input. The correlation matrix is then plotted using the above data. The correlation matrix values also range from -1 to $+1$, where -1 is a weakly related entity and $+1$ is strongly related considering every parameter in the data set mentioned in Table 1. As per the correlation matrix in Fig. 1, NDVI is related with precipitation, humidity, dew point and is negatively related with cloud coverage which is acceptable. Also, precipitation

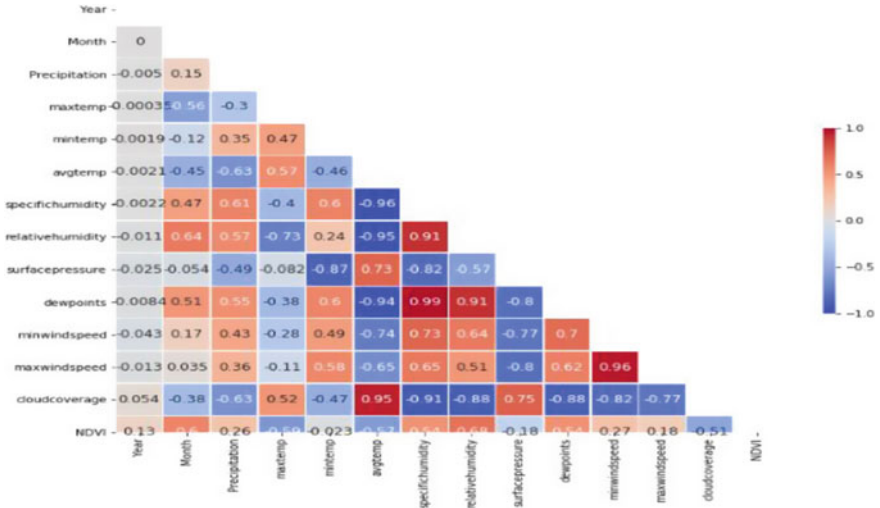


Fig. 1 Correlation matrix

is related with humidity and minimum temperature. So by using the correlation matrix, we can safely avoid the cloud coverage which is negatively related to NDVI.

The NDVI is calculated at 10 m resolution with Sentinel-2 multispectral images within the given boundary of the interested region. The graphs of the NDVI time series analysis are plotted in Fig. 2.

Observing the time series analysis of NDVI, we can clearly say that the value is minimum at the start of the year and increases to its maximum point in the month

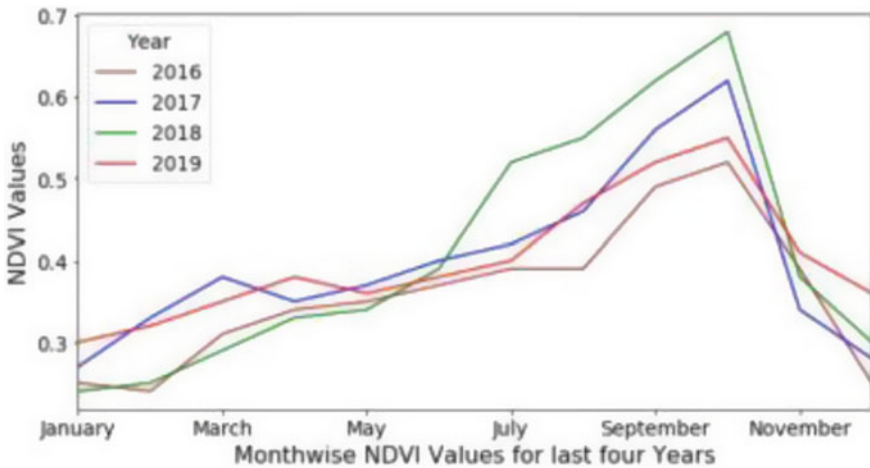


Fig. 2 NDVI time series analysis

of October. So this is the harvesting season for the sugarcane crop. The NDVI value is predicted using the above historical data and regression techniques which are the input for the stage III of the prediction model. The different algorithms such as linear, lasso, ridge regression, decision tree, random forest, *K*-nearest neighbor, SVR-rbf, Bayesian regression are used for the above data set on each and every stage.

4 Yield Prediction Model

The system architecture consists of three layers as shown in Fig. 3.

- Layer I Database layer where all the historical data reside.
- Layer II Operational layer, where all the preprocessing on the data is done. Also, the required features are extracted from the data, and the final yield is predicted using the three stage prediction model using different machine learning regression algorithms.
- Layer III Presentation layer is the final layer where the predicted results are displayed.

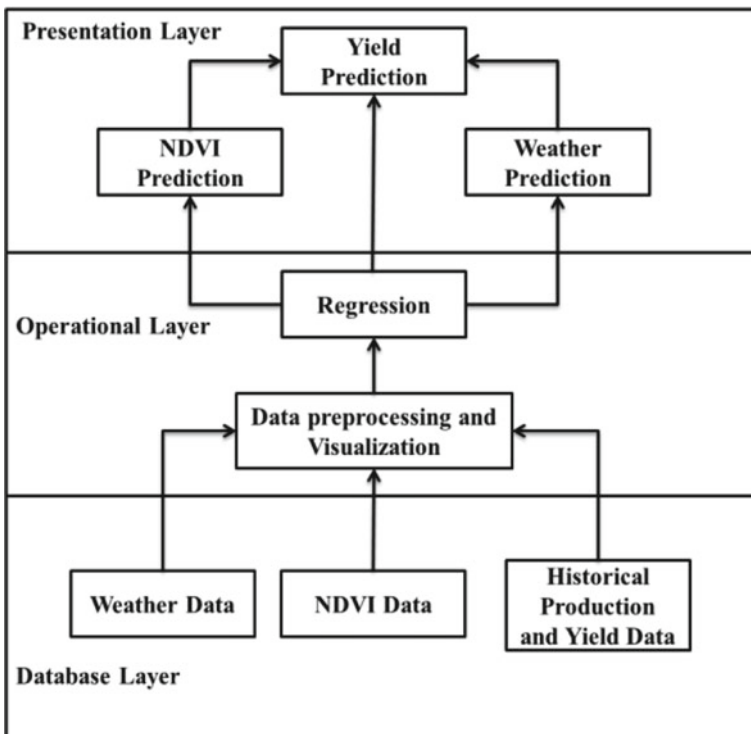


Fig. 3 Overall system architecture

4.1 Data Set Preprocessing

For the preprocessing of the data, we have used Pandas. The null values are removed, and the data set is converted to floating data type as the parameters such as precipitation, temperatures have floating values. Observing the correlation matrix, the attributes which are negatively related with each other are neglected. The Sentinel-2 satellite has a revisit time of five days; down sampling of the NDVI data is performed on the data set. This is required to match the samples with stage I and II inputs.

In supervised machine learning, the frequency of the data is maintained for each interval and then we can calculate ' t ' by using ' $t - 1$ ' and ' $t - 2$.' This approach of regression is used, and the weather parameters are predicted for next year with the given inputs. This is stage I output which is given to stage II as input parameter. In stage II, historical means NDVI along with the predicted weather parameters used by the algorithm to forecast the future NDVI values. This is stage II output. The output of both the stages is used by stage III to predict the sugarcane yield for the next season.

4.2 Training and Testing

The training of the data is performed by using a different number of training and testing samples. For the initial stage, each weather parameter is tested with regression algorithms such as lasso, DT, random forest, and KNN. For the later stages, the test size is selected as 0.1 and random state = 101. While for KNN, the value of n is tested for different inputs and is maximum for $n = 4$. The training samples for the last 20 years have varied from 70 to 90%.

After testing the accuracy with the input training samples, the predicted values of the next year are given as an input to get the desired output from the final stage.

5 Experimental Results and Analysis

The samples are tested, and results are analyzed for the prediction model. The temperature parameter the weather is predicted using K -nearest neighbor regression with the value of $n = 12$. This regressor gives maximum accuracy of 90% to minimum, maximum, and average temperature followed by Bayes with 88.03% and random forest with 86%, where random state = 42. For the precipitation, again KNN regressor gives maximum accuracy of 75.69% followed by Bayes with 69.72%. The value of the nearest neighbor is set to 1. For the specific and relative humidity, random forest and KNN are the best regressors with accuracy of 95 and 89.73%.

For the dew point, KNN gives the accuracy of 90% **followed by random forest with 88%**. KNN with neighbor = 12 is also best for wind prediction with accuracy

of 89.6% followed by Bayes with 86.513%. The above all listed weather parameters are also tested with different regression techniques like linear, lasso, ridge, DT, etc. The graphs in Fig. 4 show the comparison of different regressors.

After completion of stage I, the prediction of stage II is completed. In this phase, the KNN gives the best accuracy when the value of n is set to 4. The accuracy of KNN is 90% followed by the Bayes regressor with 86%. The final phase of the model is then tested and the decision tree gives the accuracy of 91.5%. The phase II and III are also tested with other algorithms such as linear, lasso, ridge, SVR for the comparison purpose. The comparison graphs are shown in Fig. 5.

The algorithms like linear, ridge, lasso regression have given very low accuracy for the prediction model. The total of 216 samples taken for the training of the model. The NDVI is surely dependent on the precipitation, dew point, temperature, etc. This is tested by changing the dependent variables.

For most of the samples and attributes in the data sets, the KNN is a suitable algorithm, in which context following outstanding training in feature space is graded.

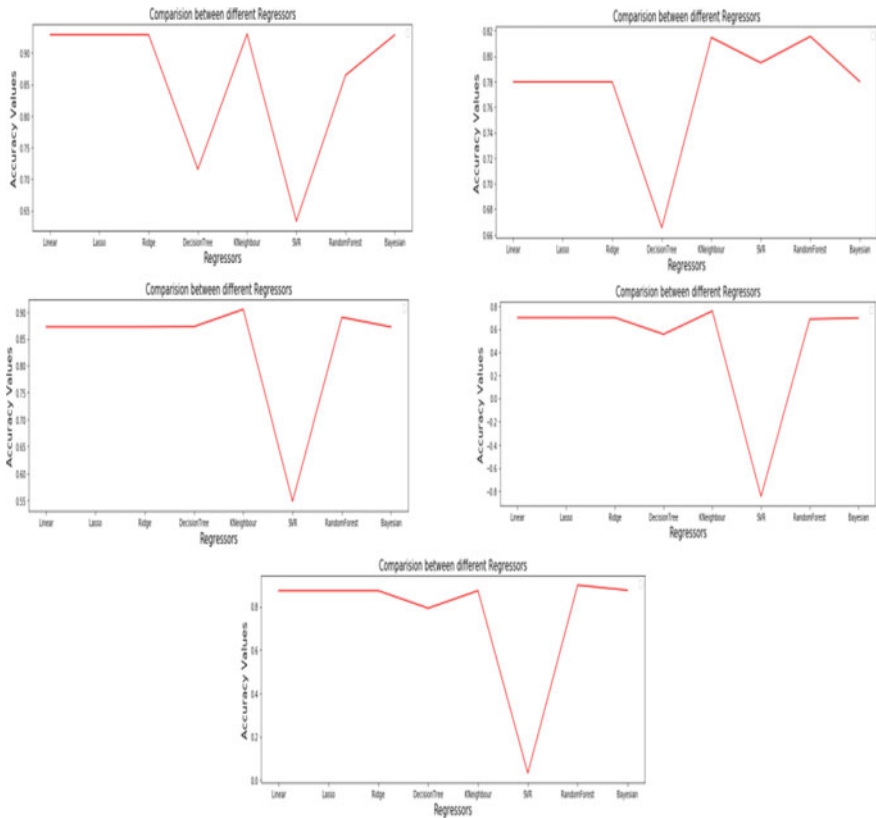


Fig. 4 Min temp, max temp, dew point, precipitation, relative humidity comparison graphs with different regressors

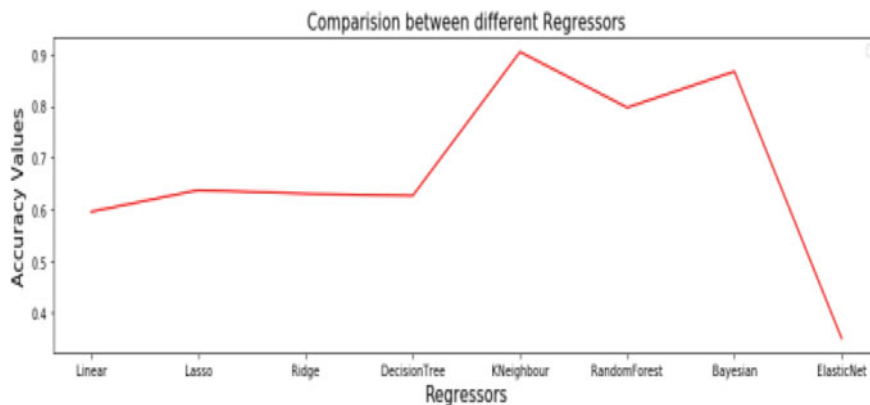


Fig. 5 Comparison of different regressors for NDVI prediction

Their role implicitly calculates the boundary of decision, as well as decision can also be explicitly calculated. The machine complexity of NN, therefore, depends on the complexity of the boundary. Neighbors are chosen for the right assignment from a collection of objects. No specific training steps are necessary, which can be considered as algorithm training. *KNN* algorithm is adaptive to the local data set structure. Higher values of k decrease noise on classification; however, it creates boundaries among classes less distinct. We have checked the best choice of k by testing various values for different training samples.

6 Conclusion

Traditionally, the crop yield is predicted using the ground weather parameters such as precipitation and temperature. In this paper, we have considered remote sensing data such as NDVI along with the ground parameters such as weather and soil data sets. So, from the experimental studies, we can conclude for most of the predictions *KNN* and random forest gives the better accuracy. Like *KNN* gives the accuracy of 90% for the temperature, 75.69% for the precipitation, 95% for the relative and specific humidity and 90% for the dew point. For the NDVI prediction, *KNN* gives 90% accuracy as well followed by Bayes regressor with 86.66%. While for the final yield prediction model, decision tree gives the maximum accuracy of 91.5%.

In this study, we have focused on the small area where the crop cycle is 12 month; we can try this for the crop cycle of 15–18 months as well. The NDVI shows significant results on the yield prediction for the sugarcane crop, so the different crop regions for different seasons such as Kharif and Rabi can also be useful to study using this remote sensing approach. Apart from NDVI, the other vegetation indices such as EVI, leaf area index, and leaf water content index can also be studied for the same purpose. Since *KNN* and DT show better overall performance, the different

crops and vegetation indices can be useful for the monitoring of the different crops and calculation of the yield. The accuracy obtained in this paper is more than that observed in [18].

References

1. T. Bhange, S. Shekapure, K. Pawar, H. Choudhari, Survey paper on prediction of crop yield and suitable crop. *Int. J. Innov. Res. Sci. Eng. Technol.* **8**(5), 5791–5795 (2019)
2. A. Kumar, N. Kumar, V. Vats, Efficient crop yield prediction using machine learning algorithms. *Int. Res. J. Eng. Technol. (IRJET)* **05**(06), 3151–3159 (2018). e-ISSN: 2395-0056
3. S. Bang, R. Bishnoi, A.S. Chauhan, A.K. Dixit, I. Chawla, in *Fuzzy Logic Based Crop Yield Prediction Using Temperature and Rainfall Parameters Predicted Through ARMA, SARIMA, and ARMAX Models*. 2019 Twelfth International Conference on Contemporary Computing (IC3), Noida, India, 2019, pp. 1–6. <https://doi.org/10.1109/IC3.2019.8844901>
4. M.A. Hossain, M.N. Uddin, M.A. Hossain, Y.M. Jang, in *Predicting Rice Yield for Bangladesh by Exploiting Weather Conditions*. 2017 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, 2017, pp. 589–594. <https://doi.org/10.1109/ICTC.2017.8191047>
5. T. Islam, T.A. Chisty, A. Chakrabarty, in *A Deep Neural Network Approach for Crop Selection and Yield Prediction in Bangladesh*. 2018 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), Malambe, Sri Lanka, 2018, pp. 1–6. <https://doi.org/10.1109/R10-HTC.2018.8629828>
6. A. Garg, B. Garg, in *A Robust and Novel Regression-Based Fuzzy Time Series Algorithm for Prediction of Rice Yield*. 2017 International Conference on Intelligent Communication and Computational Techniques (ICCT), Jaipur, 2017, pp. 48–54. <https://doi.org/10.1109/INT ELCCT.2017.8324019>
7. N.G. Hegde, S. Mujumdar, P.R. Rajath Navada, S.S. Jambarmath, R.P. Madhavi, Survey paper on agriculture yield prediction tool using machine learning. *Int. J. Adv. Res. Comput. Sci. Manag. Stud.* **5**(11), pp. 36–39 (2017)
8. X. Teng, Y. Gong, Research on application of machine learning in data mining. *IOP Conf. Ser. Mater. Sci. Eng.* **392**(062s202), 1–5 (2018). <https://doi.org/10.1088/1757-899X/392/6/062202>
9. W. Zhuo et al., in *Assimilating SAR and Optical Remote Sensing Data into WOFOST Model for Improving Winter Wheat Yield Estimation*. 2018 7th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Hangzhou, 2018, pp. 1–5. <https://doi.org/10.1109/Agro-Geoinformatics.2018.8476074>
10. H. Aghighi, M. Azadbakht, D. Ashourloo, H.S. Shahrabi, S. Radiom, Machine learning regression techniques for the silage maize yield prediction using time-series images of landsat 8 OLI. *IEEE J. Sel. Topics Appl. Earth Observations Remote Sens.* **11**(12), 4563–4577 (2018). <https://doi.org/10.1109/JSTARS.2018.2823361>
11. X. Huang, J. Liu, C. Atzberger, Q. Liu, in *Research on the Optimal Thresholds for Crop Start and End of Season Retrieval from Remotely Sensed Time-Series Data Based on Ground Observations*. IGARSS 2018—2018 IEEE International Geoscience and Remote Sensing Symposium, Valencia, 2018, pp. 7727–7730. <https://doi.org/10.1109/IGARSS.2018.8519031>
12. I. Ahmad, U. Saeed, M. Fahad et al., Yield forecasting of spring maize using remote sensing and crop modeling in Faisalabad-Punjab Pakistan. *J. Indian Soc. Remote Sens.* **46**, 1701–1711 (2018). <https://doi.org/10.1007/s12524-018-0825-8>
13. N. More, V.B. Nikam, B. Banerjee, Machine learning on high performance computing for urban greenspace change detection: satellite image data fusion approach. *Int. J. Image Data Fusion* (2020). <https://doi.org/10.1080/19479832.2020.1749142>

14. N.P. More, V.B. Nikam, S.S. Sen, in *Experimental Survey of Geospatial Big Data Platforms*. 2018 IEEE 25th International Conference on High Performance Computing Workshops (HiPCW), 2018, December (IEEE), pp. 137–143
15. R.A. Medar, V.S. Rajpurohit, A.M. Ambekar, Sugarcane crop yield forecasting model using supervised machine learning. *Int. J. Intell. Syst. Appl. (IJISA)* **11**(8), 11–20 (2019). <https://doi.org/10.5815/ijisa.2019.08.02>
16. P. Bose, N.K. Kasabov, L. Bruzzone, R.N. Hartono, Spiking neural networks for crop yield estimation based on spatio temporal analysis of image time series. *IEEE Trans. Geosci. Remote Sens.* **54**(11) (2016)
17. R. Luciani, G. Laneve, M. JahJah, Agricultural monitoring, an automatic procedure for crop mapping and yield estimation: the great rift valley of Kenya case. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* **12**(7) (2019)
18. V. Nathgosavi, A survey on crop yield prediction using machine learning. *Turk. J. Comput. Math. Educ. (TURCOMAT)* **12**(13), 2343–2347 (2021)

Automated Multiple-Choice Question Creation Using Synonymization and Factual Confirmation



M. Pranav, Gerard Deepak, and A. Santhanavijayan

Abstract E-assessment is the method of end-to-end automated evaluation in which cyber-physical technology is used to present evaluation events and document responses. This perspective requires a full-length assessment mechanism for teachers, tutors, educational agencies, granting institutions and regulators, and the public at large. Various methodologies are implemented for increasing the quality of the assessment through automation of question generation through machine learning and deep learning concepts. Though there are various amounts of contributions toward this challenge, an efficient and a stable method has not been yet formulated. This paper introduces a process that can help to make a contribution decreasing the intensity of the challenge by using abstractive LSTM series to model sequence summarization with attention mechanism for summarization along with key generation using synonymizing with factual validation and knowledge-centric distractor generation using Wiki data for the MCQ generation. The performance of the proposed model is measured and compared to baseline models, and the proposed solution was found to be superior in terms of performance with the findings of 94.87%, 96.74%, 95.84%, and 0.03%, respectively, of precision, recall, accuracy, *F*-measure, and false negative rate.

Keywords E-assessment · LSTM · Machine learning · MCQ · Wiki data

1 Introduction

Learning assessment aims to facilitate the learning process by presenting appraisal data to teachers and students that can educate and direct teaching. The learning appraisal is the mechanism by which learners and their teachers seek and analyze

M. Pranav

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, India

G. Deepak (✉) · A. Santhanavijayan

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

facts to establish where the learners are with their learning, where they need to go, and how best to get there. An efficient and reliable strategy for assessments will be capable enough to test the learner's ability to learn to achieve the different qualities of the domain in question. The challenging problems that promote analytical and contextual thinking play a primary role in the appraisal. In modern years, with the advancement of learning systems, there has been a significant evolution of the learning pattern. E-learning is a web-based portal that helps learners to learn about something without focusing on spatial distances and time. The automation in the field of e-assessment through automatic question generation will create a great impact on the faster evaluation process of e-learning.

Motivation: Recent COVID-19 pandemic situation shows as a great example where the process of education all over the world took place despite the grave situations on e-learning platforms. The traditional barriers in the education system have been eliminated by introducing an e-learning education system. The e-assessment plays a major role in the evaluation of the students' progress in a day-to-day learning process. Multiple-choice questions are the most common method that is used in the evaluation process. Attentiveness, cognition of the topic given, reasoning along scrutiny are the major requirement for these types of questions for the process of elimination. Automated question generations have great importance in the e-assessment process as it helps to save a lot of time and man efforts put into the process of manual question generation. Using automatic query generation, the goal of generating tough questions has opened the way for research in the field of e-learning and assessment. Hence, there are a lot of standards that have to be maintained to achieve challenging question generation.

Contribution: A various number of researches have been made in the field of improvising the standard of the questions, using methods of natural language processing (NLP), recurrent neural network (RNN), convolution neural network (CNN), long short-term memory (LSTM) algorithms, etc., for the processes like summarization, question framing, key generation and the creation of distractors. The distractor is one of the most primary standards that have to be improvised in the automated question generation. Thus, improving the quality of the distractors will result in an increased standard of the questions generated. An innovative and solitary anatomy for automated question formation of multiple-choice e-assessment questions has been proposed. An idiosyncratic method for the creation of distraction has been formulated for distractor generation. The other distinct contributions include the utilization of CHiC heritage data sets, factually validated synonymized key generation. The accuracy, recall, and precision for synthesis, key generation, and distraction were enhanced in a better way.

Organization: This paper has the following form. Section 2 reviews recent query generation literature. The pipeline of our proposed system is laid out in Sect. 3. The implementation and performance evaluation is depicted in Sect. 4. The paper is concluded in Sect. 5.

2 Related Work

Santhanavijayan et al. [1] used the principle of the automatic ontology using multi-swarm optimization based on the expectations of the learning algorithm for summarizing the derived text using the unsupervised quick reduct algorithm. The statistical sequence algorithm is used for the generation of keys. Deepak et al. [2] have built a model called OntoQuest for the generation of MCQs from various crawled web corpus and uses domain and granular ontologies for synonymization and anonymization of key and distractor generation and WordNet incorporation for enhanced accuracy. Srivastava et al. [3] have developed a Questionator model that uses deep learning algorithms to produce automated questions. Here, the model uses CNN as an encoder to extract information from the image and the LSTM algorithm as a decoder that transforms the extracted features into a natural language that is further used to generate questions that chatbots can query for. Gumaste et al. [4] have used the Stanford tagger to mark sentences with a gated encoder to overcome long-term language text input processing problems for query generation. Scialom et al. [5] use self-attention architectures for the generation of questions. The author uses a sequence-to-sequence paradigm of symmetrical encoder and decoder based on a self-care mechanism for the generation of questions. The model for dialog and the generation of transient questions was introduced by Pan et al. [6]. A question-and-answer dialog of the type was performed by the model suggested. It is planned to use the ReDR technique and the CoQA dataset for implementation. The automated fuzzy MCS question generation technique, hybridized with a modified sequestration algorithm, is proposed by Santhanavijayan et al. [7]. The approach uses ontology, so it is possible to boost overall output ontologically by eliminated optimization algorithms to make it less computerized. In [8–22], several ontology-driven approaches in support of the proposed literature have been depicted.

3 Proposed Architecture

The design of the proposed device structure for automated query generation is seen in Fig. 1. The proposed system consists of three phases, the first phase so user entry of topic of choice and summarization, the second phase will be the question generation and key generation, and the final phase will be the distractor generation and MCQ formation with a final review.

The main aim of this method is to improvise the key and distractor generation, rendering MCQs difficult for evaluation. This method uses a special layer for e-assessment by creating multiple-choice questions from the CHIC heritage dataset. The CHIC Patrimonial Knowledge Set (CHIC), the CLEF Cultural Heritage Data Set (CHiC) is a record collection developed by the Pledge FP7 Network of Excellence for the CLEF Culture Heritage Assessment Lab of the CLEF Initiative. The CLEF Project (Conference and Labs of the Appraisal Forum, formerly known as the

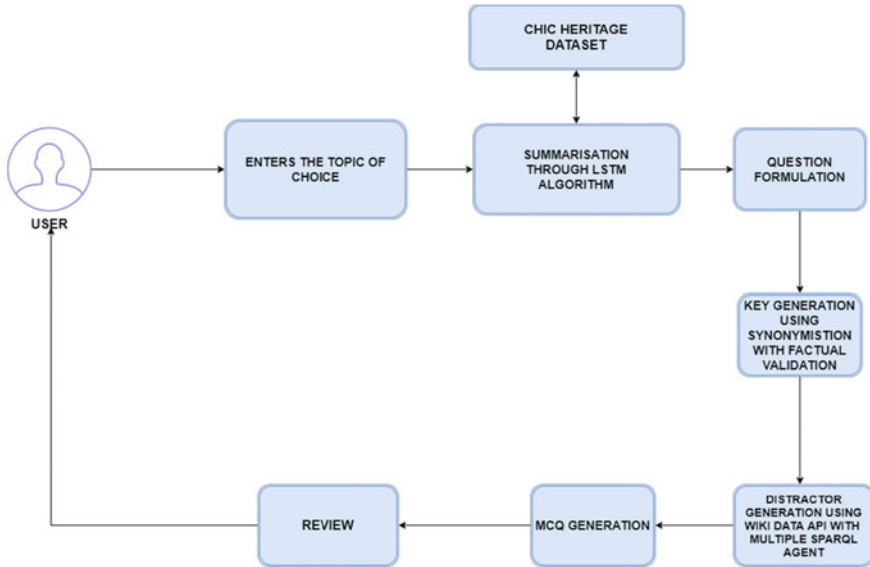


Fig. 1 System architecture

Cross-Language Evaluation Forum) is a self-organized organization whose primary mission is to encourage research, innovation, and the development of knowledge access systems, with an emphasis on multilingual and multimodal information at various levels of structure. The input provided by the user and the first corresponding domain and related sub-domains and processed as subjects and auxiliary themes of the data set documents containing all relevant sub-topics and auxiliary themes shall be extracted from the data set using the LSTM classification algorithm.

The LSTM is a type of RNN architecture. The LSTM classification model is first used for the classification of the topics from the Chic heritage data set which is then summarized using Abstractive LSTM sequence-to-sequence summarization model with Attention Mechanism. The summarization process involves text cleaning, contraction, plotting the distribution of words, data split up, tokenizing, sequence padding, encoder, and decoder along with the training of the model. Post summarization the important sentences act as the markers for generating questions.

For the question generation a pre-trained template consisting of the appending word like who what, where, why, how, etc., can be utilized from the trained Bloom's taxonomy corpus. Once the questions are formulated the key for the MCQs are generated with the same keyword or to use the synonym of the word using synonymizing algorithm along with factual validation for further accuracy. If the synonymizing process is used. The synonyms are then created from the same reference domain and the synonym is validated for the fact that the given term belongs to the defined domain.

Once the keys are generated for the MCQ, the process of producing the distractors takes place. For this method, the produced key is forwarded and the words nearest to the key are generated, but the key is not generated using the Wiki data API and multiple SPARQL agents. SPARQL is an RDF query language that is a semantic query language for databases that can be used to retrieve and manage data in a resource creation environment. SPARQL allows users to write queries against what may be technically considered “key-value” data or, more precisely, data that meets the W3C RDF standard.

Finally, the MCQ’s are created in the specified format along with the options containing keys and distractors and further, the options for each question are randomized and the questions are arranged in order of difficulty using the random function and is presented to the user for review.

4 Implementation and Performance Evaluation

The proposed system has been successfully designed and implemented on a windows 10 platform. Intel Core i7 8th generation processor with 16 GB of RAM was employed for this implementation. The complete system is programmed using Python running on Google-colab. Keras, a high-level deep learning library that uses Tensor-Flow as a backend, was used to design and train the LSTM model. NLTK a python NLP library has been imbibed for text preprocessing and tokenizing.

The experiment was performed on the English language files of the CHiC heritage dataset. The selected topic is first cleaned and preprocessed in the process of summarization. An embedding layer of 110 dimensions is added to embed our text and further is encoded with a 200 latent dimension with three LSTM layers. Finally, an LSTM layer is used for decoding with 200 latent dimensions, and an attention mechanism is used to obtain a summarized text. The SPARQL-dependent agent is used to query information, which has a state and actions. The condition of the agent is to aggregate similar information from the knowledge base, while the insightful knowledge is given by the actions. The knowledge base that was used for this is the knowledge base for Wiki data. This is used for distractor generation. The percentages of precision, recall, accuracy, F-measure, and false negative rate are calculated and used as performance evaluation metrics. The accuracy, *F*-measure, and false negative rate are computed using Eqs. (1–5), respectively.

$$\text{Precision \%} = \frac{\text{True number of Positives}}{\text{True number of positives} + \text{false number of positives}} \quad (1)$$

$$\text{Recall \%} = \frac{\text{True number of Positives}}{\text{True number of Positive} + \text{False number of Negatives}} \quad (2)$$

$$\text{Accuracy \%} = \frac{\text{Precision} + \text{Recall}}{2} \quad (3)$$

$$F\text{-Measure } \% = \frac{2(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (4)$$

$$\text{False Negative Rate} = 1 - \text{Recall} \quad (5)$$

The experiment was trained on several metrics. For this experimentation purpose, five events are considered. Figure 2 shows the performance analysis for various metrics, namely summarization, key generation, distractor generation, question and multiple-choice creation. It can be observed that the precision, recall, accuracy, F -measure, and false negative rate for summarization are 93.89%, 94.89%, 94.14%, 94.39%, and 0.51%, respectively; for key generation, it was recorded 95.38%, 97.33%, 96.81%, 96.35%, and 0.027%, respectively; in the case of distractor generation, it was 95.81%, 97.47%, 96.84%, 96.63%, and 0.025%, respectively, and Finally, for question and multiple-choice creation, the values obtained were 94.87%, 96.78%, 95.84%, 95.82%, and 0.032%, respectively.

Different comparisons of the method were made, and the outcome interpretation was prepared as shown in Fig. 3. The study reveals that firefly algorithm and particle swarm optimization were ineffective compared to the multi-swarm optimization hybrid model with the MCQ generated statistic pattern algorithm [1] and random forest synonymization with the use of lexical standards with WordNet models, where there is total utility combined with more than one algorithm proves higher efficiency.

The superiority of the proposed model concerning the other models in terms of results lies in using sequence-to-sequence LSTM algorithm for summarization where the text is summarized in a much more efficient way by making the system understand the complexity in a sequence and simplify them. Also, the use of factual confirmation for key generation and multiple SPARQL APIs for distractor generation increases the overall accuracy with comparison to the other models. In Fig. 4, it can be seen that the FNR values of different baseline approaches compared to the proposed approach. It can be noted that the FNR value of the proposed approach has a comparatively less FNR value than those of the baseline approaches hence proving the proposed concept much efficient.

5 Conclusion

Assessment is today's means of modifying tomorrow's learning. Hence, proper assessment methods will play a major role in a greater learning process. This proposed method can perform accurate predictions when compared to others because of the hybrid approach of using synonymizing with factual validation for the key generation process and a distinct knowledge-centric distractor generation using Wiki data API. The proposed system also produces a more accurate result in comparison to other models. Further, we can conclude that in the future, more enhanced techniques with hybrid approaches combined with this unique framework have a higher chance of

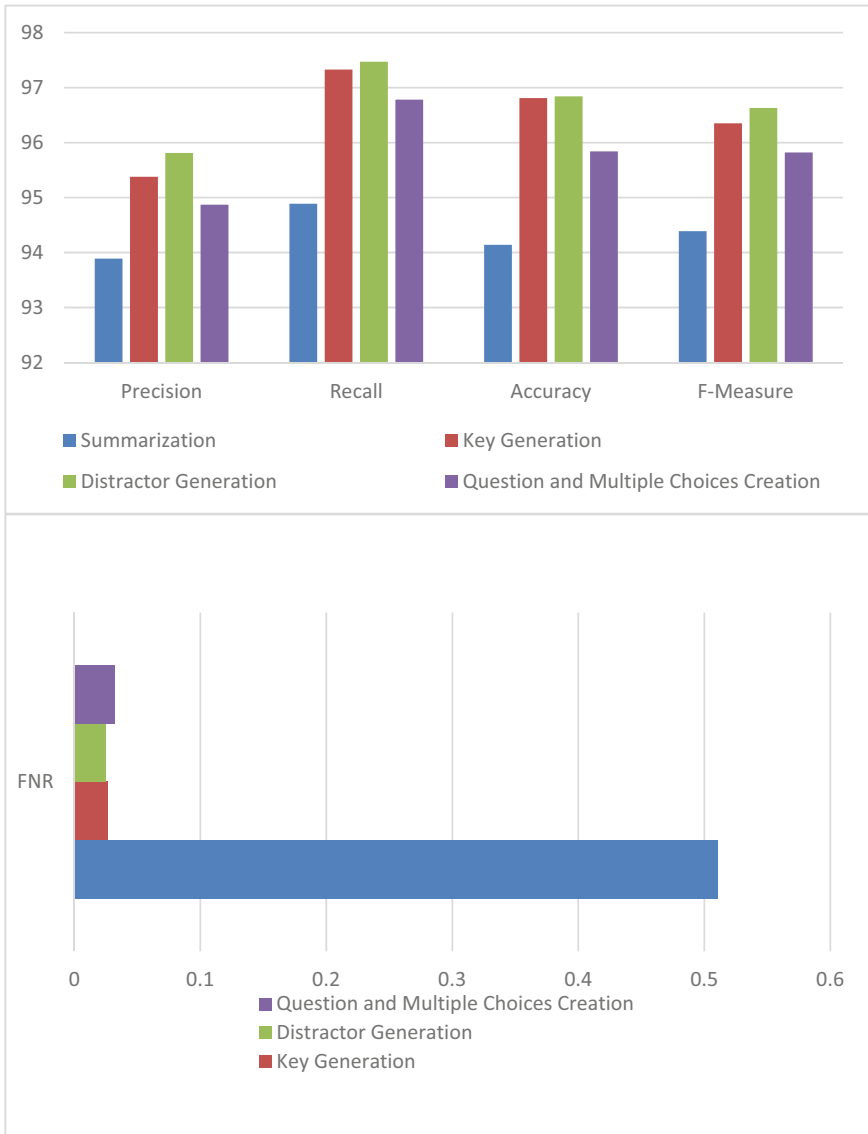


Fig. 2 Performance of the proposed approach for constituent phases for the CHiC dataset

attaining better results for the same purpose fulfillment. An average F -measure of 95.80% has been achieved with a very low FNR of 0.03 has been achieved.

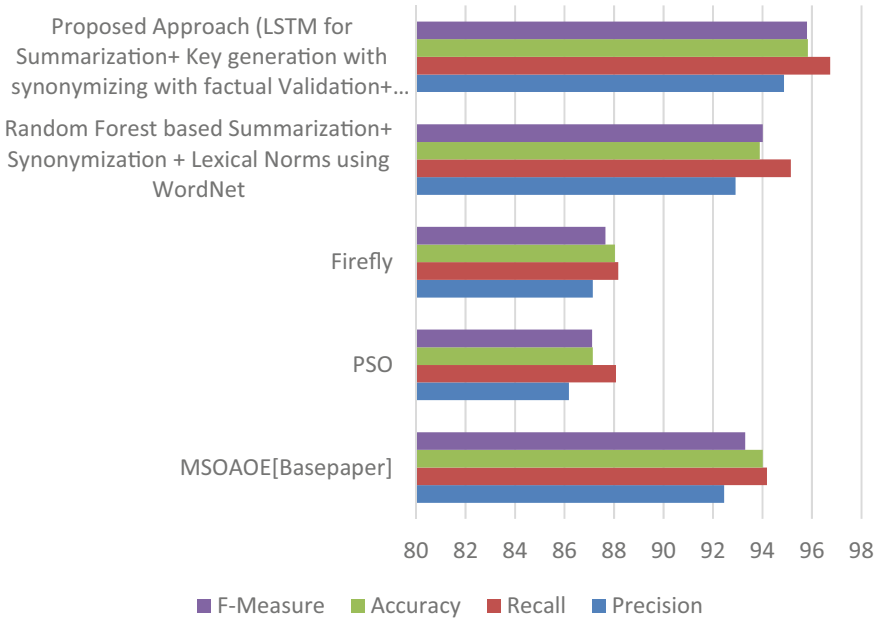


Fig. 3 Performance comparison of the proposed hybrid approach with baseline approaches

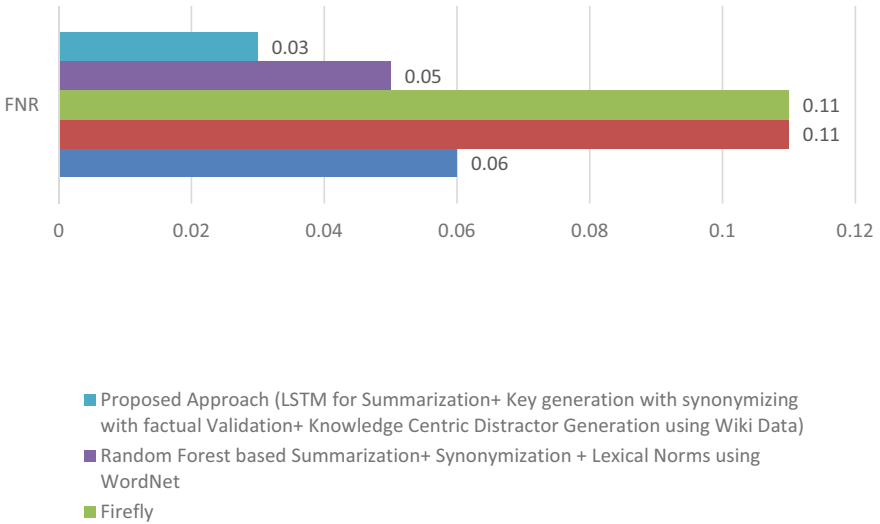


Fig. 4 FNR value comparison of the proposed hybrid approach with baseline approaches

References

1. A. Santhanavijayan, S.R. Balasundaram, Multi swarm optimization based automatic ontology for e-assessment. *Comput. Netw.* **160**, 192–199 (2019)
2. G. Deepak, N. Kumar, G.V.S.Y. Bharadwaj, A. Santhanavijayan, in *OntoQuest: An Ontological Strategy for Automatic Question Generation for e-Assessment Using Static and Dynamic Knowledge*. 2019 Fifteenth International Conference on Information Processing (ICINPRO) (IEEE, 2019 December), pp. 1–6
3. A. Srivastava, S. Shinde, N. Patel, S. Deshpande, A. Dalvi, S. Tripathi, in *Questionator-Automated Question Generation Using Deep Learning*. 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) (IEEE, 2020 February), pp. 1–5
4. M.P. Gumaste, M.S. Joshi, M.S. Khadpekar, S.R. Mali, F.Y. Be, *Automated Question Generator System: A Review*
5. T. Scialom, B. Piwowarski, J. Staiano, in *Self-Attention Architectures for Answer-Agnostic Neural Question Generation*. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (2019 July), pp. 6027–6032
6. B. Pan, H. Li, Z. Yao, D. Cai, H. Sun, Reinforced dynamic reasoning for conversational question generation. arXiv preprint arXiv: 1907.12667 (2019)
7. A. Santhanavijayan, S.R. Balasundaram, Fuzzy-MCS algorithm-based ontology generation for e-assessment. *Int. J. Bus. Intell. Data Min.* **14**(4), 458–472 (2019).
8. G. Deepak, V. Teja, A. Santhanavijayan, A novel firefly driven scheme for resume parsing and matching based on entity linking paradigm. *J. Discrete Math. Sci. Crypt.* **23**(1), 157–165 (2020)
9. V. Adithya, G. Deepak, A. Santhanavijayan, in *HCODF: Hybrid Cognitive Ontology Driven Framework for Socially Relevant News Validation*. International Conference on Digital Technologies and Applications (Springer, Cham, 2021), pp. 731–739
10. G.L. Giri, G. Deepak, S.H. Manjula, K.R. Venugopal, in *OntoYield: A Semantic Approach for Context-Based Ontology Recommendation Based on Structure Preservation*. Proceedings of International Conference on Computational Intelligence and Data Engineering: ICCIDE 2017, vol. 9 (Springer, 2017, December), p. 265
11. G. Deepak, N. Kumar, A. Santhanavijayan, A semantic approach for entity linking by diverse knowledge integration incorporating role-based chunking. *Procedia Comput. Sci.* **167**, 737–746 (2020)
12. G. Deepak, S. Rooban, A. Santhanavijayan, A knowledge centric hybrid-ized approach for crime classification incorporating deep bi-LSTM neural network. *Multimedia Tools Appl.* 1–25 (2021)
13. K. Vishal, G. Deepak, A. Santhanavijayan, in *An Approach for Retrieval of Text Documents by Hybridizing Structural Topic Modeling and Pointwise Mutual In-formation*. Innovations in Electrical and Electronic Engineering (Springer, Singapore, 2021), pp. 969–977
14. K. Shreyas, G. Deepak, A. Santhanavijayan, GenMOnto: a strategic domain ontology modelling approach for conceptualisation and evaluation of collective knowledge for mapping genomes. *J. Stat. Manag. Syst.* **23**(2), 445–452 (2020)
15. G. Deepak, A. Santhanavijayan, OntoBestFit: a best-fit occurrence estimation strategy for RDF driven faceted semantic search. *Comput. Commun.* **160**, 284–298 (2020)
16. M. Arulmozhivarman, G. Deepak, in *OWLW: Ontology Focused User Centric Architecture for Web Service Recommendation Based on LSTM and Whale Optimization*. European, Asian, Middle Eastern, North African Conference on Management & Information Systems (Springer, Cham, 2021 March), pp. 334–344
17. G. Deepak, J.S. Priyadarshini, Personalized and enhanced hybridized semantic algorithm for web image retrieval incorporating ontology classification, strategic query expansion, and content-based analysis. *Comput. Electr. Eng.* **72**, 14–25 (2018)
18. Z. Gulzar, A.A. Leema, G. Deepak, Pcrs: personalized course recommender system based on hybrid approach. *Procedia Comput. Sci.* **125**, 518–524 (2018)

19. G. Deepak, J.S. Priyadarshini, in *A Hybrid Semantic Algorithm for Web Image Retrieval Incorporating Ontology Classification and User-Driven Query Expansion*. Advances in Big Data and Cloud Computing (Springer, Singapore, 2018), pp. 41–49
20. N. Krishnan, G. Deepak, in *KnowSum: Knowledge Inclusive Approach for Text Summarization Using Semantic Alignment*. 2021 7th International Conference on Web Research (ICWR) (IEEE, 2021 May), pp. 227–231
21. N. Roopak, G. Deepak, A. Santhanavijayan, in *HCRDL: A Hybridized Approach for Course Recommendation Using Deep Learning*. International Conference on Intelligent Systems Design and Applications (Springer, Cham, 2020 December), pp. 1105–1113
22. S. Gadamshetti, G. Deepak, A. Santhanavijayan, in *LWOntoRec: Light Weight Ontology Based Novel Diversified Tag Aware Song Recommendation System*. International Conference on Intelligent Systems Design and Applications (Springer, Cham, 2020 December), pp. 743–752

EASDisco: Toward a Novel Framework for Web Service Discovery Using Ontology Matching and Genetic Algorithm



N. Krishnan and Gerard Deepak

Abstract Web services are gradually elevating as a fundamental aspect of Web applications in the era of Web 3.0. A Web service can be termed as a strategic model curated for reinforcing concordant machine-to-machine interactivity over a network. As there is a gradual transfer toward service-oriented architecture, the importance of service-based computing has turned out to be exceptionally popular. It has become a major asset in an aspect of communication within the Internet. This paper proposes an ontology-based Web service recommendation system that uses ontology matching and collective crowdsourced ontology along with a genetic algorithm for optimization. The dataset is used for training followed with classification and computing semantic similarity using the genetic algorithm which recommends the services in increasing order of similarity. The proposed approach is superior in terms of performance and recorded precision and accuracy of 96.79 and 95.39% which is found to be better than existing approaches.

Keywords Genetic algorithm · Ontology · Semantic similarity · Semantic Web text summarization

1 Introduction

With the knowledge overload and a growing need to provide relevant suggestions to consumers, Web service recommendation in such an effective and reliable way has become a significant tool. A Web service is a software technology that provides interoperable machine-to-machine communication over the internet [1]. The modern Web not only serves as a platform for static knowledge but also as an interface for Web-accessible applications, ranging from basic dynamically generated pages for

N. Krishnan

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, India

G. Deepak (✉)

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

displaying works to more sophisticated services such as building an ecosystem to buy and sell products [2].

The modern Web not only serves as a platform for static knowledge but also as an interface for Web-accessible applications, ranging from basic dynamically generated pages for displaying works to more sophisticated services such as building an ecosystem to buy and sell products. As many applications are built using various programming languages, the channel of communication between these applications is not reliable. Web services offer a shared interface or a medium for the communication of these heterogeneous apps which are built using different programming languages for different test cases.

Motivation: There is an exponential increase in the number of available Web services so browsing through all the available services and selecting the best service becomes a tedious and time-consuming task for the user. As a result, among service users, the perspective of choosing Web service becomes a significant and difficult challenge. Users would be confused by the service providers given the multitude of services with similar or equivalent functionalities. Although Web service exploration alone could not solve this issue, promising strategies to Web service availability and recommendation, which is a core subject in the area of service computing, are much more important.

To solve this problem, an ontology-based Web service recommendation system can be implemented for efficient and accurate recommendations by reducing a lot of time for execution. Extensive research has been performed on Web service discovery, and many architectures have been proposed, but most of these systems uses conventional methods and lacks hybridization, and some does not consider real-world knowledge. This paper proposes an ontology-based Web service recommendation model that encompasses hybridization and can give practical and logical solutions using real-world knowledge.

Contribution: Input query has been obtained from the user and is subjected to query preprocessing such that query words are obtained. Ontology matching is done using SemantoSim Measure by formulating ontologies in Web services thesaurus. For the obtained terms, entity aggregation is done using Google's Knowledge Base API. The Web service dataset is classified using gated recurrent unit. For the top 25% of the recommended services, semantic similarity is computed and recommended services in increasing order of semantic similarity. The precision and accuracy of the proposed system were found out to be 96.79% and 95.39%, respectively.

Organization: The remainder of the paper is arranged as follows. Section 2 contains pertinent research that has already been done on the subject. The proposed architecture is found in Sect. 3. Section 4 discusses implementation. Section 5 comprises performance evaluations and observed results. The paper is concluded in Sect. 6.

2 Related Works

Deepak et al. [3] came up with a novel system for the recommendation of Web services. The system constructs a Web dataset from popular Web sites. The normalized pointwise mutual information (NPMI) value is validated, and a library is designed using ontology clusters. The system compares services and user requests to accurately recommend a Web service. Ren et al. [4] describe a method for recommending Web services using support vector machine (SVM). The system uses a CF method that obtains a hyperplane from the information that can refine the Web services based on user interest. The user interest is calculated by the hyperplane. The system provides efficient and accurate results because quality of service (QoS) values are not required.

Ma et al. [1] have proposed a novel method for Web service recommendation. The system uses lexical inspection combined with grammar examination to provide an accurate recommendation model. Semantic Web methods are employed that increases the system's recall ratio and accuracy. Jiang et al. [2] have introduced a system that models the recommendation of Web services. The system is different from PCC measurement and utilizes the PHCF method. It calculates similarity reading and uses personalized algorithms. Chen et al. [5] have proposed a system for an efficient recommendation of Web services. Region K-nearest neighbor (KNN) uses a hybrid approach that is made for processing a large amount of data. Features of QoS are considered to construct a model. The recommendations are obtained by a memory focused approach which has proven accurate. Xia et al. [6] came up with a novel system for recommending Web services. The system evaluates the similarity reading to recommend services depending on user requirements. The summary of requests and services is compared with the help of an ontology model. The system provides the users with services that match their given interests.

Karthikeyan et al. [7] have described a method for the recommendation of cloud services. The system reads the features of cloud service using the NLP method. The user requests are improved with DNF, and fuzzy nodes are modified. This method provides an accurate model after experimental validation. Feier et al. [8] have delivered an approach for modeling Web services. The system helps to make the available data on the Web be evaluated by machines. WSMO also enables services to utilize the data on the Web. It employs ontology to enhance methods to carry out Web services. Rupasingha et al. [9] has proposed a system for enhancing recommendation service. The system uses a CF method to eliminate the lack of data. Ontology modeling is employed to obtain the correlation of requesters. It lowers fallacy in prediction and gives trust in user inputs. Costa et al. [10] have introduced a system for the recommendation of services. The system works together with context-based systems for accurate modeling of recommendations. CORES employs Infraware recommender to obtain more accurate recommendations. The system eliminates the problem resulting from a vast data archive using ontology modeling. In [11–20], several ontological models have been discussed in support of the proposed approach.

3 Proposed Architecture

The architecture of the EASDisco framework for Web services recommendation is divided into two phases as depicted in Fig. 1. In phase 1, the user feeds in the input query. The query imputed by the user is preprocessed using tokenization, lemmatization, stop word removal, and named entity recognition which is also performed, and query words are obtained. Ontology formulation is done using Web service thesaurus.

Web service thesaurus contains Web service providers such as Amazon Web services where they publish all the services they provide, so that service seekers can use the services from them. They contain WSDL files; it stands for Web services description language which contains the functionality of the Web service provided. They contain a UDDI registry; it stands for universal description, discovery, and integration. They also contain a Web service repository; it is a log where you can see all the available services in the network.

After obtaining all the terms, it is combined to form a Web service thesaurus, and the words are indexed. From the indexed words, ontology is formed using Ontocolab. The generated ontology is matched with the initially obtained query words set. That is for each entity in the query set, ontology matching is done using SemantoSim measure. With the terms matched with the ontology, entity aggregation is done with the help of Google Knowledge Base API. This will help add more similar terms to the existing ontology terms and is depicted as Phase 2 of EASDisco (Fig. 2).

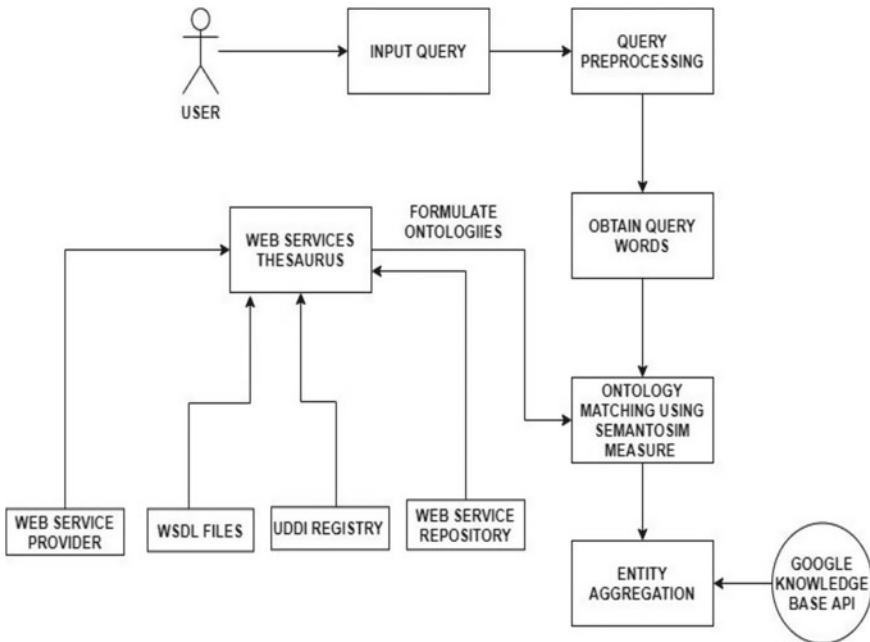


Fig. 1 Phase 1 architecture diagram of the proposed EASDisco system

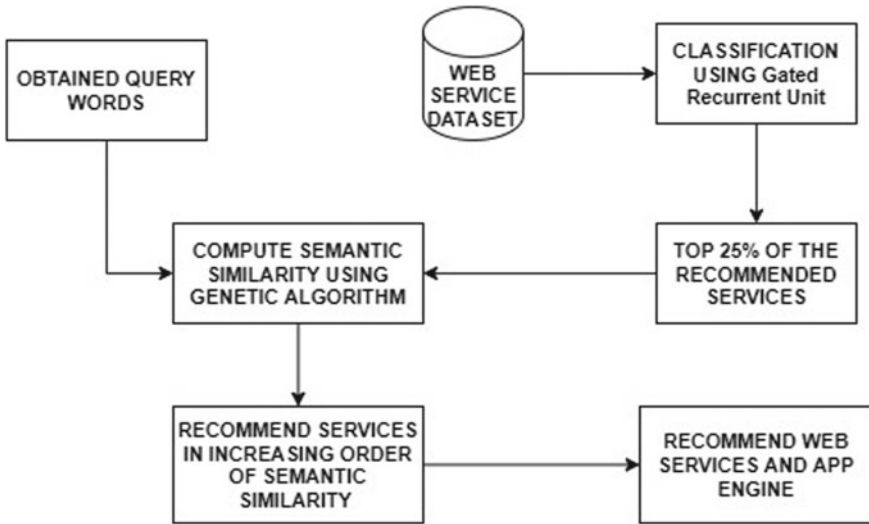


Fig. 2 Phase 2 architecture diagram of the proposed EASDisco system

In Phase 2, Web service dataset is taken. Classification is done using a gated recurrent unit. From the dataset, we can extract features using information entropy. Here, only top 25% of the data is considered from the recommended services. From the obtained query words from the previous phase and the top 25% of the recommended services, semantic similarity is calculated using a genetic algorithm. Here, each recommended service is considered as a population. The fitness function will determine the relevance between the recommended services and the query words. The selection process is done according to increasing order of semantic similarity for recommending services. The most relevant Web services are proposed along with the app engine which is required to run the specified recommended Web service.

4 Implementation

The proposed EASDisco architecture for Web service recommendation was designed and implemented on Windows 10 OS. The implementation has been done using Intel Core-i7 10th generation processor and a RAM of 16 GB. NLTK and WordNet 3.0 have been utilized to obtain the meanings of the words and to find their synonym, and to discover the root word, WordNetLemmatizer has been employed. Re and sklearn are the Python’s NLP tool packages that have been employed for preprocessing data. Keras, a Python deep learning API package running on TensorFlow, was utilized for implementing and training the GRU model.

5 Performance Evaluation and Result

The precision, recall, accuracy, *F*-measure, false discovery rate (FDR), and normalized discounted cumulative gain (nDCG) are computed and utilized as performance evaluation metrics. The accuracy, *F*-measure, FDR, and nDCG are calculated and plotted as Figs. 3 and 4. Standard formulations are used for computing the precision, recall, accuracy, *F*-measure, FDR, and nDCG. Precision, recall, accuracy, and *F*-measure quantifies the relevance of results. FDR quantifies the false positives, and nDCG measures the diversity in the results.

The proposed approaches performance was recorded and juxtaposed with that of the baseline approaches and benchmark models and have plotted the graph accordingly as shown in Fig. 3. Performance metrics that are marked on the *x*-axis are precision, recall, accuracy, *F*-measure, FDR, and nDCG, and the models in the *Y*-axis are proposed EASDisco, SVM + Collaborative Filtering [4], OntoDisco [3], WS_SSE [1], Eliminating Entity Aggregation from EASDisco, Eliminating XGBoost from EASDisco, Eliminating PSO from EASDisco, and Eliminating Ontology Matching from EASDisco. The average precision, average recall, *F*-measure, FDR, and nDCG of the proposed EASDisco approach were found to be 96.79%, 94.32%, 95.39%,

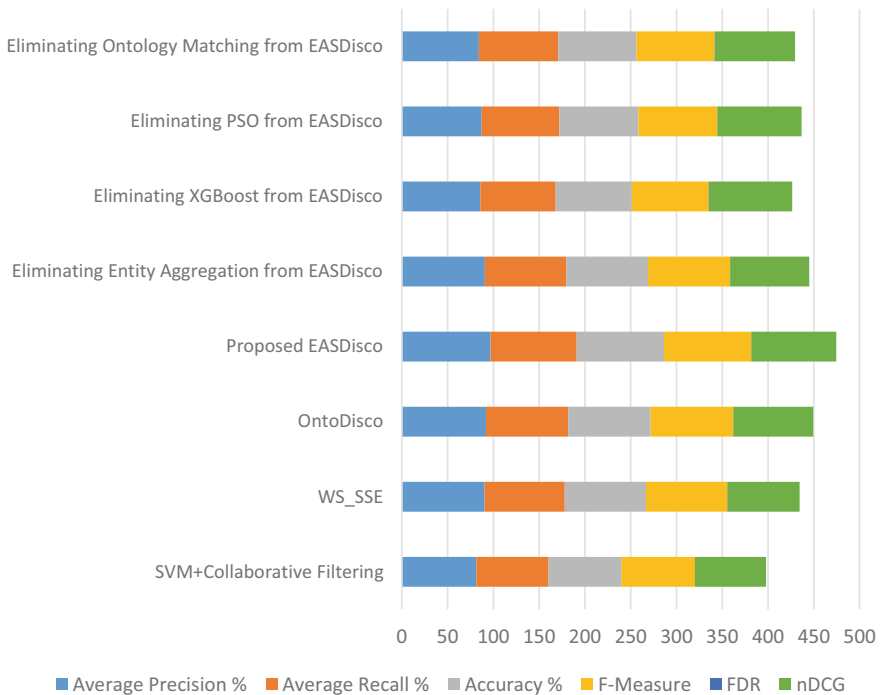


Fig. 3 Performance metrics versus percentage of performance measures of the proposed architecture

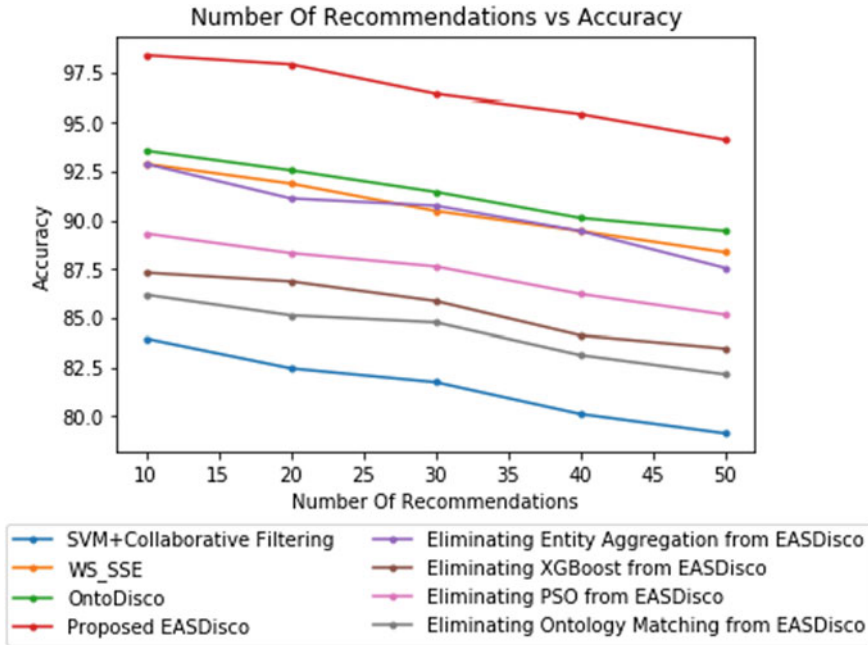


Fig. 4 Number of recommendations versus accuracy

95.54%, 0.07%, and 92.34%, respectively. For SVM + Collaborative Filtering [3], the average precision, recall, accuracy, *F*-measure, FDR, and nDCG were found to be 81.42%, 78.69%, 79.97%, 80.03%, 0.21%, and 77.42%, respectively. For WS_SSE [4], the average precision, recall, accuracy, *F*-measure, and FDR were found to be 90.41, 87.69, 88.72, 89.03, 0.12, and 78.69%.

The proposed approach was compared with baseline approaches by removing some of the primary techniques, and their performance is also evaluated. The proposed approach is superior to the compared approaches and models in terms of performance. This is because the proposed architecture uses entity aggregation along with ontology matching, and a gated recurrent unit is used for classification and genetic algorithm for final optimization.

Figure 4 depicts the comparison of Web service recommendations versus accuracy. Here, proposed EASDisco is baselined with SVM + Collaborative Filtering [4], WS_SSE [1], OntoDisco [3], Eliminating Entity Aggregation from EASDisco, Eliminating XGBoost from EASDisco, Eliminating PSO from EASDisco, and Eliminating Ontology Matching from EASDisco.

From Fig. 4, it is inferred that as the number of Web services recommended increases, the accuracy decreases. Peak accuracy is attained by the proposed EASDisco model when the number of recommendations is 10. The other models SVM + Collaborative Filtering [4], WS_SSE [4], OntoDisco [3], Eliminating Entity Aggregation from EASDisco, Eliminating XGBoost from EASDisco, Eliminating

PSO from EASDisco, and Eliminating Ontology Matching from EASDisco have an accuracy of 83.96, 92.89, 93.56, 92.89, 87.32, 89.32, and 86.19, respectively, when the number of recommendations is 10. As the number of recommendations is incremented to 50, the EASDisco system achieves an accuracy of 94.12%, while the baseline models have an accuracy of 79.12, 88.36, 89.45, 87.56, 83.45, 85.17, and 82.14%, respectively. Even while the number of recommendations is high, the EASDisco has higher accuracy than the baseline models; the reason for this is, in the proposed system, performing entity aggregation using Google Knowledge Base API and genetic algorithm which is utilized for optimization which increases the relevancy of the services and provides more accurate recommendation for the Web service.

6 Conclusion

Web services are one of the most valuable components of the World Wide Web since each service plays an important part. The proposed architecture is conceived to effectively come across various Web services that use an ontology that is formulated based on a set of key features which is related to Web services that are matched using SemantoSim Measure for which entity aggregation is done using Google's Knowledge Base API. The recommendation system has proven to be highly effective as it uses ontology matching, gated recurrent units, and computing semantic similarity using a genetic algorithm. The proposed EASDisco hybridization technique resulted in the best of all the approaches and overall accuracy of 95.39% is achieved.

References

1. C. Ma, M. Song, K. Xu, X. Zhang, in *Web Service Discovery Research and Implementation Based on Semantic Search Engine*. 2010 IEEE 2nd Symposium on Web Society (2010)
2. Y. Jiang, J. Liu, M. Tang, X. Liu, in *An Effective Web Service Recommendation Method Based on Personalized Collaborative Filtering*. 2011 IEEE International Conference on Web Services (IEEE, 2011, July), pp. 211–218
3. C.N. Pushpa, G. Deepak, A. Kumar, J. Thriveni, K.R. Venugopal, in *OntoDisco: Improving Web Service Discovery by Hybridization of Ontology Focused Concept Clustering and Interface Semantics*. 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT) (2020), pp. 1–5
4. L. Ren, W. Wang, An SVM-based collaborative filtering approach for Top-N web services recommendation. *Future Gener. Comput. Syst.* **78**(2), 531–543 (2018)
5. X. Chen, X. Liu, Z. Huang, H. Sun, in *Regionknn: A Scalable Hybrid Collaborative Filtering Algorithm for Personalized Web Service Recommendation*. 2010 IEEE International Conference on Web Services (IEEE, 2010 July), pp. 9–16
6. H. Xia, T. Yoshida, in *Web Service Recommendation with Ontology-Based Similarity Measure*. Second International Conference on Innovative Computing, Informatio and Control (icic 2007) (IEEE, 2007 September), pp. 412–412

7. N.K. Karthikeyan, R.S. Raj Kumar, Fuzzy service conceptual ontology system for cloud service recommendation. *Comput. Elect. Eng.* **69**, 435–446 (2018)
8. C. Feier, A. Polleres, R. Dumitru, J. Domingue, M. Stollberg, D. Fensel, Towards Intelligent Web Services: The Web Service Modeling Ontology (WSMO) (2005)
9. R.A. Rupasingha, I. Paik, in *Improving Service Recommendation by Alleviating the Sparsity with a Novel Ontology-Based Clustering*. 2018 IEEE International Conference on Web Services (ICWS) (IEEE, 2018 July), pp 351–354
10. A. Costa, R. Guizzardi, G. Guizzardi, J.G. Pereira Filho, in *COReS: Context-Aware, Ontology-Based Recommender System for Service Recommendation*. Proceedings of 19th International Conference on Advanced Information Systems Engineering (CAISE07) (2007 June)
11. G. Deepak, A. Santhanavijayan, OntoBestFit: a best-fit occurrence estimation strategy for RDF driven faceted semantic search. *Comput. Commun.* **160**, 284–298 (2020)
12. N. Krishnan, G. Deepak, in *Towards a Novel Framework for Trust Driven Web URL Recommendation Incorporating Semantic Alignment and Recurrent Neural Network*. 2021 7th International Conference on Web Research (ICWR) (2021), pp. 232–237. <https://doi.org/10.1109/ICWR51868.2021.9443136>
13. M. Arulmozhivarman, G. Deepak, in *OWLW: Ontology Focused User Centric Architecture for Web Service Recommendation Based on LSTM and Whale Optimization*. European, Asian, Middle Eastern, North African Conference on Management & Information Systems (Springer, Cham, 2021 March), pp. 334–344
14. G. Deepak, J.S. Priyadarshini, Personalized and enhanced hybridized semantic algorithm for web image retrieval incorporating ontology classification, strategic query expansion, and content-based analysis. *Comput. Electr. Eng.* **72**, 14–25 (2018)
15. N. Krishnan, G. Deepak, in *KnowSum: Knowledge Inclusive Approach for Text Summarization Using Semantic Allignment*. 2021 7th International Conference on Web Research (ICWR) (2021), pp. 227–231. <https://doi.org/10.1109/ICWR51868.2021.9443149>
16. Z. Gulzar, A.A. Leema, G. Deepak, Pcrs: personalized course recommender system based on hybrid approach. *Procedia Comput. Sci.* **125**, 518–524 (2018)
17. N. Krishnan, G. Deepak, KnowCrawler: AI classification cloud-driven framework for web crawling using collective knowledge, in *Artificial Intelligence Systems and the Internet of Things in the Digital Era. EAMMIS 2021. Lecture Notes in Networks and Systems*, by eds. A.M. Musleh Al-Sartawi, A. Razzaque, M.M. Kamal, vol. 239 (Springer, Cham, 2021). https://doi.org/10.1007/978-3-030-77246-8_35
18. G. Deepak, J.S. Priyadarshini, A hybrid semantic algorithm for web image retrieval incorporating ontology classification and user-driven query expansion. In *Advances in Big Data and Cloud Computing* (Springer, Singapore, 2018), pp. 41–49
19. G. Deepak, B.N. Shwetha, C.N. Pushpa, J. Thriveni, K.R. Venugopal, A hybridized semantic trust-based framework for personalized web page recommendation. *Int. J. Comput. Appl.* **42**(8), 729–739 (2020)
20. I.S. Kaushik, G. Deepak, A. Santhanavijayan, QuantQueryEXP: a novel strategic approach for query expansion based on quantum computing principles. *J. Discrete Math. Sci. Crypt.* **23**(2), 573–584 (2020)

An Approach Towards Human Centric Automatic Ontology Design



S. Manaswini, Gerard Deepak, and A. Santhanavijayan

Abstract Given the magnanimous amount of data that the web stores and with the evolution of semantic web, appropriate techniques for management of semantic information become vital. The time of accessing the required piece of information defines the efficiency of a system. Ontologies play a very important role in defining and organizing the information segments of a domain, which in turn help improve the efficiency of retrieval of required information. In this paper, the focus is to densify and generate improved ontologies from adopted seed-domain ontologies by incorporating a framework that utilizes a multisegmented methodology involving the LSTM model for classification followed by the cuckoo search metaheuristic optimization algorithm and semantic similarity computation approaches such as Kullback–Leibler divergence and SemantoSim measure—a child approach of the commonly used WebPMI, to hold context and enrich the relatedness in the improvised ontologies. This human centric approach also implements various logic rules and agents to delicately handle the semantic data at the same time preserve its integrity. Domains adopted for the purpose of experimentation are ensured to be from diverse real-world topics. The efficiency of the proposed model is seen to be higher than the adopted baseline, and the former supported with an accuracy of 96.12% and a false discovery rate of 0.043, therefore, exhibiting clean success of experimentation.

Keywords Automatic ontology generation · Cuckoo search · Domain ontology · LSTM · Kullback–Leibler divergence · SemantoSim

S. Manaswini

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India

G. Deepak (✉) · A. Santhanavijayan

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

1 Introduction

The amount of data that the Internet stores is immeasurable. The number of topics, the number resources available under each topic and the extent of relevancy for mapping allied topics are astounding. But, how the web manages to handle so much data and map according to relevance efficiently has no standardized procedure. The dynamic nature of the size and number of data topics calls for agility of algorithms and methods provisioning efficient accessibility. The web has a defined set of principles upon which the retrieval of information relies on. In the process of handling semantic information, the semantic web plays a very important role. Structuring semantic information or such contents of the web is essential for the faster retrieval of web contents, and ontologies play a critical role in the same. Ontology formulation demands a systematic, organized approach as the correctness of ontologies must be validated for it to function to its fullest potential. Briefing an ontology, it can be defined as a framework of data points that can be shared and reused across various domains. The very aim of ontology modeling in a system design is to make a collective, high-quality, inter-connected and coherent dataset for greater efficiency and accessibility [1]. The ability of ontologies to integrate and define semantic models with associated domain knowledge adds to its advantage. In simple terms, ontologies identify and define concepts and their relationships and are found to be an integral part of the semantic Web. Therefore, the validation of the designed framework becomes vital [2]. Semantic approaches to preserve ontologies that depend on context and their structures are crucial. The very aim is to model an ontology that is independent of semantic wikis and develop a semantic strategy to organize ontological hierarchies effectively [3].

Motivation: The very process of ontology modeling or authoring has been observed to be fragmented over various tools and methodologies. Traditionally, ontology authoring was based on semantic wikis or graph-based structures that exhibit no formalism to the framework of modeling and exploring ontologies, especially OWL, given its incoherence and inconsistency. In consideration of many such insights that uncover the issues in building styles such as manually crafted models, definition-driven models and non-generic domain-specific frameworks, an automated yet human-centric approach has been projected. Extensive influence of authors in drawing an ontology specific to domain and furthermore testing and validating the models could become tedious, and at the same time, the correctness of the ontology demands supervision. Also, the union of author modeled ontologies, and those from web sources quantitatively increase the number of available results in the domains. Therefore, the intent in the designed system is to enrich the density of generic ontologies over specified domains and at the same time validate the exactitude of the generated mesh.

Contribution: In this paper, the proposed scheme can be segmented into multiple stages. As requirements for experimentation, about 24 topics have been adopted

with quantized number of seed domains under each one of the former. As the intention is to increase the number of concepts and individuals, using an automated approach, a taxonomy of seed domains was parsed into a domain-based filtering methodology which utilizes long short-term memory (LSTM) [4] via a structured topic model [5]. This, then, undergoes semantic integration along with other sources of external knowledge like Wikidata, Google knowledge base which is optimized and then axiomatized and expunged of inconsistencies to generate enriched ontologies expanding the spectrum of domains.

Organization: The first section of the paper includes a brief introduction to the concept. The flow of the rest of the paper is as specified: Sect. 2 consists of work related to the field of study and experimentation. Section 3 bares the proposed architecture for the system of automatic ontology design. Section 4 consists of the implementation and experimental observation. Section 5 includes conclusion.

2 Related Work

Pushpa et al. [6] have proposed a strategy resting on hash table-based ontology organization. To compute the relevance of the ontologies, it is further supported by a semantic latent analysis. The proposed framework is further enhanced with content-based filtering for yielding 88.99% overall accuracy. Liebig et al. [7] have proposed a semantic-oriented ontology visualization combining improved results of research and experimentation in the field of HCI, such as zoomable interfaces, thumbnail previews, with aspects of ontology authoring. Kapoor et al. [8] aimed to identify all possible existing ontologies and ontology management tools that are freely available and review them in terms of interoperability, openness, easiness to update and maintain market status and penetration. Cheng et al. [9] have proposed a system in which ontologies and rules have been used to establish a knowledge base in the transportation domain. Problems such as insufficient expressivity in ontologies under spatial relationship reasoning have been eliminated by using the SWRL rule, ambiguity of concepts, hidden information, and semantic level inquiry has also been tackled catering to the needs of the non-GIS users. Elnagar et al. [10] have authored a paper proposing an automatic ontology generation framework from the perspective of organization. The main focus is to generate a domain-independent ontological scheme that converts unstructured text corpus into domain-consistent ontological form by initially generating KGs from unstructured text corpus. Alalwan et al. [11] have proposed a model for the generation of OWL ontology for databases. Metadata is obtained from the analysis of database tuples by the system, after which the resultant ontology is manually validated and compared with the conceptual database or model to capture the optimal ontology. In [12–23], several ontological approaches in support of the proposed work are discussed.

3 Proposed Architecture

The proposed architecture models a strategy to generate OWL ontologies with increased density in each of the adopted domains. The topics that have been considered for modeling include: Petroleum, post-colonial English literature, culture and heritage, law of evidence, sociology of gender, econometrics, developmental psychology, magazine journalism, experimental psychology, landscape ecology, community medicine, paleontology, petro-informatics, geology, crop, horticulture, furniture, men's fashion, community medicine, radiodiagnosis, automobile, restaurant, flower and music out of which ten topics have been illustrated in the paper for reference. A thesauri that has a specific number of seed-domain classes and instances ranging between 440 and 1400 for each of the mentioned topics has been accumulated, and the illustrated ten domains have been recorded in Table 1. The first stage involves preparation of data points for the extracted domain ontologies wherein the ontology author defines the concepts and instances to be considered for experimentation. Initially, the model focuses on expanding the spectrum of instances and concepts. For tokenization, the taxonomy of seed domains is parsed into bigrams for preserving context in sentimental analysis. The quality of features also increases with bigrams, therefore, increasing the F -measure of the model. This, then, involves structural topic modeling (STM) and is indicated in the Proposed System Design(Fig. 1).

This module enables the retention of the corpus structure without having to develop new models as it is a generalized linear model with document-level covariate information and, hence, can be used to condition raw text affecting the topic prevalence or topic content, reducing the time for ontology generation. The second stage consists of a storehouse of external knowledge such as Wikidata, LOD cloud, Google knowledge base API and Freebase has been used to endorse the input for classification of extracted concepts and individuals for content-based filtering. In the third stage, a

Table 1 Performance metrics of adopted domain ontologies

Domain ontologies	Precision %	Recall %	F -score %	Accuracy %	False discovery rate
Geological ontology	94.36	97.85	96.07	95.62	0.056
Crop ontology	93.21	98.42	95.74	95.12	0.067
Horticulture ontology	96.32	99.12	97.69	97.39	0.036
Furniture ontology	92.85	93.25	93.04	93.02	0.071
Men's fashion ontology	91.54	97.69	94.51	94.44	0.084
Community medicine ontology	94.69	97.86	96.24	95.51	0.053
Radiodiagnosis ontology	94.56	98.21	96.35	96.89	0.054
Automobile ontology	95.62	97.54	96.57	96.02	0.043

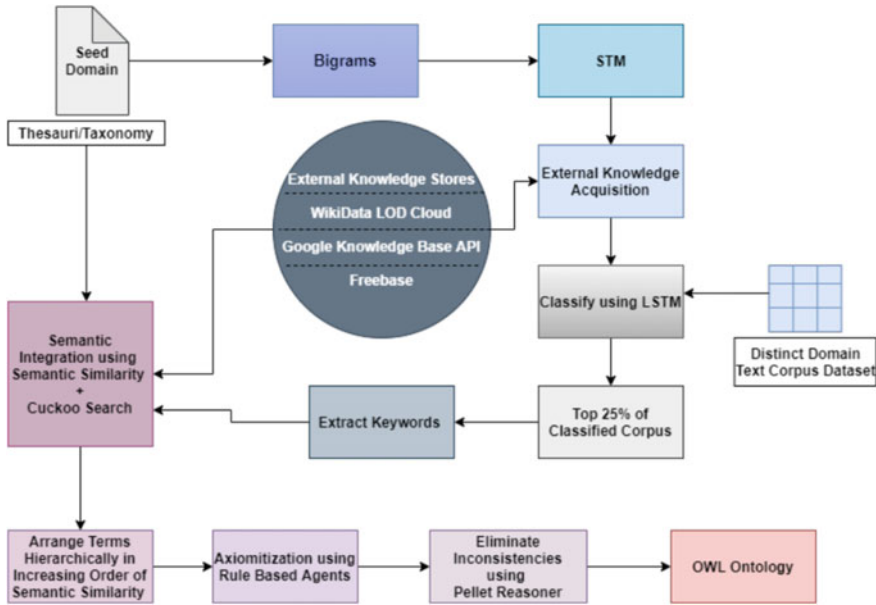


Fig. 1 Proposed system design

distinctly categorized domain text corpus dataset is passed into the LSTM classifier along with the externally accreted domain knowledge resource, of which the output is the top 25% of the documents of synonymous entities of the ontology which were identified by a frequency computation measure like TF-IDF, in each domain, between the formerly mentioned inputs. The amalgamation of the derived sets involves a series of procedures executed on the keywords extracted from the output of the classified corpus and the initially adopted seed-domain ontologies, leading on to the fourth stage. Semantic integration is a very vital step in interrelating the various entities from diverse sources but at the same time not compromising on their correctness. Same terms can denote different concepts, and different terms can be used to refer to the same hierarchy and here is where ontology mismatches are likely to occur, that are handled by distinct dimensions of semantic integration. After integration of semantically relevant concepts and individuals where relevant branches are clustered together, preserving the description of the hierarchy of the adopted ontologies, they are then to be used for axiomatization by semantic agents. To calculate the semantic similarity between the extracted words, the external resources and the initial taxonomy, two methodologies have been implemented, namely the SemantoSim measure [24] and Kullback–Leibler divergence [25], aiding to the efficiency of the approach. The former has taken inspiration from the pointwise mutual information measure, a page count-based co-occurrence approach also used to calculate semantic similarity. The aim of the measure is to explore the extent of recognizing

and relating words with similar meanings but at the same time ensuring to retain their contextual meaning.

$$\text{SemantoSim}(x, y) = \text{PMI}(x, y) + P(x, y) \log[P(x, y)]/P(x) \cdot P(y) + \log[P(y, x)] \quad (1)$$

SemantoSim is represented as in Eq. (1). The extracted terms are paired if in case there are two terms for comparison. The resultant yields the relatedness between the two terms. PMI of the two terms are calculated, and $p(x, y)$ is the probability of the co-occurrence of “ x ” with “ y ” and $p(y, x)$ vice versa. $p(x)$ and $p(y)$ denote the probability of their individual presence. Furthermore, the divergence is also calculated to measure the difference between the probability distributions of a term “ x .” It returns a non-negative score that calculates the divergence of one probability distribution from another. If A and B are two probability distributions, the Kullback–Leibler distribution is denoted by the following Eqs. (2) and (3), where “ \parallel ” denotes divergence and “ x ” being each event in the distributions. The resultant is coupled with a metaheuristic algorithm—cuckoo search for the optimization of the validated hierarchies. Despite being able to represent abstract information captured by multiple sources under each domain, it also provides a front for overlapping parts as a domain can be conceptualized in a variety of ways. For this very cause, cuckoo search has been used as it exhibits great efficiency in optimization.

$$KL(A \parallel B) \quad (2)$$

$$KL(A \parallel B) = \sum_{i=1}^n P(X_i) \log \frac{A(x)}{B(X)} \quad (3)$$

Obtained resultant is coupled with a metaheuristic algorithm—cuckoo search for the optimization of the validated hierarchies. Despite being able to represent abstract information captured by multiple sources under each domain, it also provides a front for overlapping parts as a domain can be conceptualized in a variety of ways. For this very cause, cuckoo search has been used as it exhibits great efficiency in optimization. The algorithm is based on the brood parasitism of the cuckoo and is enhanced by the levy flights. It uses a proportionate amount of both local and global explorative random walk controlled by a switching parameter. The local random walk can be represented as in Eq. (4) where “ $x_{i,t}$ ” and “ $x_{j,t}$ ” are random solutions generated by random permutation, “ pa ” is the switching parameter, “ \otimes ” is the entrywise product, $H(u)$ is the Heaviside function, “ S ” is the step size, and “ ϵ ” is a random number drawn from a uniform distribution. Levy flights are used to project the global random walk, as in Eq. (5)

$$X_{i,t+1} = X_{i,t} + \alpha S \otimes H(pa - \epsilon) \otimes (X_{j,t} - X_{k,t}) \quad (4)$$

$$X_{i,t+1} = X_{i,t} + \alpha L(S, \lambda) \quad (5)$$

where $L(u)$ is the levy flight equation and $\alpha > 0$ is the step-size scaling factor. Since levy flights are said to be extremely efficient, global random walk-based randomization techniques which in turn increase the efficiency of cuckoo search. The output terms are hierarchically arranged in increasing order of semantic similarity after which the next stage involves the axiomatization of the processed entities. It is a process by which rules are applied to the associations between the ontological entities in the structured hierarchical corpus derived. It is induced on both the individuals and concepts. The axiomatization agent is meant to identify the relation between concepts and enforce rules and axioms on the concepts to generate structured ontologies. These are then handled by the Pellet and Hermit reasoners to test the class satisfiability, query context and eliminate the inconsistencies in the generated ontology to improve its correctness. The final OWL ontology is presented to the user in place.

4 Implementation

The proposed architecture has been built on JAVA over NetBeans IDE for ease of computation. The manual ontologies were modeled over Web Protege, and the automated frameworks were built over Onto Collab. The Pellet and Hermit reasoners used to test the inconsistencies in ontologies and the semantic agents used to test the framework include JADE and AgentSpeak implemented in Java, respectively. The system has been designed using tools of NLP, namely Stanford CoreNLP. The model runs on Windows 10 operating system installed over Intel Core i7 8th gen processor supported by 16 GB RAM. The “. owl” ontology files used for experimentation are domain-specific and quantized where each of the concepts is associated with a specific number of entities. Ten out of the selected 24 domains have been illustrated for showcasing the experimentation. They are geology, crop, horticulture, furniture, men’s fashion, community medicine, radiodiagnosis and automobile.

The precision, recall, F -measure, accuracy and false discovery rate for each of the domains have been recorded under Table 1. The outcomes are based on the number of seed domains assumed for each and the versatility of each concept and the number of ways of interpreting them. The baseline HybridOnto model proposes a hybrid hash table-based ontology modeling structure, divided into three distinct phases of implementation with the very aim of improving ontological relevance incorporating a strategy called latent semantic analysis. After careful experimentation and comparison, the performance metrics have been plotted in Figs. 2 and 3, where it is clearly seen that the proposed methodology has exceeded in efficiency of modeling. The precision, recall, accuracy, F -measure and false discovery rate of the HybridOnto model are 87.68%, 92.39%, 90.12%, 89.97% and 0.123%, respectively, and for the proposed approach is 95.62%, 98.41%, 96.12%, 96.99% and 0.043%, respectively. The most crucial modules in the proposed system responsible for the increased accuracy and efficiency are the inclusion of a metaheuristic optimization algorithm along with the computation of semantic similarity.

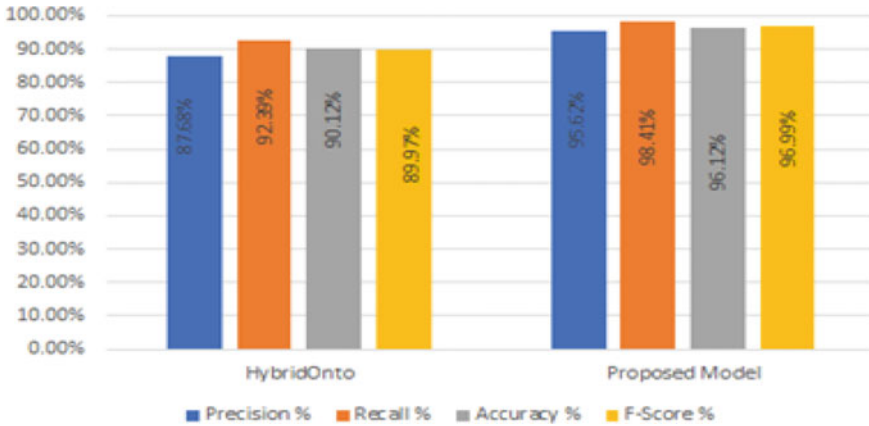
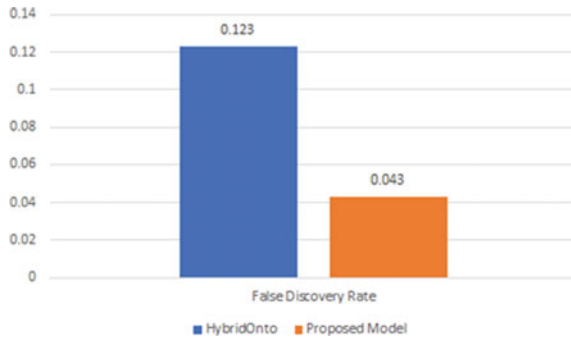


Fig. 2 Performance graph baseline approach versus proposed approach

Fig. 3 FDR HybridOnto versus proposed model



Furthermore, the dual-layer semantic computation using a divergence metric and similarity measure improves the probability of correctness in validation of the ontology. As seen from the experimentation results, the proposed model has shown an increase of 6% in accuracy and 7% in precision. The process of use of heterogeneous sources for the enrichment of the hierarchical corpus before classification and filtering by the LSTM helps prepare a profound set of entities before optimization. Semantic integration is done using semantic similarity using cuckoo search. The cuckoo search, being a swarm intelligence algorithm, has been used for optimization as the involvement of optimization helps regulate the semantic similarity. The objective function here is the resultant of Kullback–Leibler divergence and SemantoSim measure that are then optimized and put into hierarchy. The optimization is applied on the input from the seed taxonomy, and the computation of the probability measure between the events and the similarity measure using SemantoSim method aids the precision as the extent of relatedness of two entities or terms in an ontology is what is most important for its enrichment and existence. This optimization is done

to increase the quality of output of semantic similarity by increasing its efficiency and reducing the diversity to utilize and explore the randomness of optimization.

Thereby providing ground for the arranged, filtered corpus to be incorporated with the inferred and deduced logics, enhancing the process of authoring which is then followed by the process of removal of inconsistencies using the Pellet and HermiT reasoners as axiomatized semantic approaches is prone to inconsistencies and negations that could affect the genuineness of the framework designed to generate the required ontologies. Pellet and HermiT reasoners have been used for their simplicity in functionality and because we have not included extremely large numbers of domains for experimental purpose. Just as we use symbolic debuggers, tracers, inspectors and many such tools for debugging programs, ontologies also require such mechanisms for organizing and presenting the information fed into them. And, such are the reasoners. Debugging of ontologies can be tedious due to their semantics and language, which are likely to be alien in comparison. Pellet is not as mature a tool but is efficient enough to run on a countable number of real ontologies thereby compensating for the lack of fast responsiveness. The listed reasoners include several advantages such as native support for OWL and support for instances and XML schema types and cover a vast region of the OWL DL.

5 Conclusion

The graph-driven ontological entities can often cause confusion and lose the integrity of the domain knowledge. The proposed approach tackles the issue of manual authoring of ontologies at the same time preserving and improving the quality of the generated ontologies via an automated human-centric approach, which bridges the gap between automated authoring and manual authoring with an increased accuracy of 96.12%. The optimization cuckoo algorithm contributes to the semantic integration by means of the semantic similarity, and the LSTM adds to the increased efficiency by its regulated input gates supporting lexical analysis and filtering. Countable numbers of real-world domains have been adopted and experimented with, in order to add to the depth of each domain. The experimentation is seen to be successful and is supported by a false discovery rate of 0.043 in the proposed system.

References




1. M. Vigo, S. Bail, C. Jay, R. Stevens, Overcoming the pitfalls of ontology authoring: strategies and implications for tool design. *Int. J. Hum. Comput. Stud.* **72**, 835–845 (2014)
2. B. Parsia, E. Sirin, A. Kalyanpur, in *Debugging OWL Ontologies*. International World Wide Web Conference Committee (IW3C2) (2005)
3. T. Liebig, O. Noppens, in *OntoTrack: A Semantic Approach for Ontology Authoring*. Web Semantics: Science, Services and Agents on the World Wide Web, vol. 3 (2005), pp. 116–131

4. M.E. Roberts, B.M. Stewart, D. Tingley, E.M. Airolidi, in *The Structural Topic Model and Applied Social Science*. NIPS Workshop on Topic Models: Computation, Application, and Evaluation (2013)
5. S. Hochreiter, J. Schmidhuber, Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997)
6. C.N. Pushpa, G. Deepak, J. Thriveni, K.R. Venugopal, A hybridized framework for ontology modeling incorporating latent semantic analysis and content based filtering. *Int. J. Comput. Appl.* 0975–8887 (2016)
7. T. Liebig, O. Noppens, OntoTrack: a semantic approach for ontology authoring. *J. Web Seman.* **3** (2005)
8. B. Kapoor, S. Sharma, A comparative study ontology building tools for semantic web applications. *Int. J. Web Seman. Technol. (IJWesT)* (2010)
9. G. Cheng, Q. Du, in *The Design and Implementation of Ontology and Rules Based Knowledge Base for Transportation*. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XXXVII, Part B2, Beijing (2008)
10. S. Elnagar, V. Yoon, M.A. Thomas, in *An Automatic Ontology Generation Framework with an Organizational Perspective*. Hawaii International Conference on System Sciences (2020)
11. N. Alalwan, H. Zedan, F. Siewe, in *Generating OWL Ontology for Database Integration*. Third International Conference on Advances in Semantic Processing (2009)
12. *IEEE Transactions on Information Theory*, vol. 60, no. 7 (2014)
13. V. Adithya, G. Deepak, in *OntoReq: An Ontology Focused Collective Knowledge Approach for Requirement Traceability Modelling*. In European, Asian, Middle Eastern, North African Conference on Management & Information Systems (Springer, Cham, 2021 March), pp. 358–370
14. V. Adithya, G. Deepak, A. Santhanavijayan, in *HCODF: Hybrid Cognitive Ontology Driven Framework for Socially Relevant News Validation*. International Conference on Digital Technologies and Applications (Springer, Cham, 2021 January), pp. 731–739
15. G.L. Giri, G. Deepak, S.H. Manjula, K.R. Venugopal, in *OntoYield: A Semantic Approach for Context-Based Ontology Recommendation Based on Structure Preservation*. Proceedings of International Conference on Computational Intelligence and Data Engineering: ICCIDE 2017, vol. 9 (Springer, 2017 December), p. 265
16. G. Deepak, N. Kumar, A. Santhanavijayan, A semantic approach for entity linking by diverse knowledge integration incorporating role-based chunking. *Procedia Comput. Sci.* **167**, 737–746 (2020)
17. G. Deepak, S. Rooban, A. Santhanavijayan, A knowledge centric hybridized approach for crime classification incorporating deep bi-LSTM neural network. *Multimedia Tools Appl.* 1–25 (2021)
18. K. Vishal, G. Deepak, A. Santhanavijayan, in *An Approach for Retrieval of Text Documents by Hybridizing Structural Topic Modeling and Pointwise Mutual Information*. Innovations in Electrical and Electronic Engineering (Springer, Singapore, 2021), pp. 969–977
19. G. Deepak, V. Teja, A. Santhanavijayan, A novel firefly driven scheme for resume parsing and matching based on entity linking paradigm. *J. Discrete Math. Sci. Crypt.* **23**(1), 157–165 (2020)
20. G. Deepak, N. Kumar, G.V.S.Y. Bharadwaj, A. Santhanavijayan, in *OntoQuest: An Ontological Strategy for Automatic Question Generation for e-Assessment Using Static and Dynamic Knowledge*. 2019 Fifteenth International Conference on Information Processing (ICINPRO) (IEEE, 2019 December), pp. 1–6
21. G. Deepak, A. Santhanavijayan, OntoBestFit: a best-fit occurrence estimation strategy for RDF driven faceted semantic search. *Comput. Commun.* **160**, 284–298 (2020)
22. M. Arulmozhivarman, G. Deepak, in *OWLW: Ontology Focused User Centric Architecture for Web Service Recommendation Based on LSTM and Whale Optimization*. European, Asian, Middle Eastern, North African Conference on Management & Information Systems (Springer, Cham, 2021 March), pp. 334–344
23. G. Deepak, J.S. Priyadarshini, Personalized and enhanced hybridized semantic algorithm for web image retrieval incorporating ontology classification, strategic query expansion, and content-based analysis. *Comput. Electr. Eng.* **72**, 14–25 (2018)

24. G.L. Giri, G. Deepak, S.H. Manjula, K.R. Venugopal. in *OntoYield: A Semantic Approach for Context-Based Ontology Recommendation Based on Structure Preservation*. Proceedings of International Conference on Computational Intelligence and Data Engineering (2018)
25. T. van Erven, P. Harremoës, *Rényi Divergence and Kullback–Leibler Divergence* (2014)

A Review on Dataset Acquisition Techniques in Gesture Recognition from Indian Sign Language



Animesh Singh , Sunil Kr. Singh , and Ajay Mittal 

Abstract Communication for dumb and deaf community is done by the sign language using hand and finger movements. Adding grammar and emotion to a sentence require lips movement, eye gaze, and facial expressions. In this paper, we have discussed many research papers related to Indian sign language gesture recognition. For static gestures 80.2–99.73% and for dynamic gestures 72.3–99.08% result in accuracy rates found with some certain specific environmental conditions. Papers are analyzed based on dataset acquisition method, feature extraction techniques, and classification techniques with their comparative graphs tables. Some paper lag facial expression with hand gesture, hence further research is required in this direction. From the entire techniques, deep learning convolution neural network (CNN), support vector machine (SVM), and hidden Markov model (HMM) were found to give the better result as compared to others.

Keywords Sign language · Gesture recognition · Neural network (NN) · Multimodal framework · Deep learning

1 Introduction

Communication is the only way by which we can interact with anyone and it can be done using speech, face expression, lips movement, hand gesture, and signing. For the deaf and dumb community people's communication, research on recognition

A. Singh (✉) · S. Kr. Singh

Department of Computer Science and Engineering, Chandigarh College of Engineering and Technology, Sector 26, Chandigarh 160019, India
e-mail: animeshsingh@ccet.ac.in

S. Kr. Singh

e-mail: sksingh@ccet.ac.in

A. Mittal

Department of Computer Science and Engineering, University Institute of Engineering and Technology, Sector 25, Chandigarh 160014, India
e-mail: ajaymittal@pu.ac.in

of gestures plays a vital role. In this paper, we have discussed various researches done on Indian sign language gestures recognition that will help many researchers to further analyze and deal with those gaps. Section 2 discusses an overview of techniques and methods used for both dynamic and static sign language recognition. Section 3 interprets observation and conclusion of Indian sign language recognition, and Sect. 4 depicts future scope and further directions in Indian sign language gesture recognition research, and at last, it was ended with references.

2 Overview of Techniques and Discussion

This section deals with the research papers based on data acquisition techniques and methods such as Web camera (2D image and computer vision), sensor data glove, Kinect camera, leap motion sensor, multimodal [1], electromyography, and 9-camera model (8 motion capture and 1 video camera).

- (A) First, we see data acquisition through Web cameras. Futane et al. [2] predict gesture by general purpose fuzzy min–max NN and key extraction algorithm and got the result of 92.92% accuracy. Divya et al. [3] proposed a principal component analysis method to detect signs and in that they use a table lamp also to maintain the intensity of light while capturing image through a camera. Adithya et al. [4] have implemented a feed-forward neural network technique and got an accuracy of 91.11%. Prema et al. [5] propose a method of hand and finger movement tracking using meanshift and Kalman filter and acquire an accuracy of 80.20%. Anil et al. [6] capture the image using the selfie stick, hence a problem of shaking video occurs but they achieve the result of 90% accuracy. Ali et al. [7] propose data augmentation to increase the accuracy result. Athira et al. [8] use an appearance-based approach and it works and focuses on co-articulation removal to enhance the result performance. This paper gives an accuracy of 90.10% for static gestures [9] and 89% accuracy for dynamic [10] hand gestures.
- (B) Second, we see data acquisition through the sensor data glove method. Geetha et al. [11] propose the concept of key frame extraction which corresponds to the maximum curvature points (MCPs) of the global trajectory where the same gesture is performed by different persons and spatial [12] location of the key MCPs of the boundary is used for extraction of shape and it attains an accuracy of 90.90% (Figs. 1 and 2).
- (C) Third, we see about data acquisition through leap motion sensors. Poonam et al. [13] used a leap motion sensor and random forest classifier to forecast text from Indian sign language. Anshul et al. [14] capture dataset using leap motion to solve the problem of occlusion due to fingers and hands and have achieved an accuracy of 72.3%. The leap motion sensor framework shown in Fig. 3 was taken from Anshul et al. [14], for reference.

Method of Data Acquisition	Name of the Author; year of publishing	Method (Feature extraction & Classification techniques)	Number & type of gestures examined	Accuracy (in percentage)
Web camera (2D Image and computer vision)	Grzeszczuk et al.; 2000	Position of hand, two dimensional geometric features by means of distance metric	6 number of sign gestures	96%
	Ong et al.; 2006	HMM applied for lexicon sign recognition	20 different gesture of ASL	85.20%
	Elmezain et al.; 2008	Gaussian model, using hidden markov model blob analysis is done	Gestures including alpha-numeric signs	94.72%
	Futane et al.; 2011	Key extraction, general purpose fuzzy minmax NN	5 sentences with 2-3 static gestures in each	92.92%
	Divya et al.; 2012	Finger tip algorithm and PCA	20 alphabets and 9 numbers	94.00%
	Adithya V. et al.; 2013	feed forward NN	20 alphabets and 9 numbers	91.11%
	Kanchan et al.; 2014	Haar cascade classifier	3 sentences and two words	92.68%
	Prema et al.; 2015	Minshift and Kalman filter	2 words, 2 alphabets & 1 palm gesture	80.20%
	D. Anil et al.; 2016	Artificial neural network with back propagation algorithm, Mahalanobis distance	18 ISL signs	90%
	Ali et al.; 2018	Adapted deep convolutional neural network	6 ISL gestures	99.73%
	Athira et al.; 2019	Zernike moments and multi class SVM classifier	5 ISL static words and ISL dynamic gestures	90.1% for static & 89% for dynamic gesture
Cyber gloves or data gloves	Kong et al.; 2008	Linear discriminant analysis, NN based	30-40 (handshapes), 8-10 (hand orientation) 20 (hand movement trajectories)	86.80%
	Geetha et al.; 2013	Spatial location of boundary points	30 ISL banking term gestures	90.90%
	Geetha et al.; 2013	Trajectory based, eigen distance method and PCA	Not given	55% average
	Mehrotra et al.; 2015	3D points of skeleton, angular and distance features using support vector machine	37 number of Indian sign language gestures	86.16%

Fig. 1 Summary of related works contd.

(D) Fourth, we see about multimodal framework [15] that means concatenation of dataset that results from one or more sensors to increase the accuracy. Marin et al. [16] propose a combinational approach that uses leap motion sensor and Kinect device to capture sign words and achieve the result of 91.28%. Rossol et al. [17] implement multisensor system which has two leap motion sensors and they achieved a result accuracy of 90.80%. Pradeep et al. [18] propose

Method of Data Acquisition	Name of the Author; year of publishing	Method (Feature extraction & Classification techniques)	Number & type of gestures examined	Accuracy (in percentage)
Leap Motion	Poonam et al.; 2016	Random forest, block list classifier	26 ISL alphabets	Not given
	Anshul et al.; 2019	modified LSTM	35 ISL words	72.3% on signed sentences & 89.5% on isolated sign words
Multimodal	Mihail et al.; 2012	3D point cloud, PCA, angular features using k-NN	10 digit gestures	90%
	Marin et al.; 2014	fingertips angle, distance, elevation, curvature and correlation using SVM	10 ASL gesture	91.28%
	Rosol et al.; 2015	3D palm points, palm normal, hand direction etc. using SVM	3 different hand poses	90.80%
	Marin et al.; 2015	3D fingertips, curvature and correlation using SVM	10 ASL gestures	96.50%
	Pradeep et al.; 2016	Bidirectional Long Short Term Memory neural network (BLSTM-NN) and hidden markov modal	50 ISL gestures	97.85% for single hand & 94.55% for double hand
	Pradeep et al.; 2016	Coupled Hidden Markov Model	25 ISL gestures	90.80%
	Pradeep et al.; 2017	HMM & Independent Bayesian classification combination	51 dynamic ISL gestures	96.05% for single hand & 94.27% for double hand gesture
	Neel et al.; 2019	Convolutional neural network LSTM	36 static gesture for ISL alphabets and 10 ISL dynamic gesture	98.81% for static & 99.08% for dynamic gesture
Electromyography	Divya et al.; 2017	Average amplitude change, simple square integral, standard deviation	5 ISL gestures	90%
9-Camera Model (8 motion capture and 1 video camera)	Kishore et al.; 2018	kemel matching methods	20 ISL gestures	98.90%

Fig. 2 Summary of related works

multisensor data fusion system as shown in Fig. 4, and this figure is collected from Pradeep et al. [18], for reference.

HMM and bidirectional long short-term memory neural network classifiers (BLSTM-NN) are used to achieve the result of 97.85% for single hand and 94.55% for double hand gesture recognition. They proposed a multimodal framework that includes facial expression to correctly predict the signing words and gives better result as compare to unimodal framework by

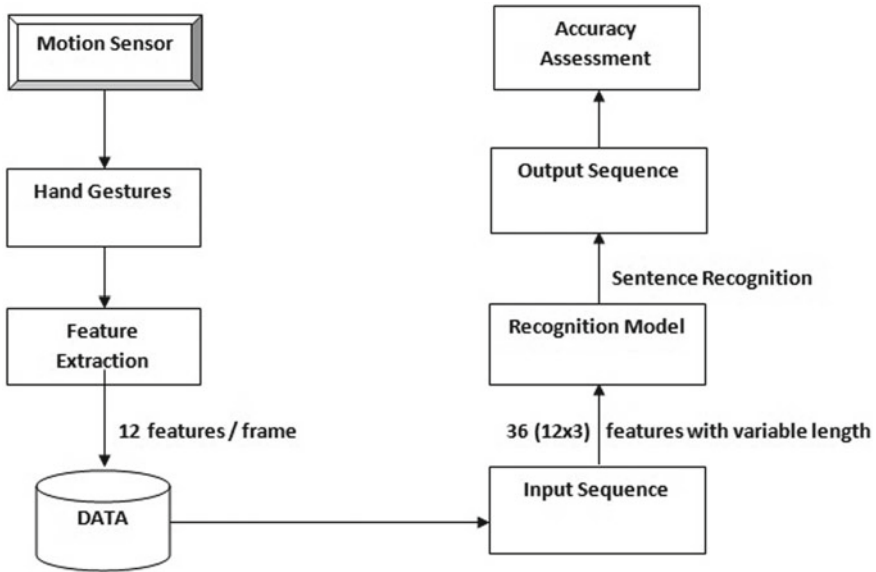


Fig. 3 Leap motion sensor framework (Anshul et al. [14])



Fig. 4 Image capturing setup (Pradeep et al. [18])

giving 96.05% accuracy for single hand and 94.27% accuracy for double hand gestures. Neel et al. [19] propose a method using convolution neural network [20, 21] for static sign and convolutional LSTMs for dynamic sign and an accuracy of 98.81% (for static gesture) and 99.08% (for dynamic gesture) are achieved.

- (E) Fifth, we see data acquisition through electromyography. Divya et al. [22] demonstrated a system that implements an electromyography hardware device in which EMG signals are captured using BIOPAC MP-45, and analysis is implemented using MATLAB.
- (F) Sixth, we found a well set up environment in the lab for capturing gestures [23]. In this paper, signer is classified into non-motion and motion joints for

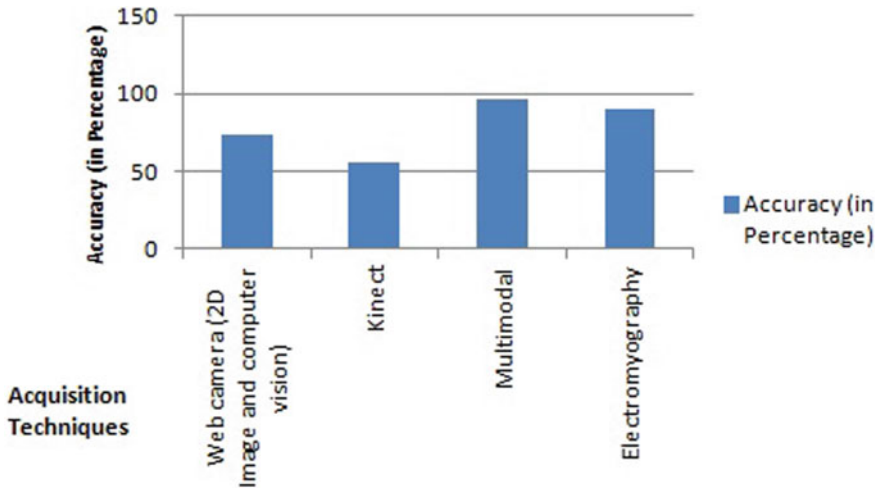


Fig. 5 Graph of acquisition techniques (static only) versus recognition accuracy

3D Indian sign language gesture. This lab arrangement consists of 9-camera model scenario which acquires the result of 98.90% accuracy.

3 Observation and Conclusion

We can say that out of all the methods, the multimodal framework [24] method to collect data is efficient. Our conclusion and observation from this paper consist of the following: (1) Application: Real-time application datasets are missing. (2) Dataset collection and lab setups: Image acquisition is done using proper lighting, plane background, and customized lab setup. (3) Hardware dependability: Specific hardwares like leap motion, Kinect camera, depth camera, etc., are used for dataset collection. (4) Result comparison: Result and accuracy are analyzed. (5) Occlusion: Less accuracy is seen in occluded images. Below are the two graphs in Figs. 5 and 6 that show recognition accuracy w.r.t. static and dynamic datasets on the basis of acquisition techniques.

4 Future Directions on Research

This paper describes that many paper lag facial expression and dataset collection from complex or moving background. Hence, more analysis and research are required in Indian sign language gesture recognition including facial expressions, which adds emotion and grammar [25] to the sign language.

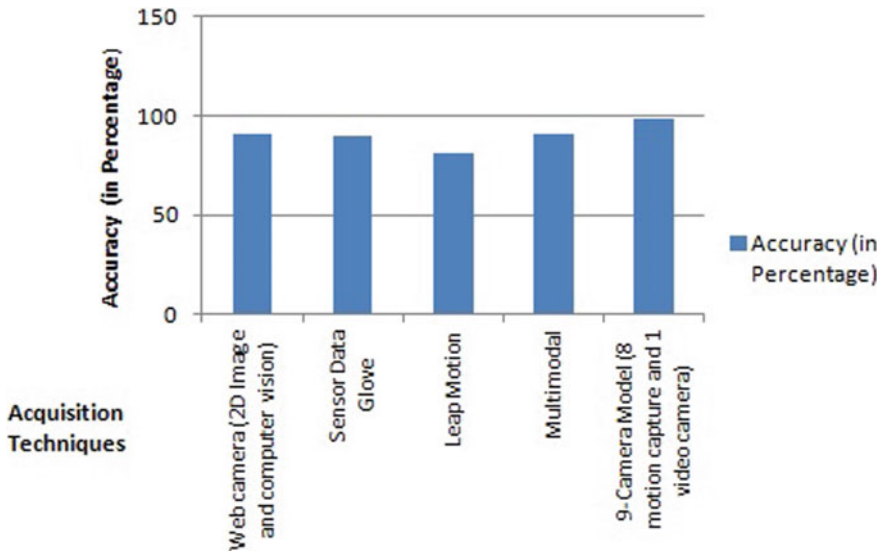


Fig. 6 Graph of acquisition techniques (dynamic only) versus recognition accuracy

References

1. R. Rastgoo, K. Kiani, S. Escalera, Multi-modal deep hand sign language recognition in still images using restricted Boltzmann machine. *Entropy* **20**(11) (2018). <https://doi.org/10.3390/e20110809>, <https://www.mdpi.com/1099-4300/20/11/809>
2. P.R. Futane, R.V. Dharaskar, in “*hasta mudra*”: An Interpretation of Indian Sign Hand Gestures. 2011 3rd International Conference on Electronics Computer Technology, vol. 2 (2011), pp. 377–380. <https://doi.org/10.1109/ICECTECH.2011.5941722>
3. D. Divya, N. Bajaj, in *Indian Sign Language Recognition*. 2012 1st International Conference on Emerging Technology Trends in Electronics, Communication Networking (2012), pp. 1–5. <https://doi.org/10.1109/ET2ECN.2012.6470093>
4. V. Adithya, P.R. Vinod, U. Gopalakrishnan, in *Artificial Neural Network Based Method for Indian Sign Language Recognition*. 2013 IEEE Conference on Information Communication Technologies (2013), pp. 1080–1085. <https://doi.org/10.1109/CICT.2013.6558259>
5. G. Prema, G. Joshi, M. Dutta, in *Comparative Analysis of Movement and Tracking Techniques for Indian Sign Language Recognition*. 2015 Fifth International Conference on Advanced Computing Communication Technologies (2015), pp. 90–95. <https://doi.org/10.1109/ACCT.2015.138>
6. K.D. Anil, P.V.V. Kishore, A.S.C.S. Sastry, P. Reddy Gurunatha Swamy, in *Selfie Continuous Sign Language Recognition Using Neural Network*. 2016 IEEE Annual India Conference (INDICON) (2016), pp. 1–6. <https://doi.org/10.1109/INDICON.2016.7839069>
7. A.A. Alani, G. Cosma, A. Taherkhani, T. McGinnity, in *Hand Gesture Recognition Using an Adapted Convolutional Neural Network with Data Augmentation*. 2018 4th International Conference on Information Management (ICIM) (2018), pp. 5–12. <https://doi.org/10.1109/INFOMAN.2018.8392660>
8. P. Athira, C. Sruthi, A. Lijiya, A signer independent sign language recognition with co-articulation elimination from live videos: an Indian scenario. *J. King Saud Univ. Comput. Inf. Sci.* (2019). <https://doi.org/10.1016/j.jksuci.2019.05.002>, <https://www.sciencedirect.com/science/article/pii/S131915781831228X>

9. A. Wadhawan, P. Kumar, Deep learning-based sign language recognition system for static signs. *Neural Comput. Appl.* **32**, 7957–7968 (2020)
10. C.C. dos Santos, J.L.A. Samatelo, R.F. Vassallo, Dynamic gesture recognition by using cnns and star rgb: A temporal information condensation. *Neurocomputing* **400**, 238–254 (2020). <https://doi.org/10.1016/j.neucom.2020.03.038>, <https://www.sciencedirect.com/science/article/pii/S092523122030391X>
11. M. Geetha, P.V. Aswathi, in *Dynamic Gesture Recognition of Indian Sign Language Considering Local Motion of Hand Using Spatial Location of Key Maximum Curvature Points*. 2013 IEEE Recent Advances in Intelligent Computational Systems (RAICS) (2013), pp. 86–91. <https://doi.org/10.1109/RAICS.2013.6745452>
12. Y. Chen, L. Zhao, X. Peng, J. Yuan, D. Metaxas, *Construct Dynamic Graphs for Hand Gesture Recognition Via Spatial-Temporal Attention* (2020) (funding Information: This work was funded partly by ARO-MURI-68985NSMUR and NSF 1763523, 1747778, 1733843, 1703883 grants to Dimitris N. Metaxas. Publisher Copyright: © 2019. The copyright of this document resides with its authors.; 30th British Machine Vision Conference, BMVC 2019 ; Conference date: 09–09–2019 Through 12–09–2019)
13. C. Poonam, T. Ghorpade, P. Padiya, in *Indian Sign Language to Forecast Text Using Leap Motion Sensor and rf Classifier*. 2016 Symposium on Colossal Data Analysis and Networking (CDAN) (2016), pp. 1–5. <https://doi.org/10.1109/CDAN.2016.7570936>
14. M. Anshul, P. Kumar, P.P. Roy, R. Balasubramanian, B.B. Chaudhuri, A modified lstm model for continuous sign language recognition using leap motion. *IEEE Sens. J.* **19**(16), 7056–7063 (2019). <https://doi.org/10.1109/JSEN.2019.2909837>
15. P. Kumar, P.P. Roy, D.P. Dogra, Independent Bayesian classifier combination based sign language recognition using facial expression. *Inf. Sci.* **428**, 30–48 (2018). <https://doi.org/10.1016/j.ins.2017.10.046>, <https://www.sciencedirect.com/science/article/pii/S0020025516307897>
16. G. Marin, F. Dominio, P. Zanuttigh, in *Hand Gesture Recognition with Leap Motion and Kinect Devices*. 2014 IEEE International Conference on Image Processing (ICIP) (2014), pp. 1565–1569. <https://doi.org/10.1109/ICIP.2014.7025313>
17. N. Rossol, I. Cheng, A. Basu, A multisensor technique for gesture recognition through intelligent skeletal pose analysis. *IEEE Trans. Hum. Mach. Syst.* **46**(3), 350–359 (2016). <https://doi.org/10.1109/THMS.2015.2467212>
18. K. Pradeep, H. Gauba, P.P. Roy, D.P. Dogra, A multimodal frame-work for sensor based sign language recognition. *Neurocomputing* **259**, 21–38 (2017). <https://doi.org/10.1016/j.neucom.2016.08.132>, <https://www.sciencedirect.com/science/article/pii/S092523121730262X> (multimodal media data understanding and analytics)
19. B.K., Neel, Y. Vishnusai, G.N. Rathna, in *Indian Sign Language Gesture Recognition Using Image Processing and Deep Learning*. 2019 Digital Image Computing: Techniques and Applications (DICTA), (2019), pp. 1–8. <https://doi.org/10.1109/DICTA47822.2019.8945850>
20. Multid-cnn: A multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in rgb-d image sequences. *Expert Syst. Appl.* **139**, 112829 (2020). <https://doi.org/10.1016/j.eswa.2019.112829>
21. K. Lim, A. Tan, C.P. Lee, S. Tan, Isolated sign language recognition using convolutional neural network hand modelling and hand energy image. *Multimedia Tools Appl.* **78** (2019). <https://doi.org/10.1007/s11042-019-7263-7>
22. B. Divya, J. Delpha, S. Badrinath, in *Public Speaking Words (Indian Sign Language) Recognition Using Emg*. 2017 International Conference on Smart Technologies for Smart Nation (SmartTechCon) (2017), pp. 798–800. <https://doi.org/10.1109/SmartTechCon.2017.8358482>
23. P.V.V. Kishore, D.A. Kumar, A.C. Sastry, E.K. Kumar, Motionlets matching with adaptive kernels for 3-d Indian sign language recognition. *IEEE Sens. J.* **18**(8), 3327–3337 (2018). <https://doi.org/10.1109/JSEN.2018.2810449>

24. E.J. Cardenas, G.C. Chavez, Multimodal hand gesture recognition combining temporal and pose information based on cnn descriptors and histogram of cumulative magnitudes. *J. Vis. Commun. Image Representation* **71**, 102772 (2020). <https://doi.org/10.1016/j.jvcir.2020.102772>
25. C. Wei, W. Zhou, J. Pu, H. Li, in *Deep Grammatical Multi-Classifer for Continuous Sign Language Recognition*. 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM) (2019), pp. 435–442. <https://doi.org/10.1109/BigMM.2019.00027>

OntoReqC: An Ontology Focused Integrative Approach for Classification of Software Requirements



R. Dheenadhayalan and Gerard Deepak

Abstract Software products start to develop based on the client requirements that the software demands and classified requirements can result in well-developed software. This paper proposes the OntoReqC framework, a requirements classifier that has an accuracy of 95.49%. OntoReqC amalgamates ontologies, domain markers and is optimized with the ant colony optimization algorithm. This model obtains functional and nonfunctional ontologies from the unified modeling language models and software requirement artifact using OntoCollab and Protégé. Domain markers are derived from software requirements specification document using linked open data cloud for domain enrichment making this model knowledge-based. Ontologies and domain markers are combined to get content development network which is then semantically aligned with requirement terms obtained from preprocessed requirements dataset. Ant colony optimization has been applied to the results, producing the classified requirements.

Keywords Domain knowledge · Knowledge centric approach · Ontology · Requirements classification · Requirement engineering

1 Introduction

Requirement engineering is the very basic part and the first stage of a software development process and is the key for developing precise software which aligns with the users' and stakeholders' exact needs. Error or miscalculation in the requirement engineering will become very costly to compensate as the software evolve. If it is not corrected in the early stages of development, a single mistake can cost more than the

R. Dheenadhayalan
Department of Computer Science and Engineering, Indian Institute of Information Technology
Kottayam, Kottayam, India

G. Deepak (✉)
Department of Computer Science and Engineering, National Institute of Technology,
Tiruchirappalli, India

allotted resources in terms of funds, labor, and time. It is important to have unambiguous, complete, and correct requirements which are some of the key factors that determine a software product's quality. Requirement engineering has many phases like requirement elicitation, analysis, specification, alignment, management, etc. This study mainly centered on requirements alignment or requirements classification.

Motivation: Requirements linking and classification can lead to an accurate testing plan which can effectively minimize the project cost and time for development [1, 2], thus making this one of the most important phases of software engineering. Better and less time-consuming method to classify the obtained requirements must be employed to speed up the software development cycle. This study aims to propose such a method that accomplishes this task.

Contribution: OntoReqC which is an Ontology-based and knowledge driven framework for requirement classification has been proposed. A software requirement dataset has been employed to train the model and optimize it using the ant colony optimization algorithm. The approach is based on generation of domain markers, functional requirements, and modeling of nonfunctional requirements from software artifacts. Ontology merging has been performed to generate a content development network and approach uses recurrent neural network for classification and linked open data cloud for auxiliary knowledge enrichment. The model is evaluated for precision, recall, accuracy, and f -measure and has found to be improved based on these factors. The proposed OntoReqC model achieves 95.49% accuracy and false negative rate of 0.03.

Organization: Sect. 2 of this paper provides the related works. Sections 3 and 4 explain the proposed methodology and its implementation, respectively. Results are analyzed, and the proposed methodology is validated in Sect. 5. Conclusions drawn from the observation are stated in Sect. 6.

2 Related Work

A method for classifying functional requirements by merging five machine learning (ML) techniques is proposed by Rahimi et al. [3], and the techniques utilized are Naïve Bayes, logistic regression, decision tree, support vector machine (SVM), and support vector classification (SVC). It is observed that the best time was exhibited by the model which uses less number of classifiers and only a less change in the accuracy was observed when comparing the faster model to the most accurate model. In a work published by Binkhonain et al. [4], authors considered 24 ML-based approaches for nonfunctional requirements classification. In these methods, 16 unique ML algorithms were found to be used commonly in which the supervised learning approach has been the most popular in providing the best solution. Authors in this paper find that ML techniques have a greater potential in classifying and identifying requirements. ML in requirement engineering may give rise to novel, expert, and brilliant

systems which can greatly support the operations on software requirements. Younas et al. [5] used a semi-supervised ML method to set apart nonfunctional requirements from the functional requirements given the requirements dataset. Textual semantic similarities of functional requirements are identified and used to differentiate them from nonfunctional requirements to better extract the nonfunctional requirements. In their study, Word2Vec model is used to determine semantic similarity distance and it is trained with repositories of Wikipedia. First, nonfunctional requirements extraction is done with traditional preprocessing, and then, POS tagged preprocessing and word augmentation are administered to result in enhanced requirements classifying system. In [6–18], several ontological models in support of the proposed literature have been depicted.

3 Proposed System Architecture

Functional ontologies of the software are generated from its unified modeling language (UML) models such as use case, sequence, state chart, and class diagrams. A collaborative ontology modeling mechanism known as OntoCollab [19] is used for the generation of functional ontologies. Nonfunctional ontologies are obtained from software requirement artifact by manually modeling them with Web Protégé in the proposed system architecture (Fig. 1).

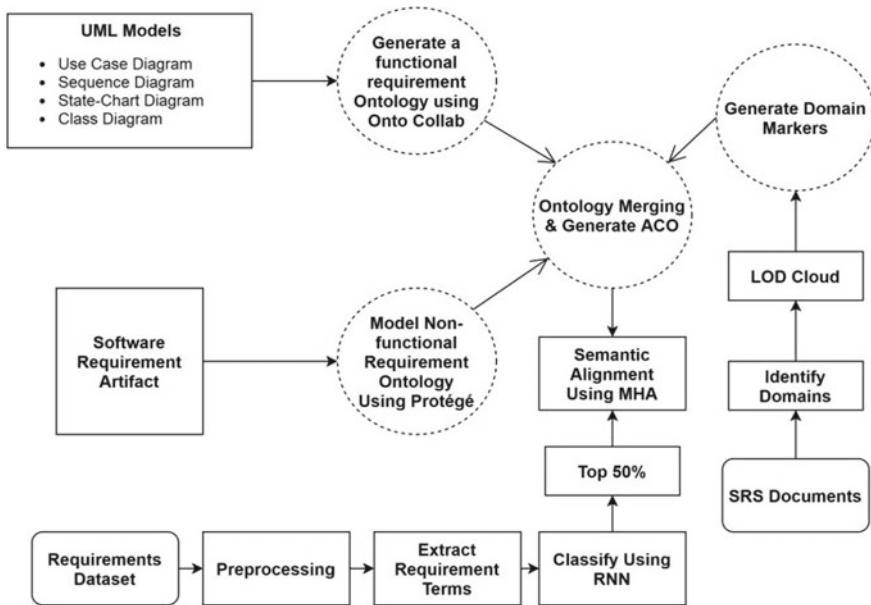


Fig. 1 Architecture for the proposed OntoReqC

Domains of the software are identified based on software requirement specification (SRS) document and entities of the derived domains are further enriched using linked open data (LOD) cloud which is queried using SPARQL endpoint, thus resulting in the required domain markers. Generated functional ontologies, nonfunctional ontologies, and domain markers are merged, and based on these data, a content development network (CDN), which is a network of relevant ontologies, is derived.

The requirement dataset of the software is preprocessed using natural language processing techniques like lemmatization, tokenization, stop word removal, and named entity recognition. Requirement terms are extracted, and classification of requirements is then done by using recurrent neural network (RNN) which maps each requirement to appropriate classes. RNN achieves this by processing the current input vector and its previous state with a recurrence formula. If x_i and h_i denote the i th input vector and the solution produced by recurrence formula for the i th input, respectively, then the recurrence formula is given by Eq. (1).

$$h_t = f(h_{t-1}, x_t) \quad (1)$$

For some instance t , where each successive instance is called a time stamp.

Merged ontologies, generated CDN, and the top 50% of the classified requirements from the requirement dataset are semantically aligned using ant colony optimization (ACO) algorithm. Using Jaccard similarity [20], normalized Google distance [21], Shannon's entropy [22], and information gain [23] as initial objective functions in ACO, requirements are classified. The reason for considering ACO as the meta-heuristic algorithm for the semantic alignment is because of its ability to generate best solution set between requirements classified by RNN and generated CDN. In this study, daemon actions check for semantic similarity by ensuring threshold in each case.

4 Implementation

The proposed methodology is implemented in Python. Google colab notebook is utilized, which runs on 12 GB NVIDIA Tesla K80 GPU. The latest software requirements dataset is taken from Kaggle datasets. Dataset is labeled and contains various functional and nonfunctional requirements of different kinds of software and the labels specifying the type of the requirements.

Algorithm 1: Proposed OntoReqC Algorithm

Input:	UML Models, Software Requirement Artifact, SRS Documents, Requirements Dataset
Output:	Requirements that are categorized into a different classification
<i>Start</i>	
Step 1:	Based on the UML model of the software, generate functional requirement Ontology using OntoCollab
Step 2:	Based on the Software Requirement Artifact manually model a non-functional requirement Ontology using Protégé
Step 3:	Obtain SRS Document of the software Identify the domain of the software Enrich the domain using LOD Cloud queried using SPARQL endpoints to obtain domain markers
Step 4:	Functional and Non-Functional Ontologies are merged
Step 5:	Merged Ontologies and domain markers are used to generate CDN, a network of relevant Ontologies
Step 6:	Requirements dataset is pre-processed using Tokenization, Lemmatization, Stop Word Removal, and Named Entity Recognition
Step 7:	Requirement terms are extracted from pre processed Requirements dataset
Step 8:	Requirement terms are then classified using RNN
Step 9:	Top 50% of the classified Requirements terms is semantically aligned with generated CDN using ACO algorithm resulting in the classified requirements Perform ACO_Metaheuristics as `while(Entities in Classified Instances && CDN! =NULL) (1) Using Jaccard Similarity, Normalized Google Distance, Shannon's Entropy as initial Objective Functions, generate_Initial_Solutions() by thresholding with 0.5. (2) Check if the entities in the Initial Solution Set correlate with Shannon's Entropy and Information Gain with a maximum difference not exceeding 0.25 as ActionUpdate() Set. (3) Update the entities matching the CDN and the Classified instances of the RNN in the pheromone_UpdateSet(). end while
<i>End</i>	

From Algorithm 1, it is clear that the proposed OntoReqC needs UML models, software requirement artifact, SRS document, and requirement dataset to produce well-classified requirements. UML models produce functional requirement ontology using OntoCollab, and software requirement artifact is used to manually generate nonfunctional requirement ontology using Protégé. Domain markers are obtained by enhancing the domain derived from SRS document using LOD cloud by SPARQL endpoint queries. Functional ontology, nonfunctional ontology, and domain markers are merged, resulting in CDN. Requirement terms are extracted from the requirement dataset which is preprocessed and then they are classified by the RNN algorithm. Top 50% of the classified requirements and the obtained CDN are semantically aligned using ACO algorithm resulting in the classified requirements which is the output of the proposed OntoReqC framework.

5 Results and Performance Evaluation

Proposed OntoReqC model along with OntoReqC without ACO, OntoReqC without ontologies, and OntoReqC without LOD cloud are also subjected to the experiment, and for control, two other methodologies, methodology for the classification of quality of requirements (MCQR) proposed by Parra et al. [24] and software requirements classification (SRC) proposed by Dias Canedo et al. [25], were also considered for evaluation. *F*-measure and false negative rate (FNR) are preferred metrics for measuring the performance of this system, and the results are recorded in Table 1.

It is seen that performance of MCQR and SRC is significantly less than OntoReqC. MCQR mainly uses the bagging and boosting ensemble method to further optimize their model. This technique considers many learners or classifiers, and the results

Table 1 Performance of the proposed OntoReqC and other chosen approaches

Classification method	Average precision (%)	Average recall (%)	Accuracy (%)	<i>F</i> -measure (%)	FNR
MCQR [7]	81.89	84.72	83.22	83.28	0.15
SRC [8]	84.69	87.37	85.32	86.01	0.13
Proposed OntoReqC	93.72	96.69	95.49	95.18	0.03
Eliminating ACO from OntoReqC	88.72	90.74	89.27	89.72	0.09
Eliminating ontologies from OntoReqC	84.22	87.62	85.38	85.89	0.12
Eliminating LOD cloud from OntoReqC	85.21	88.71	87.02	86.92	0.11

obtained from the model as a whole is the combination of results produced by the individual learners. The bagging method can be ignorant to the magnitude of the results produced by different learners and gives an average result which may lead to the decreased accuracy of the model and this method also does not fit if there is bias or underfitting in the data. Variances error and overfitting in the dataset cannot be handled by boosting, thus rendering this less effective in these cases. In the SRC model, authors utilize term frequency-inverse document frequency (TF-IDF) to classify requirements but TF-IDF can only be used in to analyze lexical level statements but not the semantic values. Neither MCQR nor SRC use ontologies or the auxiliary knowledge like which is provided by LOD cloud which makes them less effective than OntoReqC model.

From Table 1, it can be inferred that the ACO, ontologies, and LOD cloud are necessary part of the OntoReqC. This is evident from the fact that by eliminating these steps from OntoReqC, there is a significant drop in the performance of the model. LOD cloud uses auxiliary knowledge which makes this a knowledge driven model and it also populates the domain entity which proceeds to feed data with high information density into the classifier making classification highly accurate. The performance drop observed at the elimination of ACO is due to the lack of optimization of the requirements term. In ontology merging, concepts and sub-concepts are merged using concept similarity, taking in account up to five levels of nodes near to the considered node in the ontology. Individual matching is also done by using concept similarity, finally resulting in well linked ontologies for the model to make better prediction, hence making ontology merging a necessary method. A drop in accuracy of 10.11 and 8.47% can be observed in the model due to the elimination of ontologies and LOD cloud making them the most important part of the proposed methodology and eliminating ACO also caused a 6.22% degradation in the accuracy.

Figure 2 provides better visualization for comparing different search techniques and the significance of the algorithms used in OntoReqC. It can be clearly seen that OntoReqC performs significantly better than the other models. By eliminating ACO from OntoReqC, the precision drops to about 5%. If LOD cloud is not used in the proposed model, then there is a precision drop of 8.51%, and by eliminating ontologies from the algorithm, there is a 9.5% reduction in the performance of the system. From this observation, it can be concluded that ontologies are a very important part of OntoReqC and the elimination of this causes the system to lose most of its precision and then LOD cloud, which also contributes a significant change in the precision of the system. Therefore, ontologies play a vital role than LOD cloud and LOD cloud than ACO in improving precision of the proposed methodology.

6 Conclusions

By observing the results, it can be stated that the proposed OntoReqC method can be a highly functional model by utilizing ontologies, LOD cloud, and ant colony

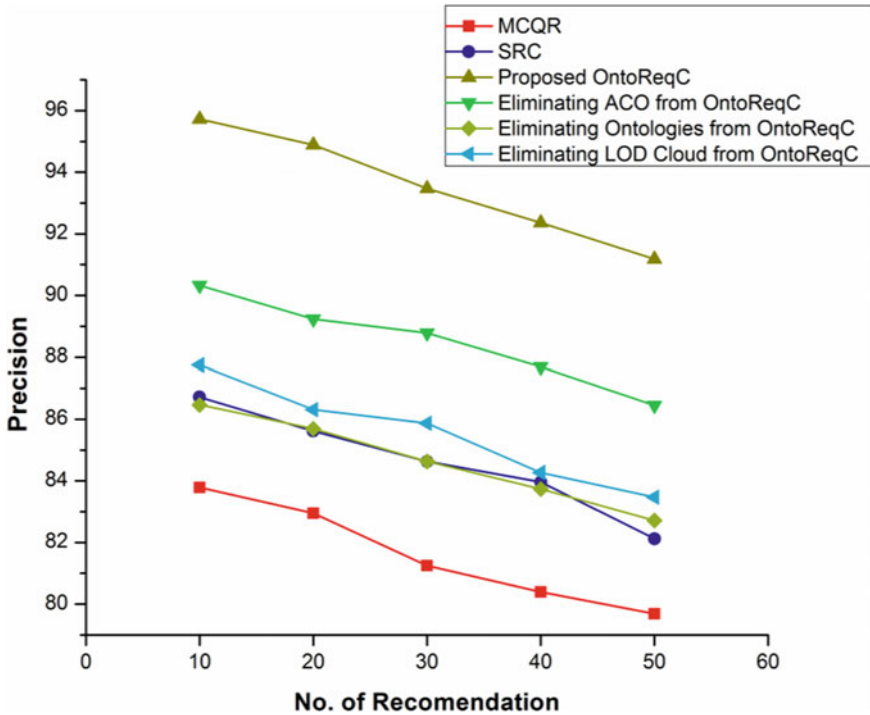


Fig. 2 Precision percentage versus no. of recommendation

optimization, providing 95.49% accurate classification of the requirements. Ontologies and LOD cloud majorly contributed to the accuracy and precision of this model. OntoReqC also exhibited a very low FNR of 0.03 compared to other considered models. Enhancements in the precision, recall, and accuracy of the model are achieved owing to well-classified ontologies, auxiliary knowledge from the LOD cloud, and optimization of the ACO. Therefore, OntoReqC can be a potential model to classify software requirements for manufacturing excellent quality products at a faster phase. Future experimentation on this model can be by changing the meta-heuristic algorithm or by adding other parameters. UML models can be summarized, and story boarding can be carried out on UML models and requirements.

References

1. Z.A. Barmi, A.H. Ebrahimi, R. Feldt, in *Alignment of Requirements Specification and Testing: A Systematic Mapping Study*. 2011 IEEE Fourth International Conference on Software Testing, Verification and Validation Workshops (IEEE, 2011 March), pp. 476–485
2. J. Kukkanen, K. Väkeväinen, M. Kauppinen, E. Uusitalo, in *Applying a Systematic Approach to Link Requirements and Testing: A Case Study*. 2009 16th Asia-Pacific Software Engineering

- Conference (IEEE, 2009 December), pp. 482–488
3. N. Rahimi, F. Eassa, L. Elrefaei, An ensemble machine learning technique for functional requirement classification. *Symmetry* **12**(10), 1601 (2020)
 4. M. Binkhonain, L. Zhao, A review of machine learning algorithms for identification and classification of non-functional requirements. *Expert Syst. Appl.* **X**, **1**, 100001 (2019)
 5. M. Younas, D.N. Jawawi, I. Ghani, M.A. Shah, Extraction of non-functional requirement using semantic similarity distance. *Neural Comput. Appl.* **32**(11), 7383–7397 (2020)
 6. V. Adithya, G. Deepak, in *OntoReq: An Ontology Focused Collective Knowledge Approach for Requirement Traceability Modelling*. European, Asian, Middle Eastern, North African Conference on Management & Information Systems (Springer, Cham, 2021 March), pp. 358–370
 7. V. Adithya, G. Deepak, A. Santhanavijayan, in *HCODF: Hybrid Cognitive Ontology Driven Framework for Socially Relevant News Validation*. International Conference on Digital Technologies and Applications (Springer, Cham, 2021 January), pp. 731–739
 8. G.L. Giri, G. Deepak, S.H. Manjula, K.R. Venugopal, in *OntoYield: A Semantic Approach for Context-Based Ontology Recommendation Based on Structure Preservation*. Proceedings of International Conference on Computational Intelligence and Data Engineering: ICCIDE 2017, vol. 9 (Springer, 2017 December), p. 265
 9. G. Deepak, N. Kumar, A. Santhanavijayan, A semantic approach for entity linking by diverse knowledge integration incorporating role-based chunking. *Procedia Comput. Sci.* **167**, 737–746 (2020)
 10. G. Deepak, S. Rooban, A. Santhanavijayan, A knowledge centric hybridized approach for crime classification incorporating deep bi-LSTM neural network. *Multimedia Tools Appl.* 1–25 (2021)
 11. K. Vishal, G. Deepak, A. Santhanavijayan, in *An Approach for Retrieval of Text Documents by Hybridizing Structural Topic Modeling and Pointwise Mutual Information*. Innovations in Electrical and Electronic Engineering (Springer, Singapore, 2021), pp. 969–977
 12. G. Deepak, V. Teja, A. Santhanavijayan, A novel firefly driven scheme for resume parsing and matching based on entity linking paradigm. *J. Discrete Math. Sci. Crypt.* **23**(1), 157–165 (2020)
 13. G. Deepak, N. Kumar, G.V.S.Y. Bharadwaj, A. Santhanavijayan, in *OntoQuest: An Ontological Strategy for Automatic Question Generation for e-Assessment Using Static and Dynamic Knowledge*. 2019 Fifteenth International Conference on Information Processing (ICINPRO), (IEEE, 2019 December), pp. 1–6
 14. G. Deepak, A. Santhanavijayan, OntoBestFit: a best-fit occurrence estimation strategy for RDF driven faceted semantic search. *Comput. Commun.* **160**, 284–298 (2020)
 15. M. Arulmozhivarman, G. Deepak, in *OWLW: Ontology Focused User Centric Architecture for Web Service Recommendation Based on LSTM and Whale Optimization*. European, Asian, Middle Eastern, North African Conference on Management & Information Systems (Springer, Cham, 2021 March), pp. 334–344
 16. G. Deepak, J.S. Priyadarshini, Personalized and enhanced hybridized semantic algorithm for web image retrieval incorporating ontology classification, strategic query expansion, and content-based analysis. *Comput. Electr. Eng.* **72**, 14–25 (2018)
 17. Z. Gulzar, A.A. Leema, G. Deepak, Pcrs: Personalized course recommender system based on hybrid approach. *Procedia Comput. Sci.* **125**, 518–524 (2018)
 18. G. Deepak, J.S. Priyadarshini, in *A Hybrid Semantic Algorithm for Web Image Retrieval Incorporating Ontology Classification and User-Driven Query Expansion*. Advances in Big Data and Cloud Computing (Springer, Singapore, 2018), pp. 41–49
 19. C.N. Pushpa, G. Deepak, J. Thriveni, K. Venugopal, in *OntoCollab: Strategic Review Oriented Collaborative Knowledge Modeling Using Ontologies*. 2015 Seventh International Conference on Advanced Computing (IEEE, 2015 December) (ICoAC), pp. 1–7
 20. P. Jaccard, The distribution of the flora in the alpine zone. 1. *New Phytol.* **11**(2), 37–50 (1912)
 21. R.L. Cilibrasi, P.M.B. Vitanyi, The Google similarity distance. *IEEE Trans. Knowl. Data Eng.* **19**(3), 370–383 (2007)

22. C.E. Shannon, W. Weaver, *The Mathematical Theory of Communication* (Univ of Illinois Press, 1949). ISBN 0-252-72548-4
23. J.R. Quinlan, Induction of decision trees. *Mach. Learn.* **1.1**, 81–106 (1986)
24. E. Parra, C. Dimou, J. Llorens, V. Moreno, A. Fraga, A methodology for the classification of quality of requirements using machine learning techniques. *Inf. Softw. Technol.* **67**, 180–195 (2015)
25. E. Dias Canedo, B. Cordeiro Mendes, Software requirements classification using machine learning algorithms. *Entropy* **22**(9), 1057 (2020)

SemUserProfiling: A Hybrid Knowledge Centric Approach for Semantically Driven User Profiling



Rituraj Ojha and Gerard Deepak

Abstract Information is shared by a large number of people around the globe in the form of chats, posts, blogs, tweets, and news. For information to reach the right audience and for companies to target the right people, user profiling and term profiling are much needed. In this paper, SemUserProfiling which is an entity enrichment and structural topic modeling (STM)-based approach is proposed. The proposed approach uses the Twitter dataset, and user profile Up as an input. The Twitter dataset is preprocessed, and scenario is retrieved for each term in the preprocessed dataset. Furthermore, the entity enrichment takes place using linked open data (LOD) cloud and classification of the entity set happens using eXtreme gradient boosting (XGBoost) algorithm using categorical domain ontologies. Similarly, user profile Up is preprocessed and subjected to topic modeling using STM, and entity integration takes place using LOD cloud. Finally, the semantic similarity is calculated between the enriched user profile terms and classified enriched entity set using entropy and normalized Google distance (NGD) under frog leap algorithm. The proposed SemUserProfiling yields a high accuracy of 99.02% and a negligible false discovery rate of 0.011.

Keywords Entity enrichment · Scenario retrieval · Topic modeling · User profiling

1 Introduction

Social media is a digital platform that helps to create and share information, interests, ideas, and to develop new social connections with other Internet profiles. Users generally access social media sites through desktop apps, mobile apps, or Wweb-based apps. Few famous social media platforms include Twitter, Facebook, Quora,

R. Ojha

Department of Metallurgical and Materials Engineering, National Institute of Technology, Tiruchirappalli, India

G. Deepak (✉)

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

Instagram, Snapchat, Reddit, and LinkedIn. Professionals use these platforms to share knowledge, express ideas, and resolve doubts.

On social media, users usually form groups to share and receive knowledge of their interests. The number of groups increase as the number of users with different interests increase on the platform. The user posts his ideas in the group, which has its id and timestamp. The platform provides features such as likes and comments on every post to encourage and provide feedback. The user's post reach depends on the relevance of the post with the people reading, and thus, categorizing the user is essential for social media to recommend the right post to the right audience.

Motivation: There may be times during gathering of users when discussions or debates start on different range of topics. Therefore, profiling becomes vital to attract similar audiences and encourage knowledge. There is no framework which achieves profiling in social media with high accuracy. Also, in the era of semantic Web, there is a need for a better profiling algorithm which is semantic driven, and thus, building this user profiling approach was a major motivating factor.

Contribution: The SemUserProfiling based on entity enrichment and STM for profiling users is proposed. The Twitter dataset drives the proposed approach. The dataset is preprocessed, and scenario is retrieved for each term. Moreover, the entity enrichment takes place using the LOD cloud and classification of the entity set happens using XGBoost algorithm using categorical domain ontologies. Similarly, user profile Up is preprocessed and subjected to topic modeling using STM, and entity integration takes place using LOD cloud. Finally, the semantic similarity is calculated between the enriched user profile terms and classified enriched entity set using entropy and NGD under frog leap algorithm. The values of metrics like precision, accuracy, recall, and F -measure of the proposed framework are increased.

Organization: The flow of the remaining paper is as follows. A condensed summary of the related works is provided in Sect. 2. Section 3 depicts the architecture of the proposed system. Section 4 describes the architecture implementation. Section 5 presents the evaluation of results and performance. The conclusion of the paper is presented in Sect. 6.

2 Related Works

Several people have done research in the area of profiling users. López-Monroy et al. [1] have proposed a methodology for profiling documents. The proposed approach captures discriminative and sub-profile information of terms, and then, these representations are accumulated to represent the content of the document. Van Dam and Van De Velden [2] have proposed a framework to learn and understand the user's data shared on social media platform Facebook. The proposed approach helps individuals to find the profiles of the target users based on their interest toward a particular field.

Chen et al. [3] have proposed a technique for estimating the location of users present on Twitter social media platform. The proposed approach can find location up to city level using the user's network, data shared by them, and tie strength. Greco and Polli [4] have proposed a methodology to analyze the textual data available on social media platforms and identify the sentiments and opinion of users on the particular topic. This paper proposes a better tool focused mainly on brand management.

Mishra et al. [5] have proposed an approach for profiling users for detection of abusive and hateful comments on social media platforms. The proposed technique uses community-based profiling characteristics for the social media users. Mishra et al. [6] have proposed SNAP-BATNET, a deep learning framework to find the suicidal tendency of the user. The proposed approach uses the data in the form of social graph embeddings. It also profiles users based on features from their previous data.

Wisniewski et al. [7] have proposed a methodology to suggest Facebook users with a set of distinct privacy options based on the user profiling. The proposed approach uses advanced factor analysis methods to present several privacy management strategies. Kosmajac and Keselj [8] have proposed a technique for bot and gender identification for Twitter using feature extraction, user behaviors fingerprints, syntactic information, and transformation methods.

Singh et al. [9] have proposed a user behavior profiling approach to detect threats present within an organization. The proposed approach uses an ensemble hybrid machine learning model. This machine learning model uses multistate long short-term memory and CNN. Chen et al. [10] have proposed a methodology for classifying the sentiments of a user in a document for a particular product. This is done using their proposed hierarchical neural network.

Guimaraes et al. [11] have proposed a technique for determining characteristics and age of a person by analyzing user profile and historical data. The proposed techniques use several approaches for this problem and the deep convolutional neural network archives the best accuracy. Menini et al. [12] have proposed a system for identifying cyberbullying on social media platforms like Instagram and Twitter by combining classification and social network analysis techniques. Papers [13–19] have proposed several approaches based on ontologies and knowledge bases.

3 Proposed System Architecture

The architecture for user profiling is depicted in Fig. 1. The socially aware user profiling takes place in several steps. Initially, the Twitter dataset (tweet dataset) is preprocessed using tokenization, lemmatization, stop word removal, and named entity recognition (NER). Tokenization is a process of splitting the texts in the dataset into pieces called tokens. Byte pair encoding (BPE) is used for the tokenization process. Lemmatization involves grouping together several inflected kinds of the same word so that they can be analyzed as a single term. During stop word removal, the common ubiquitous words are removed as they add no value for the analysis and

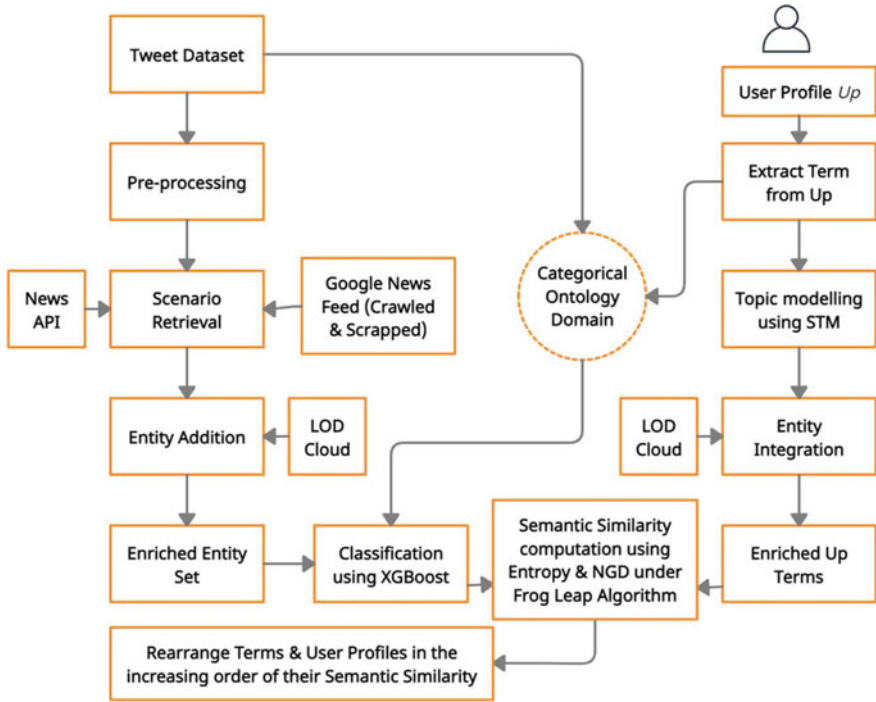


Fig. 1 Proposed system architecture

only increase the dimension of the feature set. NER is the process of finding and categorizing the data or entity into predefined categories.

The next step involves retrieving scenarios for the preprocessed tweets. The individual terms from the preprocessed tweets are taken and for these individual terms, the scenario is obtained by crawling and scraping Google news feed and other news APIs, namely mediastack, news API, and Webhose. Furthermore, real world similar entities are also added from the LOD cloud. After these processes, we obtain the enriched entity set. The next step involves classifying the enriched entity set using categorical domain ontologies using the XGBoost algorithm. XGBoost is an ensemble machine learning model that is based on the decision tree. Features of XGBoost include handling of missing values automatically, supports parallel processing, tree pruning, and regularized boosting to prevent overfitting. Equation (1) presents how gradient tree boosting works.

$$F_m(X) = F_{m-1}(X) + \alpha_m h_m(X, r_{m-1}) \tag{1}$$

where α_i and r_i represent the regularization parameters and residuals calculated with the i th tree. The function for predicting residuals, r_i using X for the i th tree, is represented by h_i . To calculate the α_i , we use r_i and calculate the following: $\arg \arg =$

$\sum_{i=1}^m [L(Y_i, F_{i-1}(X_i)) + \alpha h_i(X_i, r_{i-1})]$ where $L(Y, F(X))$ is a differentiable loss function.

The user profile Up is preprocessed, and the terms are extracted from it. We generate the categorical domain ontologies between the tweet dataset and the extracted terms from user profile, and these categorical domain ontologies are used for the classification process using XGBoost algorithm. Furthermore, STM is applied using document corpus and extracted user profile data. The document corpus is crawled using beautiful soup API. This helps our model to get more topics that are similar to the user profile terms. Topic modeling is a technique to discover different topics present in the document corpus and get information about the hidden patterns exhibited by the document corpus. The STM is used for topic modeling that accommodates corpus structure through covariates at document-level. This model consolidates and expands three models: the Dirichlet-multinomial regression topic model, the correlated topic model, and the sparse additive generative topic model [20].

Furthermore, the user profile data are enriched with the linked data from LOD cloud. Finally, the semantic similarity is calculated between the enriched user profile terms and classified enriched entity set using entropy and normalized Google distance under the frog leap algorithm, which is the preferred optimization metaheuristics. Entropy is represented by Eq. (2) and it represents the uncertainty or surprise that a variable can output. NGD between search terms x and y is represented by Eq. (3) and it is a semantic similarity measure. The calculation is derived by measuring Google returning the number of hits for a set of keywords. The NGD produces results such that keywords with similar meaning tend to be close as compared to keywords with dissimilar meaning, which are far apart. The frog leap algorithm is an optimization algorithm based on the actions observed in a family of frogs when they search for the place that has the highest quantity of available food. Their population consists of the group of frogs that are further divided into subsets called memplexes. Each of the memplexes perform a local search, and the ideas are later passed between them during shuffling. The shuffling process and the local search are continued until a defined convergence criterion is fulfilled [21].

$$\text{Entropy}(P) = - \sum_{i=1}^N P_i \log_2 P_i \quad (2)$$

$$\text{NGD}(x, y) = \frac{\max\{\log \log f(x), \log \log f(y)\} - \log f(x, y)}{\log \log N - \min\{\log \log f(x), \log \log f(y)\}} \quad (3)$$

Finally, after all these steps, the terms and user profiles are rearranged. The rearrangement takes place in the increasing order of their semantic similarity.

4 Implementation

The algorithm for the proposed approach is depicted in Algorithm 1, which takes Twitter dataset and user profile as input. The Twitter dataset is preprocessed using tokenization, lemmatization, stop word removal, and NER. The preprocessing is done using several Python libraries. The scenario is retrieved for each term using Google news feed API and several other news APIs, namely mediastack, news API, and Webhose. Moreover, the entity enrichment takes place using LOD cloud and classification of entity sets happen using XGBoost algorithm using categorical domain ontologies. Similarly, user profile Up is preprocessed and is subjected to topic modeling using STM, and entity integration takes place using LOD cloud. Eventually, the semantic similarity is calculated between the enriched user profile terms and the classified enriched entity set using entropy and NGD under frog leap algorithm, to obtain the rearranged user profiles and terms in increasing order of the calculated semantic similarity. The proposed SemUserProfiling was successfully implemented and evaluated using Windows 10 operating system, equipped with 8th generation Intel Core i5 and 16 GB RAM, in Google collaboratory environment with Nvidia graphics card.

Algorithm 1: Proposed SemUserProfiling Algorithm

Input:	Twitter Dataset & User Profile Up
Output:	Rearranged User profiles & Terms
	<i>begin</i>
Step 1:	Twitter dataset D is subjected to preprocessing. D is tokenized, lemmatized, and stop word removal and NER is performed on it.
Step 2:	while ($D.next() \neq \text{NULL}$) D set \leftarrow scenario is retrieved using Google News Feed API and News API end while
Step 3:	while ($D.next() \neq \text{NULL}$) D set \leftarrow relevant entities using LOD Cloud end while
Step 4:	User profile Up is pre-processed and the terms are extracted from it
Step 5:	Categorical Domain Ontologies generated using Up terms and Twitter dataset
Step 6:	6.1: STM (document corpus, user profile data Up) 6.2: Up set.append(terms relevant to Up)
Step 7:	while ($Up.next() \neq \text{NULL}$) Up set \leftarrow relevant entities using LOD cloud end while
Step 8:	Classification of D set using XGBoost algorithm (using Categorical Domain Ontologies)
Step 9:	SemanticSimilarity(D set, Up set)
Step 10:	Rearranged User profiles & Terms in increasing order of the calculated semantic similarity.
	<i>end</i>

5 Results and Performance Evaluation

The performance of the proposed SemUserProfiling is measured by considering precision, recall, and accuracy. Other measures including false discovery rate, F -measure, and normalized discounted cumulative gain are also used. The performance is evaluated for 6129 queries, and the ground truth has been collected.

$$\text{Precision} = \frac{\text{Retrieved} \cap \text{Relevant}}{\text{Retrieved}} \quad (4)$$

$$\text{Recall} = \frac{\text{Retrieved} \cap \text{Relevant}}{\text{Relevant}} \quad (5)$$

$$\text{Accuracy} = \frac{\text{Proportion Corrects of each query passed ground truth test}}{\text{Total number of queries}} \quad (6)$$

$$F\text{-Measure} = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (7)$$

$$\text{False Discovery Rate} = 1 - \text{Positive Predictive Value} \quad (8)$$

$$\text{nDCG} = \frac{\text{DCG}_\alpha}{\text{IDCG}_\alpha} \quad (9)$$

$$\text{DCG} = \sum_{i=1}^{\alpha} \frac{\text{rel}_i}{\log_2(i+1)} \quad (10)$$

Equations (4–6) represent precision, recall, and accuracy, respectively. Furthermore, Eqs. (7–10) represent F -measure, false discovery rate (FDR), normalized discounted cumulative gain (nDCG), and discounted cumulative gain, respectively. The reason for considering accuracy, precision, recall, and F -measure metrics for evaluation is because they measure the relevance of the results, and FDR computes the number of false discoveries made by the proposed system. The nDCG represents the diversity of the relevant results.

Table 1 presents the performance comparison of the proposed SemUserProfiling model with Wisely et al. [22], PSMU [23], and other similar baseline models. It is evident from the table that the proposed approach achieves the highest average accuracy with the precision of 98.87%, recall of 99.87%, accuracy of 99.02%, F -measure of 99.36%, and nDCG value of 0.97. Also, the FDR is least with the value of 0.011. Wisely et al. achieves a low precision of 84.89%, recall of 92.32%, accuracy of 87.32%, F -measure of 87.52%, high FDR of 0.151, and a low nDCG value of 0.81. Furthermore, PSMU achieves a low precision of 95.32%, recall of 98.31%, accuracy of 96.71%, F -measure of 96.79%, high FDR of 0.047, and low nDCG value of 0.84.

Figure 2 represents the precision % versus number of recommendations for

Table 1 Performance comparison of the proposed SemUserProfiling with other approaches

Search technique	Average precision %	Average recall %	Accuracy %	F-measure %	FDR	nDCG
Wisely et al. [22]	84.89	90.32	87.32	87.52	0.151	0.81
PSMU [23]	95.32	98.31	96.71	96.79	0.047	0.84
Fuzzy C-means clustering + LDA	90.12	93.33	92.24	91.69	0.099	0.90
content-based filtering using max entropy classifier	87.32	91.14	88.71	89.19	0.127	0.86
Proposed SemUserProfiling	98.87	99.87	99.02	99.36	0.011	0.97

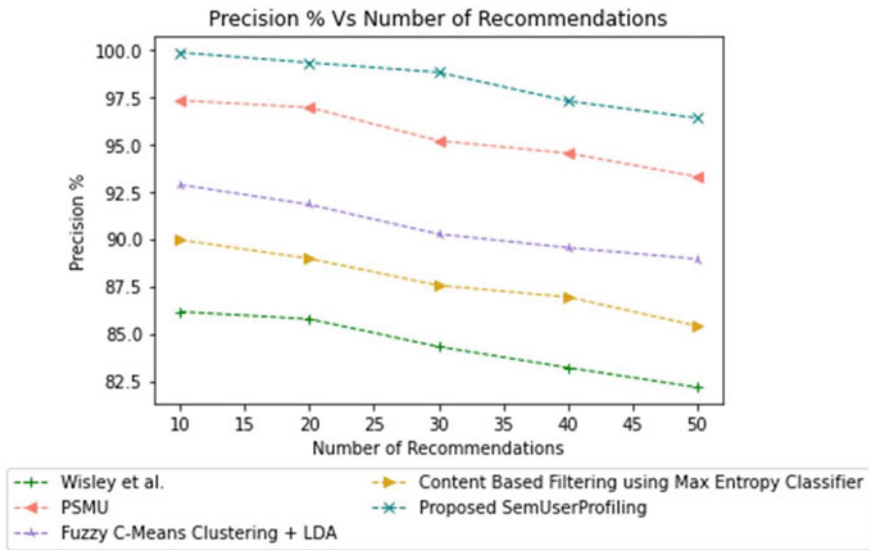


Fig. 2 Precision % versus number of recommendations

each model. The proposed SemUserProfiling approach is better than other baseline approaches due to several reasons. Firstly, it is a classification-based approach. The precision, recall, accuracy, and *F*-measure are high mainly because we are enriching the entities. Entity enrichment happens using news API and Google news feed API along with LOD cloud. Furthermore, the entity classification takes place using XGBoost algorithm and the semantic similarity is computed using entropy and NGD under the frog leap algorithm. Also, the user profile is subjected to topic modeling and entity integration. These are the main reasons for better performance of the proposed approach. Higher nDCG value means higher diversity in results. The reason for the high nDCG value of the proposed approach is due to topic modeling

using STM, entity integration using LOD cloud, and scenario retrieval using Google news feed API and other news APIs.

Content-based filtering using max entropy classifier approach is not as good as the proposed approach because it does not have any entity integration approach, and the relevant entities are not added. Also, there is no scenario retrieval. Fuzzy C-means clustering + LDA used the topic modeling approach but the fuzzy C-means clustering alone is not sufficient to reach such high accuracy as the proposed approach. In the PSMU [23] approach, the author is using the NER technique to cluster keywords which is later used to cluster the users. There is no process to further enrich the entity set, and also there is no procedure for scenario retrieval for latest information from several news APIs. Similarly, the Wisely et al. [22] approach uses the user's browsing history to crawl the content with no entity enrichment or scenario retrieval techniques. Hence, the baseline models are not as good as the proposed approach.

6 Conclusions

In the world where social media plays an important role in information sharing, user profiling is much needed. SemUserProfiling has been proposed where a user profiling model which is a semantically driven approach and is based on entity enrichment and STM. The Twitter dataset is preprocessed, and scenario is retrieved for each term using Google news feed API and several other news APIs. Moreover, the entity enrichment takes place using the LOD cloud and classification of the entity set happens using XGBoost algorithm using categorical domain ontologies. Similarly, user profile Up is preprocessed and subjected to topic modeling using STM, and entity integration is achieved using LOD cloud. Eventually, the semantic similarity is calculated between the enriched user profile terms and the classified enriched entity set using entropy and NGD under frog leap algorithm, to obtain the rearranged user profiles and terms in increasing order of the semantic similarity. The proposed approach achieves the precision of 98.87%, recall of 99.87%, accuracy of 99.02%, F -measure of 99.36%, false discovery rate of 0.011, and nDCG of 0.97. The overall accuracy of the proposed approach is much better than the existing approaches.

References

1. A.P.López-Monroy, M. Montes-y-Gómez, H.J. Escalante, L. Villasenor-Pineda, E. Stamatatos, Discriminative subprofile-specific representations for author profiling in social media. *Knowl. Based Syst.* **89**, 134–147 (2015)
2. J.W. Van Dam, M. Van De Velden, Online profiling and clustering of Facebook users. *Decis. Support Syst.* **70**, 60–72 (2015)
3. J. Chen, Y. Liu, M. Zou, Home location profiling for users in social media. *Inf. Manag.* **53**(1), 135–143 (2016)

4. F. Greco, A. Polli, Emotional text mining: customer profiling in brand management. *Int. J. Inf. Manag.* **51**, 101934 (2020)
5. P. Mishra, M. Del Tredici, H. Yannakoudakis, E. Shutova, in *Author Profiling for Abuse Detection*. Proceedings of the 27th International Conference on Computational Linguistics (2018 August), pp. 1088–1098
6. R. Mishra, P.P. Sinha, R. Sawhney, D. Mahata, P. Mathur, R.R. Shah, in *SNAP-BATNET: Cascading Author Profiling and Social Network Graphs for Suicide Ideation Detection on Social Media*. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop (2019 June), pp. 147–156
7. P.J. Wisniewski, B.P. Knijnenburg, H.R. Lipford, Making privacy personal: profiling social network users to inform privacy education and nudging. *Int. J. Hum Comput Stud.* **98**, 95–108 (2017)
8. D. Kosmajac, V. Keselj, in *Twitter User Profiling: Bot and Gender Identification*. International Conference of the Cross-Language Evaluation Forum for European Languages (Springer, Cham, 2020 September), pp. 141–153
9. M. Singh, B. M. Mehtre, S. Sangeetha, in *User Behavior Profiling Using Ensemble Approach for Insider Threat Detection*. 2019 IEEE 5th International Conference on Identity, Security, and Behavior Analysis (ISBA) (IEEE, 2019 January), pp. 1–8
10. H. Chen, M. Sun, C. Tu, Y. Lin, Z. Liu, in *Neural Sentiment Classification with User and Product Attention*. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (2016 November), pp. 1650–1659
11. R.G. Guimaraes, R.L. Rosa, D. De Gaetano, D.Z. Rodriguez, G. Bressan, Age groups classification in social network using deep learning. *IEEE Access* **5**, 10805–10816 (2017)
12. S. Menini, G. Moretti, M. Corazza, E. Cabrio, S. Tonelli, S. Villata, in *A System to Monitor Cyberbullying Based on Message Classification and Social Network Analysis*. Proceedings of the Third Workshop on Abusive Language Online (2019 August), pp. 105–110
13. G. Deepak, N. Kumar, A. Santhanavijayan, A semantic approach for entity linking by diverse knowledge integration incorporating role-based chunking. *Procedia Comput. Sci.* **167**, 737–746 (2020)
14. G. Deepak, S. Rooban, A. Santhanavijayan, A knowledge centric hybridized approach for crime classification incorporating deep bi-LSTM neural network. *Multimedia Tools Appl.* 1–25 (2021)
15. K. Vishal, G. Deepak, A. Santhanavijayan, in *An Approach for Retrieval of Text Documents by Hybridizing Structural Topic Modeling and Pointwise Mutual Information*. Innovations in Electrical and Electronic Engineering (Springer, Singapore, 2021), pp. 969–977
16. G. Deepak, V. Teja, A. Santhanavijayan, A novel firefly driven scheme for resume parsing and matching based on entity linking paradigm. *J. Discrete Math. Sci. Crypt.* **23**(1), 157–165 (2020)
17. G. Deepak, N. Kumar, G.V.S.Y. Bharadwaj, A. Santhanavijayan, *OntoQuest: An Ontological Strategy for Automatic Question Generation for e-Assessment Using Static and Dynamic Knowledge*. 2019 Fifteenth International Conference on Information Processing (ICINPRO) (IEEE, 2019 December), pp. 1–6
18. G. Deepak, A. Santhanavijayan, OntoBestFit: a best-fit occurrence estimation strategy for RDF driven faceted semantic search. *Comput. Commun.* **160**, 284–298 (2020)
19. M. Arulmozhivarman, G. Deepak, in *OWLW: Ontology Focused User Centric Architecture for Web Service Recommendation Based on LSTM and Whale Optimization*. European, Asian, Middle Eastern, North African Conference on Management & Information Systems (Springer, Cham, 2021 March), pp. 334–344
20. M.E. Roberts, B.M. Stewart, D. Tingley, E.M. Airoldi, in *The Structural Topic Model and Applied Social Science*. Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation, vol. 4 (2013 December), pp. 1–20
21. M. Eusuff, K. Lansey, F. Pasha, Shuffled frog-leaping algorithm: a memetic meta-heuristic for discrete optimization. *Eng. Optim.* **38**(2), 129–154 (2006)

22. W.L. Dennis, A. Erwin, M. Galinium, in *Data Mining Approach for User Profile Generation on Advertisement Serving*. 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE) (IEEE, 2016 October), pp. 1–6
23. G.U. Vasanthakumar, D.R. Shashikumar, L. Suresh, in *Profiling Social Media Users, a Content-Based Data Mining Technique for Twitter Users*. 2019 1st International Conference on Advances in Information Technology (ICAIT) (2019 July), pp. 33–38

Prediction of Stroke Disease Using Different Types of Gradient Boosting Classifiers



Astik Kumar Pradhan, Satyajit Swain, Jitendra Kumar Rout,
and Niranjan Kumar Ray

Abstract Stroke is a critical medical state which occurs when blood supply to a region in the brain is hindered or minimized, depriving the brain tissues from oxygen and nutrients. As a result, the brain cells start to die within minutes. The majority of strokes occur due to an unpredicted disruption in the brain and heart's pathways. Prior recognition of the various stroke warning signs can help minimize the severity of the stroke. In this paper, various machine learning techniques are utilized to identify stroke diseases early using various clinical features. For predicting a stroke, three distinct classifiers, namely eXtreme gradient boost (XGBoost), light gradient boosting machine (LGBM), and CatBoost have been used on existing dataset. Using UC-ROC score, it has been found that the XGBoost classifier outperforms the other two classifiers. The accuracy value for XGBoost classifier has been recorded as 96%, which is highest compared to other two boosting classifier.

Keywords Stroke disease prediction · Machine learning · Boosting · Extreme gradient boost · Light gradient boosting machine and CatBoost

1 Introduction

Health is regarded as a vital aspect of everybody's life, and thus a framework to track diseases and their links are needed. Patient case summaries, emergency medical reports, as well as other manually compiled records contain the majority of disease-related details. To decode the sentences in them, knowledge discovery and machine learning (ML) techniques can be used [1]. Different ML and text analysis approaches are developed and constructed for the extraction of features and classification [2]. A stroke is a potentially lethal disease that happens when blood supply to different regions of the brain is barged in or reduced, and the cells within these areas are

A. K. Pradhan · S. Swain · N. K. Ray
Kalinga Institute of Industrial Technology, Bhubaneswar, Odisha 751024, India

J. K. Rout (✉)
National Institute of Technology, Raipur, Chhattisgarh 492010, India

deprived of nutrients and oxygen and begin to die [3, 4]. To avoid more injury to the harmed portion of the brain as well as other complications in other parts of the body, early diagnosis and careful management are needed. Stroke detection is important, and it must be managed to avoid permanent harm or death. Hypertension, BMI, heart disease, and mean glucose level have been used as factors in this study to predict stroke. Stroke claims the lives of a significant number of people and it is on the rise in emerging countries. By detecting and addressing issues early, machine learning can help patient's well-being. Three distinct boosting algorithms, such as eXtreme gradient boosting (XGBoost), light gradient boosting machine (LGBM), and CatBoost classifier, have been used in this study to predict stroke disease. To assess the performance of each classifier and the ranking of predictions, the AUC-ROC score has been used.

The rest of the paper is laid out as follows. Section 2 discusses about the related works carried out in this field. Section 3 demonstrates the research methodology which includes dataset interpretation, data preprocessing, splitting of the dataset, and ML boosting classifiers. The results have been discussed in Sect. 4. Finally, the paper concludes in Sect. 5.

1.1 Motivation and Objective

According to many neurologists, there is currently no treatment that can fully heal a stroke. The number of individuals who die as a result of a stroke is 10 times higher in developing nations and it is increasing twice as fast in the world. This research aims to see how efficiently boosting machine learning models can be used to predict the stroke using clinical features. The flaws of data preprocessing, feature selection, and prediction of stroke are all addressed by the approach.

2 Literature Review

Several works related to this field have been carried out till date. Bentley et al. [1] aim to see whether ML used in CT images can predict the outcome of stroke thrombolysis using support vector machine (SVM). Cheon et al. [3] used a deep neural network with principal component analysis (PCA) to classify 15,099 stroke patients using medical care usage and health behavior data. A research was made by Colak et al. [4] to foresee the results of stroke through knowledge discovery process techniques, artificial neural networks, and support vector machine models. Kamal et al. [5] have included an overview of recent advances and applications of ML in neuro-imaging, with an emphasis on acute ischemic stroke. Kim et al. [6] have analyzed the efficiency of natural language processing and ML techniques for classifying brain MRI radiology reports onto acute ischemic stroke (AIS) and non-AIS phenotypes. Liu et al. [7] have established a hybrid ML method to detect cerebral

stroke in clinical diagnosis using the physiological data having insufficiency and class variance. Monteiro et al. [8] have used ML methods to analyze the functional results of ischemic stroke patients, three months following admittance. Sirsat et al. [9] have built a stroke prediction framework that uses real-time bio-signals and artificial intelligence to detect stroke. Similar kinds of works were also reported by Govindarajan et al. [10], Heo et al. [11], and Lin et al. [12].

Despite several works in the current literature, it has been found that no work has been done yet comparing XGBoost with other boosting classifiers, which is, thus, carried out in this paper.

3 Methodology

The entire flow of work starting from input data preprocessing to predicting the best classifier in terms of optimization parameters is presented in this section. The work flow diagram for the entire process is shown in Fig. 1.

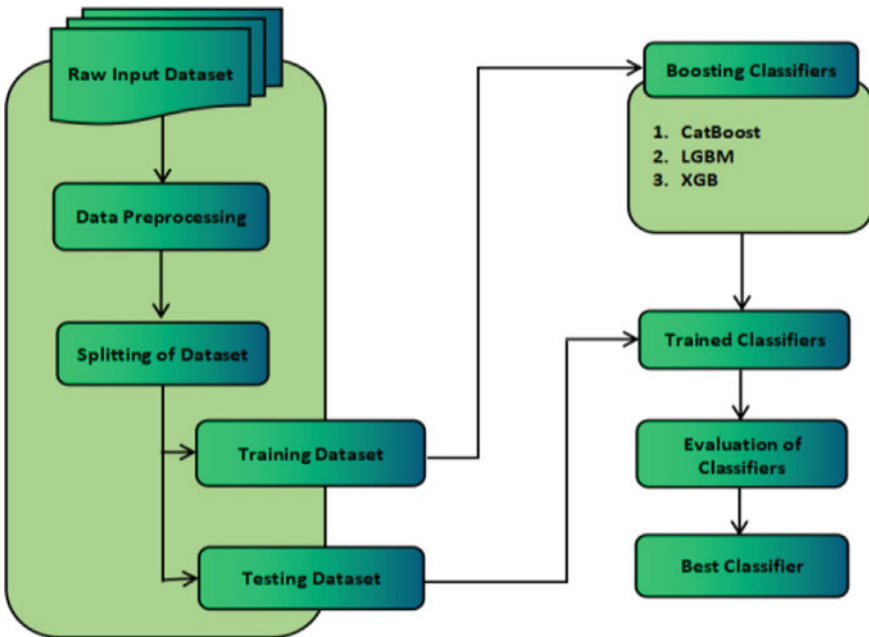


Fig. 1 Workflow diagram for the proposed prototype

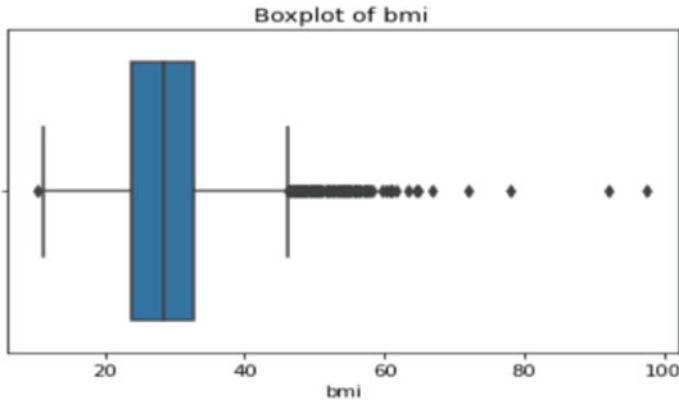


Fig. 2 Boxplot of BMI

3.1 Dataset Interpretation

For the prediction of stroke disease, the input data have been collected from Kaggle repository [13]. The dataset consists of 5110 rows and 10 columns. The attributes are (i) Gender, (ii) Age, (iii) Ever_married, (iv) Work_type, (v) Smoking_status, (vi) Hypertension, (vii) Heart_disease, (viii) Avg_glucose_level, (ix) BMI, (x) Stroke.

All these attributes have been used for the prediction of stroke disease by using different classifiers. But since some attributes in the dataset contain some null values, they have been corrected in the preprocessing phase.

3.2 Data Preprocessing

Missing values were replaced by the mean/median of the other values. Two attributes, such as BMI and avg_glucose_level, contain some outliers as shown in Figs. 2 and 3 were subsequently removed. After that, normalization and label encoding are applied on the categorical data, discovering the entire dataset as a numerical value and obtaining a standardized dataset for further examination. Following this, the dataset has been split into two groups (80%:20%) for training and testing, respectively.

3.3 Boosting Classifiers

Gradient boosting is an effective ML technique that gives state-of-the-art results in a number of real-world applications. It is a method of creating an ensemble predictor in a functional space by performing gradient descent. In this section, the three ML

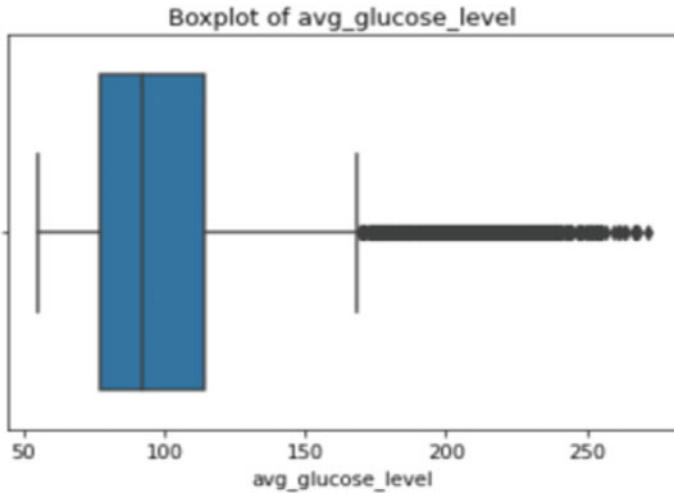


Fig. 3 Boxplot of avg_glucose_level

models used for the experimentation purpose, i.e., CatBoost, LGBM, and XGBoost, are presented.

3.3.1 CatBoot

It is a gradient boosting algorithm centered on decision trees that yields categorical values such as text, audio, and video. It can be used to translate categories into numbers without any specific preprocessing. It removes the requirement for detailed hyper-parameter tuning and decreases the possibility of overfitting, resulting in more generalized models. CatBoost employs a quick encoding process that is alike average encoding. It minimizes overfitting in the following manner: (i) It starts by transforming the set of input observations in an arbitrary manner, producing various arbitrary permutations. (ii) The label values are converted from a floating point or category into an integer. (iii) Equation 1 is used to convert the values of all categorical features to numeric values.

$$\text{target}_{\text{avg}} = \frac{\text{count}_{\text{inclass}} + \text{prior}}{\text{count}_{\text{total}} + 1} \tag{1}$$

where $\text{count}_{\text{inclass}}$ is the count of the label value being matched to one for targets having the present categorical attribute value, prior denotes the initial value for the numerator that is predicted via the early parameters, and $\text{count}_{\text{total}}$ gives the overall count of targets (till the present one) which have a categorical attribute value resembling the present one. Let $p = (p_1, \dots, p_n)$ be the permutation. Then, mathematically, $y_{p,i}$ can be represented using an equation as given in (2).

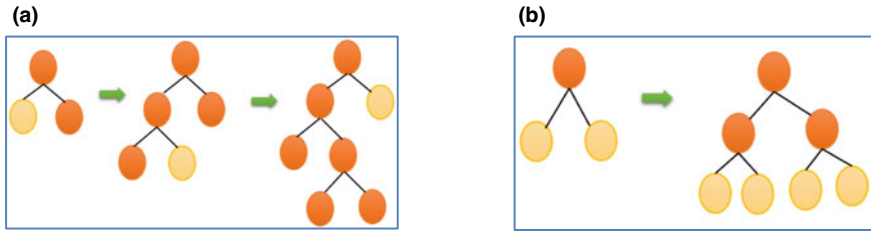


Fig. 4 Leaf-wise tree growth (LGBM), **b** level-wise tree growth

$$y_{p,i} = \frac{\sum_{k=1}^{p-1} [y_{k,i} = y_{p,i}] X_k + z \cdot S}{\sum_{k=1}^{p-1} [y_{k,i} = y_{p,i}] + z} \tag{2}$$

3.3.2 Light GBM

Light GBM is a tree-based gradient boosting framework. Light GBM raises trees vertically, while other algorithms develop the trees horizontally. This means that light GBM grows trees leaf-by-leaf, while other algorithms grow level-by-level. It will develop the leaf with the highest delta loss. Leaf-wise algorithms can reduce more loss than level-wise algorithms when increasing the same leaf. The diagram Fig. 4a, b illustrates how LGBM and some other boosting algorithms are implemented.

3.3.3 XGBoost

The richness of this effective algorithm is its scalability. Unlike CatBoost and LGBM, XGB cannot accommodate categorical features on its own, instead, it supports numerical values in the same way as random forest does. Before supplying categorical data, various encodings such as label encoding, average encoding, or one-hot encoding must be performed. It has the ability to tune complicated models using both $L1$ and $L2$ regularization, averting overfitting. The minimized objective function L (loss function and regularization) at iteration i is given in (3).

$$L^{(i)} = \sum_{j=1}^k y \left(z_j, \bar{z}_j^{(i-1)} + l_i(A_j) \right) + \sigma(l_i) \tag{3}$$

where y is the loss term, z_j is a real value label known from the training dataset, and σ is the regularization term. The above equation can be seen as $f(x + \Delta x)$, where $x = z_j^{(i-1)}$.

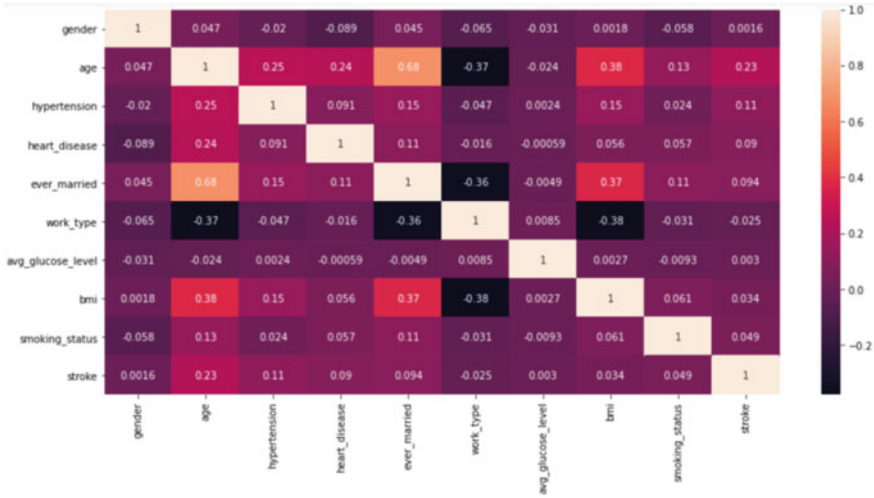


Fig. 5 Correlation matrices between the stroke attribute and the other attributes

4 Results and Discussion

In this section, the results obtained using XGBoost, LGBM, and CatBoost used for the prediction of stroke disease are analyzed and compared in terms of the different performance metrics such as precision, recall, accuracy, and *F1* score.

4.1 Results of Correlation

The relationship between the stroke attribute and the other attributes is shown in Fig. 5. As it can be observed from the chart, no particular metric has a significant impact on stroke. Factors such as heart_disease, gender, BMI, age, hypertension, avg_glucose_level, and smoking_status have a major impact on stroke, while factors like Ever_married and Work_type have the least influence on stroke.

4.2 Evaluation of Performance

In this section, the testing dataset is used to analyze how accurate the classifiers are at classifying the data. The confusion matrices for the three different gradient boosting classifiers for stroke prediction have been shown in Table 1.

In Table 2, precision, recall, and the *F1* score of each classifier have been summarized with accuracy and AUC-ROC score. It has been found that XGBoost classifier

Table 1 Confusion matrices of three different gradient boosting classifiers

Classifiers	Predicted →	Not stroke	Stroke
	Actual ↓		
CatBoost	Not stroke	760	64
	Stroke	35	831
LGBM	Not stroke	767	57
	Stroke	32	834
XGBoost	Not stroke	784	40
	Stroke	26	840

Table 2 Result analysis of three different gradient boosting classifiers

Classifiers	Class labels	Precision	Recall	F1 score	Accuracy (%)	AUC-ROC
CatBoost	Not stroke	0.96	0.92	0.94	94	0.988
	Stroke	0.93	0.96	0.94		
LGBM	Not stroke	0.96	0.93	0.95	95	0.989
	Stroke	0.94	0.96	0.95		
XGBoost	Not stroke	0.95	0.97	0.96	96	0.993
	Stroke	0.97	0.95	0.96		

XGBoost classifier performs better as compared to other classifiers with F1-score of 0.96, accuracy of 96% and AUC-ROC of 0.993. So the result part for this specific case is made bold to highlight it.

performs quite well with an accuracy of 96%. LGBM classifier holds the second position in terms of performance with 95% accuracy value. However, the CatBoost model performs the least well as compared to the other two classifiers with 94% accuracy.

The AUC-ROC curve gives an output measurement for prediction and classification problems. It provides the relation among the true positive rate (TPR) and the false positive rate (FPR). Naturally, the greater the TPR and smaller the FPR for each threshold, the better is the outcome since curve-based classifiers are preferable. The more top-left the curve is the greater the area end as a bi-product, the higher is the AUC-ROC score. The AUC-ROC score of XGBoost classifier comes out to be 0.993, which is higher as compared to the other two classifiers. The AUC-ROC curve of three boosting classifiers has been shown in Fig. 6.

5 Conclusion and Future Scope

In this paper, three distinct classifiers, namely XGBoost, LGBM, and CatBoost, were used to determine a person’s stroke occurrence performance. The XGBoost classifier has been found to be the most accurate, with a 96% accuracy rate followed by LGBM

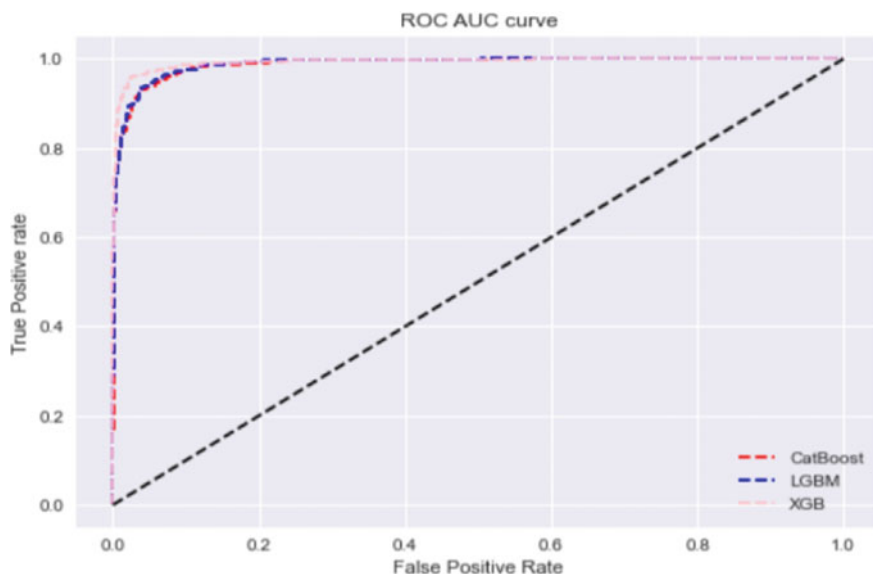


Fig. 6 AUC-ROC curve of three gradient boosting classifiers

classifier with accuracy of 95% whereas CatBoost model does not perform well and it has an accuracy of 94%. The AUC-ROC score for each classifier was determined, and the AUC-ROC score in case of XGBoost classifier was found to be 0.993. This had a higher rating in comparison with the other two classifiers. Although any classifier is required to fix the errors in the predecessors, boosting is sensitive to outliers. As a result, the approach is overly reliant on outliers. Another shortcoming is scaling up the method is nearly impossible. The link between these chronic diseases and the risk of having a stroke in a human was explored. Thus to improve the performance measures in future, the DL-based imagery like neural CT scans and MRI can be combined along with an established framework.

References

1. P. Bentley, J. Ganesalingam, A.L.C. Jones, K. Mahady, S. Epton, P. Rinne, P. Sharma, O. Halse, A. Mehta, D. Rueckert, Prediction of stroke thrombolysis outcome using CT brain machine learning. *NeuroImage Clin.* **4**, 635–640 (2014)
2. B.K. Mengiste, H.K. Tripathy, J.K. Rout, in *Analysis and Prediction of Cardiovascular Disease Using Machine Learning Techniques*. Advances in Systems, Control and Automations: Select Proceedings of ETAEERE 2020 (Springer Singapore, 2021), pp. 133–141
3. S. Cheon, J. Kim, J. Lim, The use of deep learning to predict stroke patient mortality. *Int. J. Environ. Res. Public Health* **16**(11), 1876 (2019)
4. C. Colak, E. Karaman, M.G. Turtay, Application of knowledge discovery process on the prediction of stroke. *Comput. Methods Programs Biomed.* **119**(3), 181–185 (2015)

5. H. Kamal, V. Lopez, S.A. Sheth, Machine learning in acute ischemic stroke neuroimaging. *Front. Neurol.* **9**, 945 (2018)
6. C. Kim, V. Zhu, J. Obeid, L. Lenert, Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke. *PloS One* **14**(2), e0212778 (2019)
7. T. Liu, W. Fan, C. Wu, A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. *Artif. Intell. Med.* **101**, 101723 (2019)
8. M. Monteiro, A.C. Fonseca, A.T. Freitas, T.P. Melo, A.P. Francisco, J.M. Ferro, A.L. Oliveira, Using machine learning to improve the prediction of functional outcome in ischemic stroke patients. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **15**(6), 1953–1959 (2018)
9. M.S. Sirsat, E. Fermé, J. Câmara, Machine learning for brain stroke: a review. *J. Stroke Cerebrovasc. Dis.* **29**(10), 105162 (2020)
10. P. Govindarajan, R.K. Soundarapandian, A.H. Gandomi, R. Patan, P. Jayaraman, R. Manikandan, Classification of stroke disease using machine learning algorithms. *Neural Comput. Appl.* **32**(3), 817–828 (2020)
11. J.N. Heo, J.G. Yoon, H. Park, Y.D. Kim, H.S. Nam, J.H. Heo, Machine learning–based model for prediction of outcomes in acute stroke. *Stroke* **50**(5), 1263–1265 (2019)
12. C.H. Lin, K.C. Hsu, K.R. Johnson, Y.C. Fann, C.H. Tsai, Y. Sun, L.-M. Lien et al., Evaluation of machine learning methods to stroke outcome prediction using a nationwide disease registry. *Comput. Methods Programs Biomed.* **190**, 105381 (2020)
13. Stroke Prediction Dataset, <https://www.kaggle.com/fedesoriano/stroke-prediction-dataset>. Accessed 10 Dec 2020

Bi-objective Task Scheduling in Cloud Data Center Using Whale Optimization Algorithm



Srichandan Sobhanayak, Isaac Kennedy Alexandre Mendes,
and Kavita Jaiswal

Abstract Workflow scheduling in clouds refers to mapping workflow tasks to the cloud resources to optimize some objective function. Workflow scheduling is a crucial component behind the process for optimal workflow enactment. It is a well-known NP-hard problem and is more challenging in the heterogeneous computing environment. Cloud environments confront several issues, including energy consumption, implementation time, emissions of heat and CO₂ and running costs. The increasing complexity of the workflow applications forces researchers to explore hybrid approaches to solve the workflow scheduling problem. Efficient and effective cloud workflow planning is one of the most important approaches to address the above difficulties and make optimal use of resources. This study suggests energy awareness, based on the methodology whale optimization algorithm (WOA). Our objective is to decrease the energy consumption and maximize the throughput of computational workflows which impose a considerable loss on the quality of service guarantee (QoS). The proposed method is compared with other standard state-of-the-art techniques to analyze its performance.

Keywords Whale optimization algorithm · Cloud computing · Energy · Throughput · Cost · Physical machine

S. Sobhanayak (✉) · I. K. A. Mendes
Department of Computer Science, IIIT Bhubaneswar, Bhubaneswar, Odisha 751003, India
e-mail: srichandan@iiit-bh.ac.in

I. K. A. Mendes
e-mail: a119002@iiit-bh.ac.in

K. Jaiswal
Department of Computer Science, NIT Raipur, Raipur, Chhattisgarh 492001, India
e-mail: a116010@iiit-bh.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022
P. Verma et al. (eds.), *Advances in Data Computing, Communication and Security*,
Lecture Notes on Data Engineering and Communications Technologies 106,
https://doi.org/10.1007/978-981-16-8403-6_31

347

1 Introduction

The term cloud computing is derived from distributed computing, parallel computing, and grid computing. In cloud computing, resources such as storage, memory, processors, and application are seen as services [1]. Nowadays, hundreds or even thousands of tasks have been created for corporate and scientific applications. The workflow model has a commonly utilized usage of guided acyclic graphs (DAG) [2] for representing such applications. Nodes comprise of tasks and edges reflect the dependencies between them in a workflow application. Large workflows may be submitted within an acceptable period to distributed computing environments. Traditional distributed settings such as clusters and grids cost experienced people to maintain them functioning and require them. Cloud computing is an alternative for running workflows since it is easy to set up and can provide and release resources with little administrative efforts. Unfortunately, high energy consumption and resource utilization have become a critical issue in clouds. The energy consumed in data centers accounts for 1.4% of the world's total power usage with a growth rate of 12% per year. As a result, reducing the energy consumption of data centers is increasingly attractive, while maximizing throughput is an important issue as it reduces monetary costs and protects our natural environment.

There has already been a lot of effort in recent years on optimizing the implementation of large-scale scientific workflow applications. It is an issue known as [3] for NP-complete. Various applications need high dependability, high throughput and minimal completion time. Others, however, are subject to a specified monetary budget or energy consumption restriction which may not always be acceptable at the greatest possible level of QoS. There are several research on workflow scheduling challenges in heterogeneous contexts. Unlike the mainly makespan minimization cluster and grid environment, workflow scheduling in clouds is also focused on criteria like as cost, energy consumption, fault toleration, security, etc. [4]. The study will design a cloud-based she algorithm that optimizes energy consumption and computational workflows throughput, in particular, to reduce energy consumption and maximize throughput using a metaheuristic technique called the whale optimization algorithm (WOA). WOA, on the other hand, is based on whales, specifically humpback whales that use a hunting method called bubble net to trap the prey and hunt them. An individual humpback whale in the search space is already a competitor for a solution in a so-called optimization problem, which may be represented as a search agent in the WOA algorithm. WOA uses these search agents to find out the global solution. The WOA can have the upper hand in performance over other task scheduling and optimization algorithms and solve real-world problems and various cloud computing problems. So we aim to use this algorithm to optimize energy consumption and throughput.

For the reasons outlined above, this paper addresses scheduling scientific workflows in cloud environments employing WOA. The main contribution of our paper is as follows:

Table 1 Literature survey

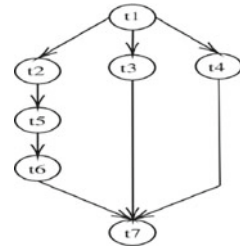
Paper reference	Task scheduling algorithm	Sub type	Application type	Objective function	Simulator used
[5]	Particle Swarm Optimization algorithm	Fuzzy TOPSIS	Heterogeneous workload	To achieve eminent resource utilization, minimize migration cost	CloudSim
[6]	Genetic algorithm	Harmony algorithm	Heterogeneous workload	To reduce makespan and computing energy	Cloudsim
[7]	Genetic algorithm	Whale optimization algorithm	Heterogeneous workload	To efficiently scheduling tasks in a given cloud	Map Reduce
[8]	Particle Swarm Optimization algorithm	MCT algorithm and LJFP algorithm	Heterogeneous workload	To minimize makespan, total execution time, degree of imbalance, and energy consumption	MATLAB
[9]	Load-balancing algorithm	Threshold algorithm	Heterogeneous workload	To exaggerate task scheduling and reduce makespan	GCC compiler
[10]	Hungarian algorithm	Pair based algorithm	Heterogeneous workload	The overall layover time is minimized	MATLAB

- Design a novel fitness function considering energy and throughput.
- Proposes an efficient WOA to optimize the energy and throughput.
- The algorithm's superior performance has been shown via rigorous simulation contrasted with heuristical/meta-heuristic techniques.

Rest of the paper is organized as follow Sect. 2 Literature Survey, Sect. 3 models and description that describes the cloud application and models, Sect. 4 problem statement definition, Sect. 5 we talk about the Proposed Scheduling Technique, Sect. 6 Simulation and Results, and Sect. 7 is conclusion and future work.

2 Literature Survey

The literature survey that we have done is summarized in Table 1. Looking into the literature survey table, no one has considered energy and throughput as their optimal parameter for whale optimization; this motivated us to choose these objectives to optimize.

Fig. 1 Scientific workflow

3 Models and Descriptions

In this section, we provide different models used in our proposal.

3.1 Cloud Model

It is regarded to be a server if it has m PMs, which are represented by the string $PM = pm1, pm2, pm3, \dots, pm_m$. It is possible to estimate the computing power of each PM in millions of instructions per second (MIPS). It is possible to use physical machines. All PMs are entirely interconnected and may be housed in one or more cloud data centers, depending on their configuration. This is the amount of time required to transfer data from the t_i task to the t_j task, which is represented as overhead communication time. The link between the output data size of a t_i task and the bandwidth between the PMs is worth noting; this is referred to as overhead time. When both t_i and t_j are operating on the same PM, the overhead communication is deemed to be zero for the sake of this calculation. To make things easier, we'll ignore transfer time and instead focus on computer-intensive workflows and fully interconnected heterogeneous networks with a high bandwidth that are much shorter in duration than execution time. One PM is responsible for carrying out one or more tasks in accordance with the algorithm of scheduling that has been calculated.

3.2 Task Model

A direct acyclic graph (DAG) [11], $G = (T, E)$, as shown in Fig. 1, where T is the set of tasks and E is a set of edges. The workflow application is displayed using a direct acyclic graph (DAG). The $t_i \rightarrow t_j$ edge shows the previous t_i relation to the successor t_j . Therefore, the t_j task won't commence unless the t_i task is complete. Each one of the task t_i is designated in million instructions belonging to its computational load (MI). The label $t_i \rightarrow t_j$ on each edge also identifies the output data produced by t_i . This data is necessary to start doing the t_j task. The task without a predecessor is

called the entry task (t_{entry}) and the task without a successor is called the exit task (t_{exit}). If more than one entry workflow task exists, a new faux entry task that has null computational burden and no output data is produced. Once this is done, all of the entry tasks are connected to the pseudo task. Likewise, if necessary, a pseudo-exit task might be constructed.

3.3 Throughput

Throughput $TP(S)$ is the number of data outputs created by a mapping scheme S in the workflow per time unit at the conclusion of the task. The throughput is defined by the longest or sometimes called “bottleneck” task in the workflow and determined by identifying the longest or slowest task in the execution of workflow using the following equation:

$$TP(S) = \frac{1}{\max_{i \in (1, n)} ECT(i, u)} \quad (1)$$

where $ECT(i, u)$ represents the time it takes for task T_i to be completed on a physical machine M_u , and $ECT(i, u)$ represents the time it takes for task T_i to be completed on a physical machine M_u . The two values are compared as $ECT(i, u) = \frac{w_i}{f_{r_u}}$, where w_i represents the computational requirements of task T_i .

3.4 Energy Model

Energy consumption of the PM in the cloud data center calculation depends on the CPU utilization of the PM (server). The following power model proposed the overall CPU utilization of the servers in data center Tian et al. [12].

$$P_o = P_{\min} + (P_{\max} - P_{\min})U \quad (2)$$

where P_o represents the power consumption of a PM, P_{\max} represents the maximum power used while the server is completely used, and P_{\min} represents the power consumption when the server is completely idle meaning not running; The CPU utilization is denoted by the letter U . In a real-world context, the utilization of the CPU may alter over time as a result of the fluctuation of the workload. As a result, the CPU utilization is a function of time and is denoted by the symbol $U_i(t)$ in mathematical notation. To put it another way, the total energy consumption (E_i) of a PM may indeed be expressed as the integral of its power consumption function over the time span $[t_0, t_1]$:

$$E_i = \int_{t_0}^{t_1} P(U_i(t_i)) \quad (3)$$

4 Problem Statement

The problem statement can be defined as follows: “Lets we have m PM and n scientific workflow. We are working on a huge scientific project, work flow apps with complex inter task relationships and a diverse cloud computing environment, we would like to devise a map - based or scheduling approach which allocates every other task mostly in the work flow application to a located key PM in order to optimize energy consumption, and throughput.”

The task scheduling problem is always NP-Complete problem [12]. For simplicity reasons, we mainly focus on CPU-intensive apps and only look at CPU-related energy usage. The maximum CPU capacity provided by a single PM is specified by the capacity parameter $g \geq 1$. Each task t_i requires a capacity C_i , that is a natural number ranging from 1 to g . The workflow scheduling problem can be mathematically defined as follows:

$$\text{Minimize } \eta = \alpha_1 E + \alpha_2 \frac{1}{\text{TP}(S)} \quad (4)$$

Subject to:

$$\text{i. } \sum_{j=1}^n B_{i,j} = 1, \quad i = 1, \dots, m; \quad (5)$$

$$\text{ii. } \sum_{j=1}^k V_{i,j} = 1, \quad i = 1, \dots, n; \quad (6)$$

$$\text{iii. } \alpha_1 + \alpha_2 = 1 \quad (7)$$

The constraint (i) states that any task in the workflow can be allocated to only one PM, and the constraint (ii) states that the resource requirement of the task should not exceed the available resource in a PM. Constraint (iii) limits the sum of α_1 , and α_2 , to 1, which is weighting factor that balances energy and throughput.

4.1 Encoding of the Solution

The aim of the task scheduling is to allocate all the scientific tasks to the physical machines while considering the throughput and energy consumption. Assume that, an array of m tasks and n PM are there such that $m < n$. The solution is represented as an array such that the array is the number of the tasks, here it is m . The values inside the array is randomly assigned between 1 and n . The allocation of the values determines the assignment of task to a PM. Lets take an example: if the value inside the array is 5 and corresponding index of the array is 7, this signifies the assignment of task 7 to PM 5. Similarly, all the tasks are allocated to the PM following precedence constraint.

5 Proposed Scheduling Technique

The metaheuristic WOA is designed based on the behavior of humpback whales to optimally allocate scientific workflow to physical machines present in the cloud data center. The WOA begins with a group of random solutions that represents a whale. Considering the initial solution best, it explores all solution search space to find out the global best solution. Fitness function is represented as follows:

$$\eta = \alpha_1 E_i + \alpha_2 \frac{1}{TP(S)} \tag{8}$$

Initialization: During this, the search agent population is initialized, and k random solution is selected, and the fitness function 8 is used to evaluate its fitness. The initial population is defined as, W_j ($j = 1, 2, \dots, k$) and the best search agent is represented as W^* . Then the algorithm passes through following three steps:

1. *Encircling the prey:* When humpback whales surround the target early, they cannot determine the best position in the search space. The solution that is currently being considered the best solution can be defined as the prey sought by the WOA. The whale nearest to the target prey will be chosen as the primary search agent, enabling the other whales to progress toward the prey that is being targeted while moderately updating their positions or locations. These behaviors are displayed in two ways:

$$\vec{D}\vec{V} = \left| \vec{O} \times \vec{LOS}^*(ci) - \vec{P}\vec{V}(ci) \right| \tag{9}$$

and

$$\vec{P}\vec{V}(ci + 1) = \vec{LOS}^*(ci) - \vec{E} \times \vec{D}\vec{V} \tag{10}$$

The distance vector between two points like the search agent to the target prey is denoted $\vec{D}\vec{V}$, the current number of iterations is represented ci , the local optimum

solution is represented as \overrightarrow{LOS}^* , and the position vector is denoted \overrightarrow{PV} . The coefficient vectors are designated by the \overrightarrow{O} and \overrightarrow{E} , and they are calculated as follows:

$$\overrightarrow{O} = 2 \times \overrightarrow{rn} \tag{11}$$

and

$$\overrightarrow{E} = 2 \overrightarrow{d} \times \overrightarrow{rn} - \overrightarrow{d} \tag{12}$$

where rn denotes a random number that ranges from 0 to 1 and d denotes a decremented linear value that ranges from two to zero [0,2] that emanate from the iteration numbers ci over the maximum iteration number ni_{max} that is described as:

$$\overrightarrow{d} = 2 - \frac{2 \overrightarrow{ci}}{\overrightarrow{ci}_{max}} \tag{13}$$

2. *Bubble net attack*: This supports the idea of shrinking encircling and spiral position updating; we can define these as:

- (a) *Shrinking Encircling*: Shrinking surroundings mark the exploitation phase. We may see from Eq. (10) that the whale are going to shrink their circle when $|E| < 1|$. Implying that the others whales will circle the prey to reach the whale that is in the position that is considered best at the time. The whale will take greater steps if the $|E|$ value is bigger and reciprocally.
- (b) *Spiral position updating*: Every single whale will compute his or her distance from the current first whale that is optimal; after that, only then will he or she swim in a spiral so-called shaped pattern.

When the whale updates his/her position, this process can be represented as:

$$\overrightarrow{PV}(ci + 1) = \overrightarrow{DV}' \times e^{yx} \times \cos(2\pi y) + \overrightarrow{LOS}^*(ci) \tag{14}$$

where $\overrightarrow{DV}' = \left| \overrightarrow{O} \times \overrightarrow{LOS}^*(ci) - \overrightarrow{PV}(ci) \right|$ is a vector that shows the distance between the individual whale and the best whale that is currently best found, x is a constant, and y is a random number with the value ranging from minus one to one $[-1, 1]$. To be able to copy these two behaviors exactly, it is presumed that the whale as 0.5 possibility of updating is location-based around the idea of the contraction path and the spiral path that is shown as:

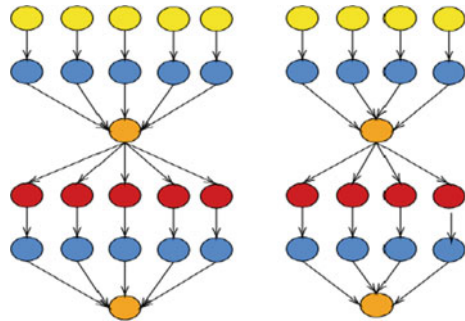
$$\overrightarrow{PV}(ci + 1) = \begin{cases} \overrightarrow{LOS}^*(ci) - \overrightarrow{E} \times \overrightarrow{DV}' & z < 0.5 \\ \overrightarrow{DV}' \times e^{yx} \times \cos(2\pi y) + \overrightarrow{LOS}^*(ci) & z > 0.5 \end{cases} \tag{15}$$

where z denotes a number that is generated randomly ranging from 0 to 1.

Table 2 WOA simulation parameters

Algorithm	Parameter	Value	Description
GA	ρ	0.7	Probability of applying crossover
	p	0.05	Probability of applying mutation
PSO	w	0.9	Initial weight
	ac_1	1.8	Acceleration constant
	ac_2	1.8	Acceleration constant
WOA	b	1	Spiral searching path parameter
	P_{max}	51	The largest population
	P_{min}	10	Initial population
	α	0.25	Individual generate parameter
	γ	20	Nonlinear factor

Fig. 2 CyberShake



3. *Search for prey*: This is the stage of exploration. To ensure that we can come close to a globally optimum solution. When $|E| > 1$, the search agents are distanced from one another, in this case, we replaced the current optimum search agent's position with a search agent chosen at random, as follows:

$$\vec{D}\vec{V} = \left| \vec{O} \times \vec{P}\vec{V}_{rand} - \vec{P}\vec{V} \right| \tag{16}$$

$$\vec{P}\vec{V}(ci + 1) = \vec{P}\vec{V}_{rand} - E \times \left| \vec{O} \times \vec{P}\vec{V}_{rand} - \vec{P}\vec{V}(ci) \right| \tag{17}$$

where $\vec{P}\vec{V}_{rand}$ is a randomly selected position vector by the search agent.

4. *Termination*: If the search agents go outside of the search region, then W^* is updated and the process continues until termination criteria achieved. At end of this process we obtain our optimal solution and fitness function.

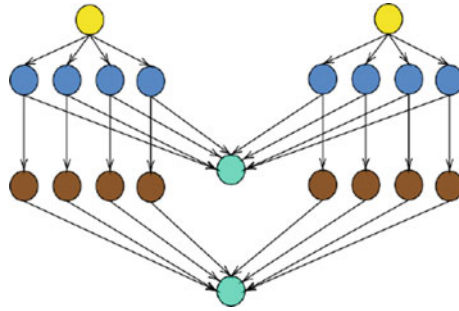


Fig. 3 LIGO

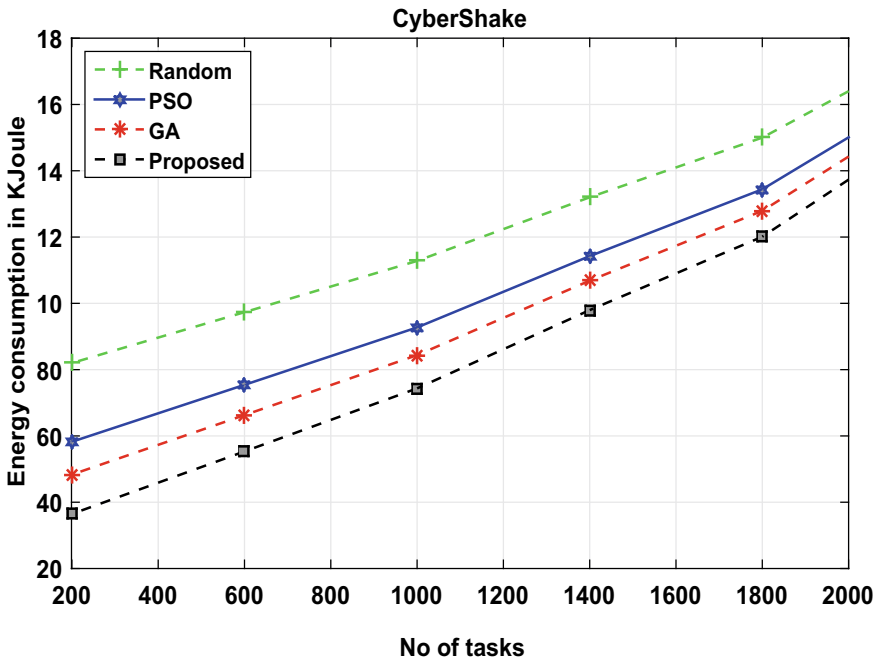


Fig. 4 Energy consumption versus no. of tasks using CyberShake

6 Simulation and Results

The proposed algorithm is implemented using a personal computer with an Intel Core i5 9th generation processor and 8GB of RAM running Windows 10 to test the effectiveness of the suggested approach. The proposed method is implemented in MATLAB R2021a, and its effectiveness is measured in terms of energy and throughput. The performance of the proposed WOA is compared with other state-of-the-art metaheuristic algorithms, such as the random algorithm (Random), particle swarm

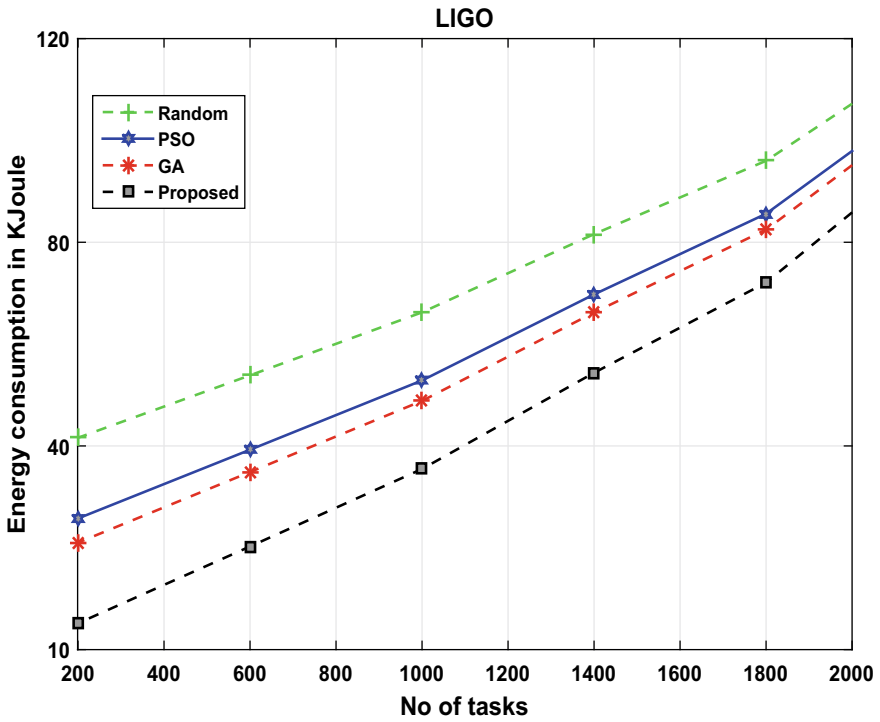


Fig. 5 Energy consumption versus no. of tasks using in LIGO

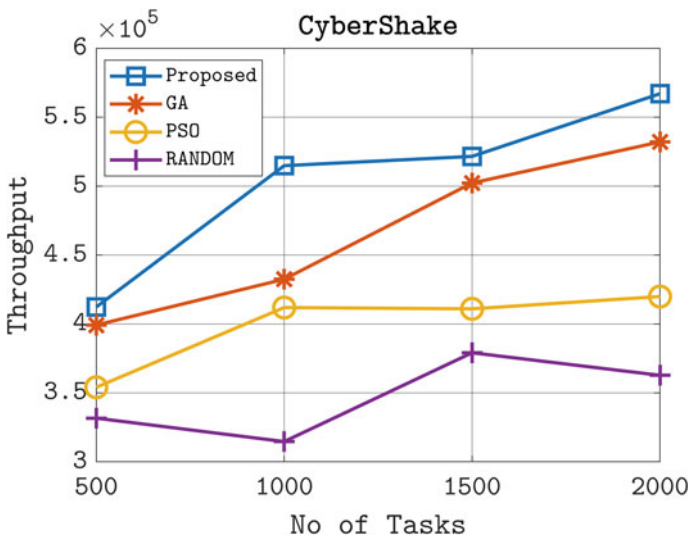


Fig. 6 Throughput versus no. of tasks using CyberShake

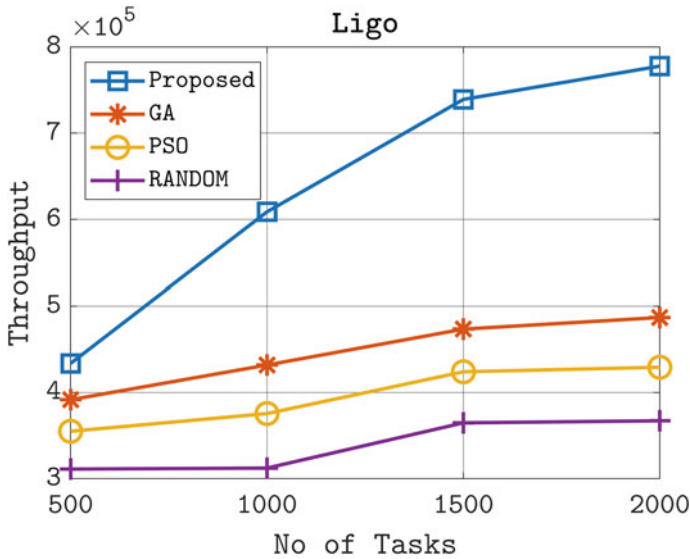


Fig. 7 Throughput versus no. of tasks using LIGO

optimization algorithm (PSO) [13], and genetic algorithm (GA) [14]. The simulation parameter used to evaluate the performance is depicted in Table 2. Finally, we use task sizes from 100 to 2000 and to analyze the performance. The detailed analysis is depicted in Figs. 4, 5, 6, and 7. We have considered two different scientific workflow as defined in the Laser Interferometer Gravitational Wave Observatory (LIGO) that tries to find gravitational waves that are created by a variate of events in the universe as mentioned on the so-called Einstein's theory of general relativity and LIGO is computationally intensive in nature, and CyberShake that is used to give category of tasks in a workflow environment. CyberShake is an input and output and network intensive in nature depicted in Figs. 2 and 3. From the simulation results, we found that the proposed method performs better than GA, PSO, and random method due to its high convergence rate, novel fitness function, and better search to locate the solution.

7 Conclusion and Future Work

On the basis of the bi-objective model and the whale optimization algorithm (WOA), this study proposes the scientific workflow scheduling algorithm for scheduling scientific tasks to the physical machines of the cloud data center. In this paper, we will discuss how to minimize the energy consumption of a cloud data center while simultaneously maximizing its throughput. We devise a unique fitness function that takes a weighted sum approach to optimization. The whale optimization algorithm is then

introduced, which is used to schedule the scientific workflows LIGO and Cyber-Shake to the physical machines in the most efficient manner. The whale optimization algorithm starts with the assumption that the current solution is the best and then searches for the optimal solution using the best search agent available. The suggested algorithm's performance was compared to current methods, such as GA, PSO, and the random technique, for the assessment metrics of energy and throughput. The proposed strategy, we conclude, optimally schedules the scientific workflow to the physical machines while costing the least amount of energy and allowing for the greatest amount of throughput from the experimental results output. In future, we plan to incorporate data-intensive workflow optimization because scientific applications are becoming increasingly data-intensive, and the scientific community is attempting to address the issues posed by big data.

References

1. F.E. Farkar, A.A.P. Kazem, Bi-objective task scheduling in cloud computing using chaotic bat algorithm. *Int. J. Adv. Comput. Sci. Appl. (IJACSA)* **8**(10) (2017)
2. M. Sonntag, D. Karastoyanova, E. Deelman, Bridging the gap between business and scientific workflows: humans in the loop of scientific workflows, in *2010 IEEE Sixth International Conference on e-Science*, Dec 2010, pp. 206–213
3. J.D. Ullman, NP-complete scheduling problems. *J. Comput. Syst. Sci.* **10**(3), 384–393 (1975) [Online]. Available at: <https://www.sciencedirect.com/science/article/pii/S0022000075800080>
4. M. Masdari, S. ValiKardan, Z. Shahi, S.I. Azar, Towards workflow scheduling in cloud computing: a comprehensive analysis. *J. Netw. Comput. Appl.* **66**, 64–82 (2016) [Online]. Available at: <https://www.sciencedirect.com/science/article/pii/S108480451600045X>
5. J. Kumar Samriya, N. Kumar, An optimal SLA based task scheduling aid of hybrid fuzzy Topsis-PSO algorithm in cloud environment. *Mater. Today: Proc.* (2020) [Online]. Available at: <https://www.sciencedirect.com/science/article/pii/S2214785320376495>
6. M. Sharma, R. Garg, Higa: harmony-inspired genetic algorithm for rack-aware energy-efficient task scheduling in cloud data centers. *Eng. Sci. Technol. Int. J.* **23**(1), 211–224 (2020) [Online]. Available at: <https://www.sciencedirect.com/science/article/pii/S2215098618312023>
7. M. Sanaj, P. Joe Prathap, An efficient approach to the map-reduce framework and genetic algorithm based whale optimization algorithm for task scheduling in cloud computing environment. *Mater. Today: Proc.* **37**, 3199–3208 (2021). *International Conference on Newer Trends and Innovation in Mechanical Engineering: Materials Science* [Online]. Available at: <https://www.sciencedirect.com/science/article/pii/S2214785320367535>
8. S.A. Alsaaidy, A.D. Abbood, M.A. Sahib, Heuristic initialization of PSO task scheduling algorithm in cloud computing. *J. King Saud Univ. Comput. Inf. Sci.* (2020) [Online]. Available at: <https://www.sciencedirect.com/science/article/pii/S1319157820305279>
9. M. Lavanya, B. Shanthi, S. Saravanan, Multi objective task scheduling algorithm based on SLA and processing time suitable for cloud environment. *Comput. Commun.* **151**, 183–195 (2020) [Online]. Available at: <https://www.sciencedirect.com/science/article/pii/S014036641930492X>
10. S.K. Panda, S.S. Nanda, S.K. Bhoi, A pair-based task scheduling algorithm for cloud computing environment. *J. King Saud Univ. Comput. Inf. Sci.* (2018) [Online]. Available at: <https://www.sciencedirect.com/science/article/pii/S1319157818302970>

11. Y. Xu, K. Li, L. He, T.K. Truong, A DAG scheduling scheme on heterogeneous computing systems using double molecular structure-based chemical reaction optimization. *J. Parallel Distrib. Comput.* **73**(9), 1306–1322 (2013)
12. W. Tian, M. He, W. Guo, W. Huang, X. Shi, M. Shang, A.N. Toosi, R. Buyya, On minimizing total energy consumption in the scheduling of virtual machine reservations. *J. Netw. Comput. Appl.* **113**, 64–74 (2018)
13. S. Liu, Y. Yin, Task scheduling in cloud computing based on improved discrete particle swarm optimization, in *2019 2nd International Conference on Information Systems and Computer Aided Education (ICISCAE)* (IEEE, 2019), pp. 594–597
14. F. Yiqiu, X. Xia, G. Junwei, Cloud computing task scheduling algorithm based on improved genetic algorithm, in *IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)* (IEEE, 2019), pp. 852–856

NDVI-Based Raster Band Composition for Classification of Vegetation Health



Rishwari Ranjan, Ankit Sahai Saxena, and Hemlata Goyal

Abstract The arid part in the Indian subcontinent displays a significant variance in vegetation and climate. The varying climate, lack of perennial rivers, rainfall, and harsh weather conditions only allow sparse vegetation to grow, since proper mapping of such areas is essential for the livelihood of people. In this work, we classify normalized difference vegetation index (NDVI) values to get vegetation health, for this, the study area is taken as Jodhpur district in Rajasthan. With the use of Google Earth API and Python script is designed for extraction of the study area, with the specification of the time frame and download the .tiff images in a more convenient way. USGS Earth Explorer and Landsat 8 imagery raster band 4,5 composite dataset is used to compute pre-monsoon and post-monsoon months to extract the NDVI values in .tiff image format. Analysis of the results concluded that a significant spike in the less dense vegetation category was due to rainfall in the post-monsoon months according to NDVI extraction values.

Keywords NDVI · Band · Landsat · Google Earth · Vegetation · Rainfall

1 Introduction

Fauna available in arid zones is useful for the people and the livestock in that area. Vegetation in this area is scarce due to low rainfall and large areas. Owing to the constantly changing conditions, mapping of vegetation in this area becomes a difficult task for the authorities [1]. Rajasthan state is mainly an arid state, and the climate is generally marked by low rainfall with limited rainy days and sparse distribution. Rain is the dominant factor that affected directly to the vegetation of any region [2] having 100 to 500 ml of rainfall in this arid state [3]. The vegetation is mainly thorny in these areas.

With advancement in satellites and remote sensing technology, reflectance data are increasingly being used in vegetation [4, 5]. Field base interpretation and remote

R. Ranjan · A. S. Saxena · H. Goyal (✉)

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

sensing are used to study the vegetation in these areas. Jodhpur is selected as study area for this research, since rainfall is erratic and mainly occurs from July to September [1, 6].

As the crop production rate depends on the geography of the region, for example, weather conditions, temperature, cloud cover, moisture, soil type, soil composition, harvesting methods, etc., different combinations characteristics can be used to predict the vegetation cover in an area.

1.1 Normalized Difference Vegetation Index (NDVI)

The NDVI helps to visualize an image with greenness. Using the contrast between two bands in multispectral raster dataset and the pigment chlorophyll absorption in the red band and the high reflectivity that plant materials give in the near-infrared (NIR) band. The NDVI uses the separation of the two wavelengths from the multi-raster database, the pigment chlorophyll gets inclusion with a red band and the building materials in near-infrared band (NIR). NDVI is specially utilized to monitor droughts and presage agricultural production. NDVI is chosen to monitor vegetation worldwide [7] and can be calculated by the following formula.

$$\text{NDVI} = ((\text{Infrared} - \text{Red Bands}) / (\text{InfraRed} + \text{Red Bands})) \quad (1)$$

NDVI values range from -1.0 to 1.0 , usually signifying green, while the worst values produced in cloud cover, ice cover or snowfall, water, and near zero values are usually produced on rock and barren ground. The shrub and grass have moderate ($0.2-0.3$) and high ($0.6-0.8$) temperatures indicating tropical and subtropical rainforests.

2 Literature Review

As we were working with an arid region like Jodhpur in “Modis-derived NDVI-based time series analysis of vegetation in the Jodhpur area” [8] by Yadav et al. talked about how Modis NDVI is best for quick vegetation assessment in arid regions. Also, changes can be observed when analyzing post-monsoon and pre-monsoon vegetation changes correlating with the rainfall pattern, especially in arid regions. Since we decided to use Landsat 8, “compare NDVI extracted from Landsat 8 imagery with that from Landsat 7 imagery” [9] by Dandan Xu et al. elaborate upon how values of NDVI calculated with Landsat 8 are higher when we deal with smaller regions as compared to Landsat 7 as the bands in Landsat 8 are narrower, however, when dealing with larger areas like forests, we see these differences in value shrink. “Exploring Landsat 8” [10] by Tri Dev Acharya et al. talks about the various bands and their combination in Landsat 8 and their various new technological and new science opportunities. It explores that the new NIR bands in Landsat 8 are split from Landsat 7, which results

in less absorption by atmospheric water. In “characteristics of Landsat 8 OLI-derived NDVI by comparison with multiple satellite sensors and in-situ observations,” [11] Yinghai Ke et al. explored the various improvement Landsat 8 NDVI offers better agreement as compared to Landsat 7 NDVI on vegetated than non-vegetated lands.

Google Earth Engines are a planetary-scale ubiquitously available platform which is easy to use and provide excellent speed [12]. Providing interactive data exploration its API helped us build our visual NDVI Jupyter Notebook. The API combines the power of massive data and power of computation from cloud computing which works under the covers as an intrinsically parallel image processing system [13].

We used SAGA as our image processing tool as in “SAGA GIS for computing multispectral vegetation indices by Landsat TM for mapping vegetation greenness” [14] was mentioned SAGA is over all quite efficient in vegetation mapping of desired regions.

In “feature extraction using normalized difference vegetation index (NDVI): A case study of Jabalpur city” [15] by Ashish Kumar Bhandari et al. remote sensed data was used to calculate the NDVI values and different values were used to classify the areas. This concluded that NDVI is highly beneficial in detecting surface features for useful mapping by authorities. In “sugarcane crop yield forecasting model using supervised machine learning” [16] by Ramesh Medar et al., long-term time series (LTTS), NDVI, and supervised machine learning were used in the modeling of weather and soil attributes.

3 Study Area

Jodhpur district holds the coordinates at 26°N and 27°37' North Latitude and 72°E and 73°52' East Longitude, which fall in the arid zone of the state topologically and comprises 11.60% of the complete arid region [17], and without any perennial river in the district, it is quite arid. Our study area Jodhpur faces frequent droughts due to erratic rainfall in the area. Most of the rainfall 82% occurs in bulk over the months of July to September shown in Fig. 1.

4 Dataset and Technology Used

4.1 Satellite Data

Landsat satellite imagery offers high-quality and varied images of the earth's surface. Landsat images as remote sensing images are special images, containing vast amounts of land, water, vegetation information collected according to lat-long with band 1–7 with visible and invisible light.

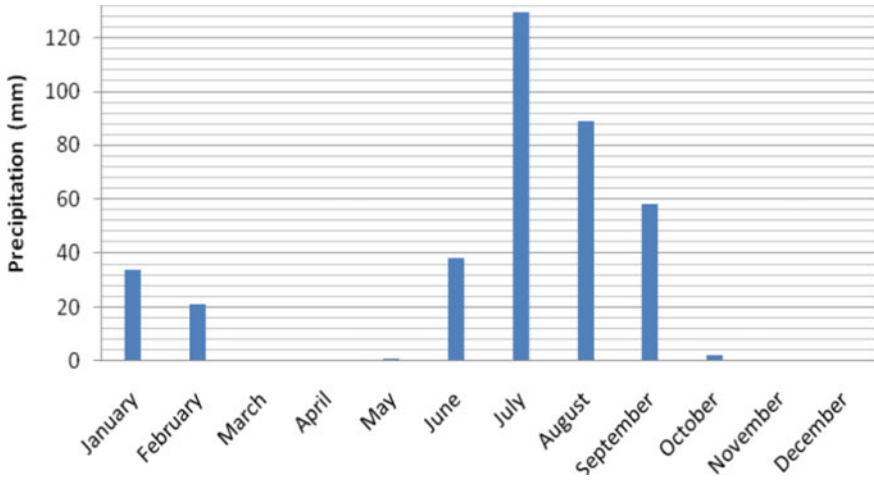


Fig. 1 Monthly precipitation data for Jodhpur

Table 1 Landsat 8 bands and respective characteristics

Band no.	Name	Wavelength (μ)	Characteristics and use
1	Visible blue	0.45–0.52	Maximum water penetration
2	Visible green	0.52–0.60	Good for measuring plant vigor
3	Visible red	0.63–0.69	Vegetation discrimination
4	Near-infrared	0.76–0.90	Biomass and shoreline
5	Middle infrared	1.55–1.75	Moisture content of soil
6	Thermal infrared	10.4–12.5	Soil moisture, thermal mapping
7	Middle infrared	2.08–2.35	Mineral mapping

As of now, there are two present Landsat satellite imagery satellites such as Landsat 8 (2013–2021) and Landsat 7 (1999–2021). Landsat 8 satellite imagery is chosen for this work, since regular international installations are provided free of cost and availability of Landsat’s archives dates from 1972. Landsat 8 bands with their characteristics of the access information are given in Table 1

4.2 Procurement of Satellite Images

USGS Earth Explorer. The USGS Earth Explorer data portal has an abundance of geospatial datasets. With an interactive map, the user can access Landsat imagery with ease [18].

Google Earth Engine. Google Earth Engine [19] hosts a multi-petabyte collection of geospatial datasets and satellite imagery which have planetary-scale analysis capabilities to detect changes, map trends, and quantify differences on the earth’s surface. Google Earth Engine hosts the Landsat collections which are part of the Google cloud public data program. Google Earth Engine [19] also has APIs [20] which enable the analysis of various datasets.

We used the Landsat imagery in our work from Earth Engine with their API. Their NDVI image collection of Landsat 8 collection 1 Tier 1 is created from all the scenes in the 8-day period starting from the first day of the year and end to the year’s 360th day.

4.3 Processing of Landsat Data Using SAGA

SAGA [21], which stands for system automated geoscientific analysis, is an open-source, robust software for analysis of geospatial data. We used the Landsat imagery downloaded by our script to analyze NDVI patterns in SAGA. The datasets used were strategically chosen from pre-monsoon and post-monsoon months to get best results.

Procurement flow of satellite data using Python script for the NDVI extraction of selected time frame for the specified lat-long is shown in Fig. 2.

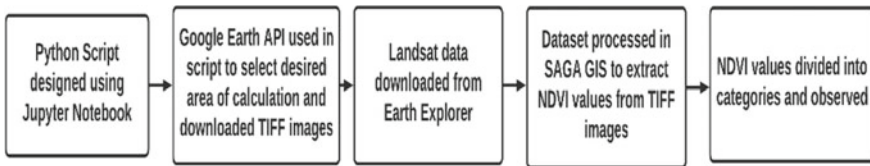


Fig. 2 Procurement flow of satellite data using Python script

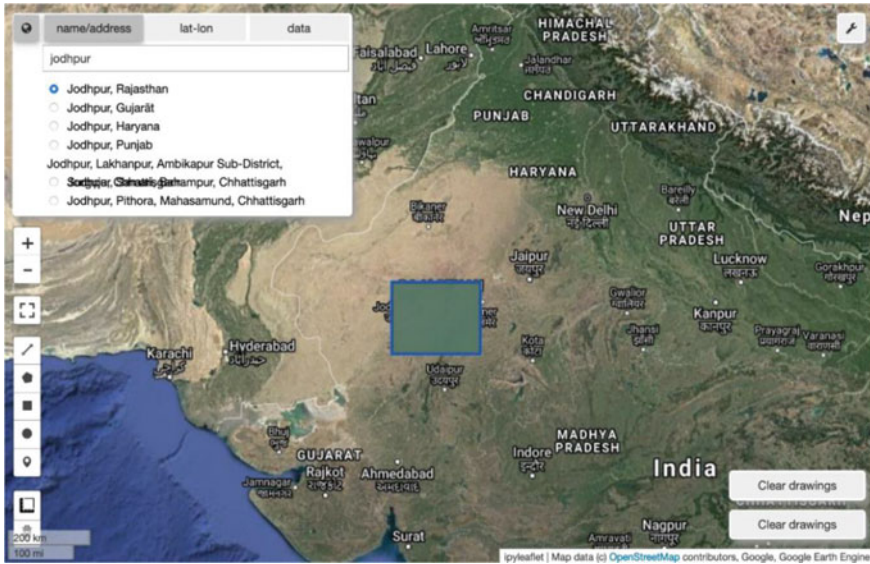


Fig. 3 Interactive map in Jupyter Notebook

5 Results and Discussion

5.1 Visual NDVI Jupyter Notebook

A Python script using Jupyter Notebook written by us was used to get TIFF images which in real-time lets you:

- Select the desired area of NDVI calculation with the help of an interactive map with search and draw features. The coordinates of the polygon drawn are used to fetch the data from Google Earth Engine.
- We also need to select a desired time range for which we need our data. This can be done using the date selection slider widget in the Python script.
- The API will fetch the desired data from the Earth Engine database with the required filters which will be downloaded to your PC's download folder for offline use and further processing in other software (Fig. 3).

5.2 Extracting NDVI Bands from TIFF Images

We downloaded the Landsat 8 dataset of the Jodhpur area using Python script. This dataset is processed in SAGA GIS to acquire the NDVI bands from the Geotiff images. We loaded both the .tiff images of Landsat 8 band 4 and band 5 as we are using Landsat 8 imagery to calculate the NDVI values in SAGA. Fig. 4a, b depicts NDVI values

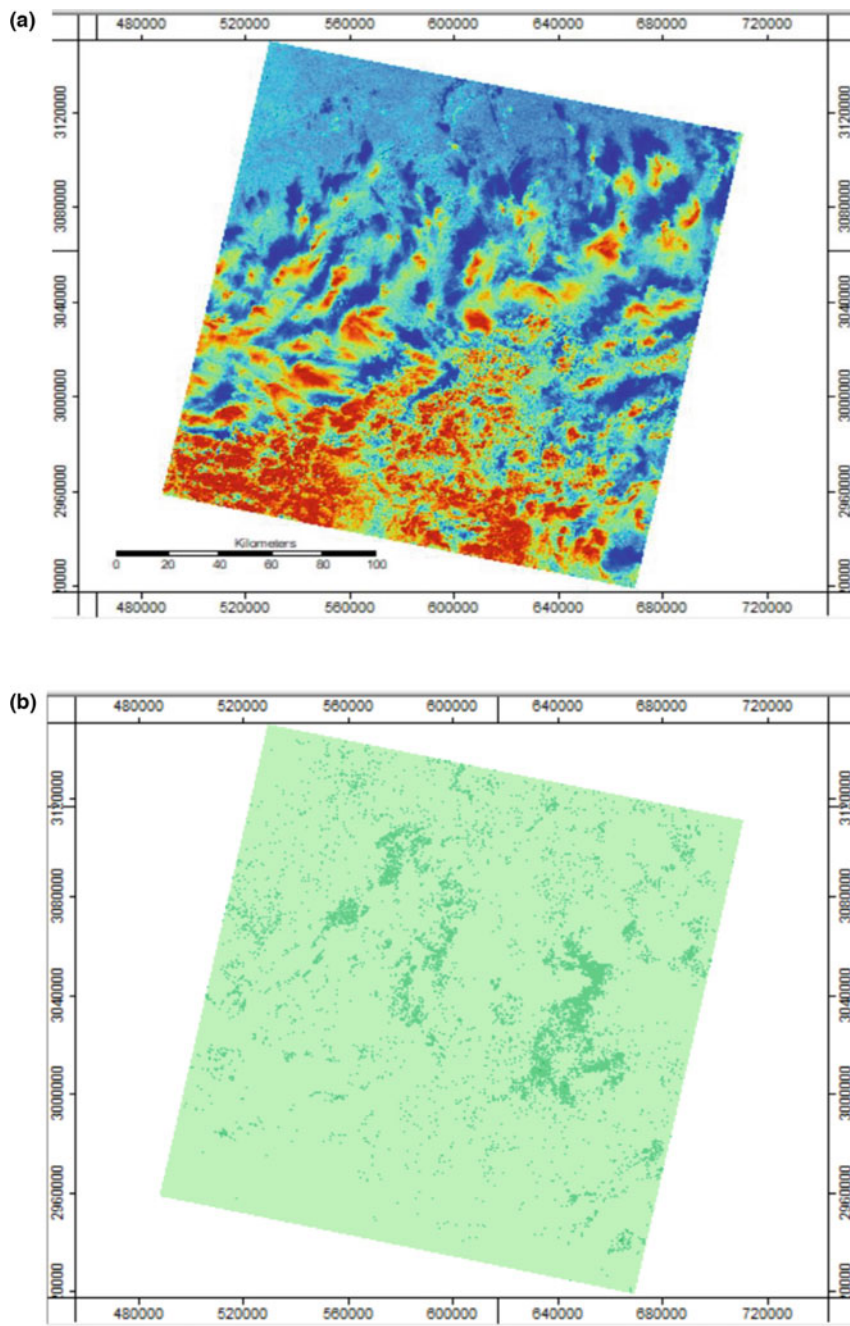


Fig. 4 a Before NDVI band calculation. b After NDVI band calculation





	Color	Name	Description	Minimum	Maximum
1		No vegetation	No vegetation	-1.000000	0.200000
2		Less dense veg	Less dense vegetation	0.200000	0.400000
3		Moderately dense veg	Moderately dense vegetation	0.400000	0.600000
4		Dense healthy veg	Dense healthy vegetation	0.600000	1.000000

Fig. 5 Categories of NDVI

before the band calculation and after NDVI band 4,5 calculation, respectively.

5.3 Change in Categorical NDVI Values

We have analyzed both the results of pre-monsoon and post-monsoon and segregate the obtained values into four classes to observe changes in pre-monsoon and post-monsoon vegetation categories as depicted in Fig. 5.

We have considered May month as pre-monsoon and December month as post-monsoon month with time lag of 4 months for comparison and analysis of the result.

A spike is observed in the category of less dense vegetation after the major rainfall months of June-July as shown in Fig. 6b which is not appearing in before pre-monsoon month as depicted in Fig. 6a. This justifies the fact that healthier vegetation flourishes in the subsequent post-monsoon months of healthy rain.

6 Conclusion and Future Scope

The objective of this paper was to analyze vegetation cover in Jodhpur area based on NDVI values. It was seen that rainfall is a major factor in contributing toward the growth in vegetation cover of an area according to extraction and classes of NDVI values. The NDVI values provided a useful source in assessing the green cover through band 4–5 composition. SAGA-based analysis for categorical NDVI shows the spike in vegetation after monsoon months to prove rainfall was the major factor in a better vegetation cover in Jodhpur due to lack of other factors like absence of perennial rivers.

With more available data and more extensive analysis of the regions, a predictive model can be constructed which will help in predicting vegetation health in future which will help the local farmers with their irrigation planning and crop harvesting.

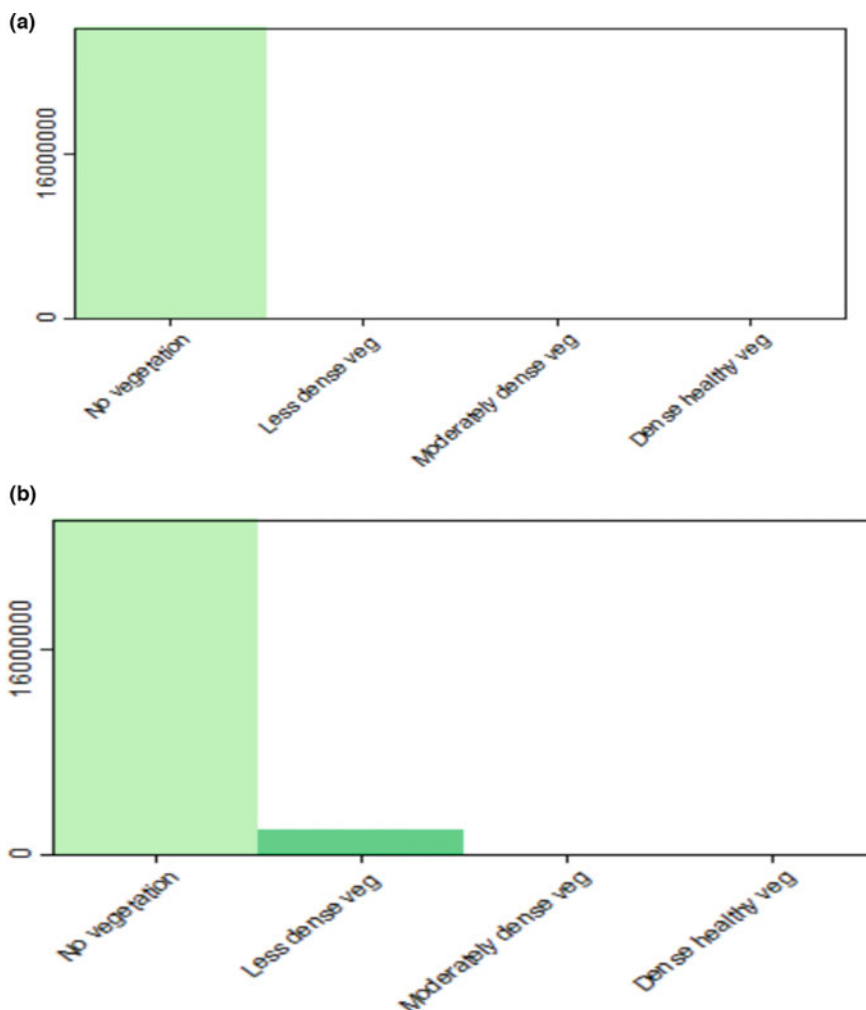


Fig. 6 **a** Comparison of vegetation in May 2019 (pre-monsoon). **b** Comparison of vegetation in December 2019 (post-monsoon)

References

1. E. Burchfield, J.J. Nay, J. Gilligan, Application of machine learning to the prediction of vegetation health, in *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol, XLI-B2 (2016), pp. 465–469
2. H. Goyal, C. Sharma, N. Joshi, Estimation of monthly rainfall using machine learning approaches, in *2017 International Conference on Innovations in Control, Communication and Information Systems (ICICCI)* (IEEE, 2017, August), pp. 1–6
3. S.K. Yadav, S.L. Borana, Modis Derived NDVI Based Time Series Analysis of Vegetation in The Jodhpur Area.”, in *ISPRS—International Archives of the Photogrammetry, Remote Sensing*

- and Spatial Information Sciences*, vol. 5XLII-3/W6 (2019), pp. 535–539. <https://doi.org/10.5194/isprs-archives-xtlii-3-w6-535-2019>
4. H. Goyal, N. Joshi, C. Sharma, An empirical analysis of geospatial classification for agriculture monitoring. *Procedia Comput. Sci.* **132**, 1102–1112 (2018)
 5. H. Goyal, C. Sharma, N. Joshi, An integrated approach of GIS and spatial data Mining in big Data. *Int. J. Comput. Appl.* **169**(11), 1–6 (2017)
 6. A. Kundu, S. Dwivedi, D. Dutta, Monitoring the vegetation health over India during contrasting monsoon years using satellite remote sensing indices. *Arab. J. Geosci.* **2**(9), 1–15 (2016)
 7. H. Goyal, N. Joshi, C. Sharma, Feature extraction in geospatio-temporal satellite data for vegetation monitoring, in *Emerging Trends in Expert Applications and Security* (Springer, Singapore, 2019), pp. 177–187
 8. K. Chi et al., Modelling the vegetation response to climate changes in the Yarlung Zangbo River basin using random forest. *Water* **12**(5), 1433 (2020). <https://doi.org/10.3390/w12051433>
 9. D. Xu, X. Guo, Compare NDVI extracted from Landsat 8 imagery with that from Landsat 7 imagery. *Am. J. Remote Sens.* **2**(2), 10–14 (2014)
 10. T.D. Acharya, I. Yang, Exploring Landsat 8. *Int. J. IT Eng. Appl. Sci. Res. (IJIEASR)* **4.4**, 4–10 (2015)
 11. N. Gorelick et al., Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **202**, 18–27 (2017)
 12. P. Lemenkova, SAGA GIS for computing multispectral vegetation indices by Landsat TM for mapping vegetation greenness. *Contemp. Agric.* **70**(1–2), 67–75 (2021)
 13. A.K. Bhandari, A. Kumar, G.K. Singh, Feature extraction using normalized difference vegetation index (NDVI): a case study of Jabalpur city. *Procedia Technol.* **6**, 612–621 (2012)
 14. R. Medar, V. Rajpurohit, A. Ambekar, Sugarcane crop yield forecasting model using supervised machine learning. *Int. J. Intell. Syst. Appl.* **11**(8), 11–20 (2019)
 15. T.S. Chouhan, *Space Technology and GIS for Disaster Monitoring and Mitigation* (Scientific Publishers, 2018)
 16. <https://earthexplorer.usgs.gov/>
 17. J.N. Schmid, *Using google earth engine for Landsat NDVI time series analysis to indicate the present status of forest stands* (Georg-August-Universität Göttingen, Basel, Switzerland, 2017)
 18. https://developers.google.com/earth-engine/guides/python_install
 19. http://www.saga-gis.org/saga_tool_doc/7.8.0/index.html
 20. https://library.wmo.int/doc_num.php?explnum_id=7768
 21. https://developers.google.com/earth-engine/api_docs

FAMDM: An Approach to Handle Semantic Master Data Using Fog-based Architecture



Saravjeet Singh, Jaiteg Singh, and Jatin Arora

Abstract In digital era, an exponential growth of online transactions leads to a growing need for efficient data storage and management mechanisms. Distributed access to data from different sources generates data quality issues for an organization's data. Master data management (MDM) has been used to provide a single copy of reference data to avoid data redundancy and quality issues. Master data of an enterprise is organizational data that need less frequent changes, and MDM techniques are used to manage master data, which are a critical component of any organization. Large organizations use standard MDM solutions but due to cost constrains, many small and medium size Enterprises (SMEs) are unable to adopt the MDM solutions. MDM solutions used by SMEs are either based on stand-alone or cloud-based approaches. These opted solutions faced poor response time and distributed access issues, to handle these issues, this paper provides an approach to handle semi-structured master data using semantic techniques. The proposed approach uses fog-based architecture to improve the response time. The proposed approach was validated using Web server-based simulation environment, and comparative analysis of the proposed approach with standard cloud-based approach was provided in this paper.

Keywords Data access · Meta data · Data processing · Data sharing · Ontologies

1 Introduction

Digital data in the world are increasing at a very faster pace. Organizations are generating and storing much more data than their previous trends. To perform a real-time operations on a huge amount of an organization's data are very challenging tasks. The organizations that locally store data face many data management issues. To manage data effectively, organizations' data are categorized into domain data and

S. Singh (✉) · J. Singh · J. Arora
Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura,
Punjab 140401, India

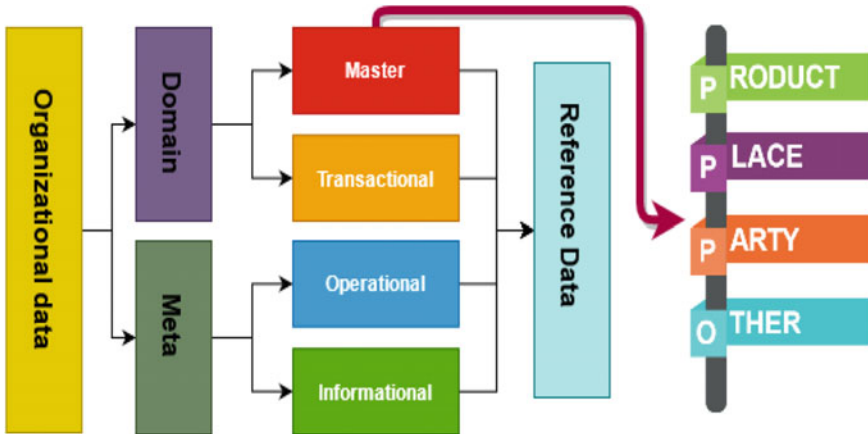


Fig. 1 Categories of organizational data and types of mater data values

metadata. Detailed categories of organizational data and their relationship are shown in Fig. 1.

Domain data include core operational data. It covers complete details of the functional data of an organization. Domain data are further divided into two parts such as master data and transactional data. Master data provide information about business perspectives. It gives the base to the transactional data. Master is always non-transactional data and it provide basic attributes of the business. It includes person, place, price, item, and other enterprise-specific data. Transactional data are actual business data. It includes all cash flow and business activities. These activities involve sale and purchase history, employee and client transactions, location and product-related transaction [1]. Master data can be categorized into 4 basic types as shown in Fig. 2. These are party, place, product, and other organization specific data [1–3]. Master data are required for organization operations such as quality evaluation, update, and enhancement. The term “master data” refers to an organization’s operational data that require only minor changes. Changes in master data are extremely difficult to accomplish, and specific provisions are necessary to perform changes in master data. Master data are a kind of operational data that have high worth, characterize center data that helps in basic dynamic, and taking care of business forms over the venture. Master data management (MDM) is a process used to handle the master data. MDM is responsible for the creation, updating, and deletion of the master data [4, 5].

MDM is very challenges process due regulatory and constrained associated with master data. Master data management used for customer relation management, client integration, employee relationship management, quality management, and other management activities. MDM is a process, which is used to enhance the quality of data and process flow in an organization. MDM is used to control the quality of data and creates restrictions on data usage. Master data management helps to

Master data	Place	Address Information
		Site / Outlet information
	Person	Supplier
		Stake Holders
Customers		
Product	Product Details	
	Price	
Other	License Details	
	Certificates etc.	

Fig. 2 Categories of master data

maintain the data integrity and provides a single copy of master data. Many organizations are using MDM for business purposes but due to cost constrain, many small and medium size enterprise (SME) is unable to adopt the MDM process. These SMEs create their own MDM solution, and these solutions are based on cloud or stand-alone system architecture. In this study, we proposed a fog architecture-based semantic MDM (FAMDM), this architecture to handle the MDM using semantic techniques. A complete description of FAMDM is provided in the next sections of this paper. Next section of this paper provides a brief history, the third section provides FAMDM details, the fourth section provides result and discussion, and the last section provides future scope and conclusion.

2 Brief History

Concept of Master data came into existence in late 90s. Siebel used the concept of data organization and further this term derived the concept of master data. SAP started working in the domain of master data and after that many big ventures like IBM, Oracle, Infosys, IBM, Google, Informatics, TCS, and so on started working on MDM. Apart from these big ventures, research community also participated in this field [6–9]. MDM is very frequently used for big organizations but for SMEs, it is not that famous. One of the biggest challenges to implement MDM for SME is its high cost. Few organizations developed their applications to handle the data using MDM approach but these solutions do not provide 360-degree view of data. Moreover, some customized solutions provide MDM facilities but fail to cover all

aspects of the organization's data. SME-based MDM solutions use cloud architecture or stand-alone system architecture.

Murthy et al. provided step by step approach to be followed to implement the MDM process for an organization. This study provided case study to explain the dependencies involved in MDM and how these dependencies can be resolved [10]. Many MDM frameworks were provided by research community to handle master data of organizations. These frameworks provided decision model to identify the appropriate MD architecture, rules to implement data protection issues, and ways to handle SME's data [1, 11, 12]. Zhao et al. provided method to evaluate the big data generated using the data network of any organization [13]. Ganesan et al. provided MDM solution using graph-based method to identify people connected to COVID positive person [14].

3 Proposed FAMDM Technique

Organizational data are increasing at a very faster rate and have characteristics of big data. Relational data management techniques are not capable to handle huge amounts of unstructured or semi-structured data. Semantic techniques provide an effective way to handle semi-structured data. Computer systems are increasingly moving to cloud infrastructure and utilizing cloud technology in various ways as the Internet and mobile computing evolve. The massive demand for unified cloud storage is raising serious problems such as reduced spectral performance, high latency, low connectivity, and security concerns [15]. Large enterprises handle master data using standardize software and require huge resources and costs. SMEs cannot afford these expensive solutions and create their own MDM solutions. Fog architecture-based semantic MDM (FAMDM) was proposed in this paper.

FAMDM uses a fog computing approach to handle the semantic technique-based MDM. According to FAMDM, the cloud layer handles all the data that include transactional, master, Reference, and data warehousing. Resource descriptor framework was used to store the master data. Ontologies were created to handle the master data. Master data management steps were performed using semantic queries. Standardized elements and data of food chain SMEs were considered for this experiment. Figure 3 shows the core element of master data using semantic techniques.

A complete ontology was created to handle the food chain master data (as shown in Fig. 4). Table 1 provides the brief detail of considered master data elements. Created semantic master data management solution was created and managed at cloud layer. After creating the consistent view of master data, a copy of master data is saved at the fog layer, which is near to users and users can access master data using edge elements. Semantic queries were used to provide the data to the fog layer and were also responsible to provide the atomic view of the data. Figure 5 shows the data distribution and master data access at different layers. Figure 6 shows a sample query to access the master data.

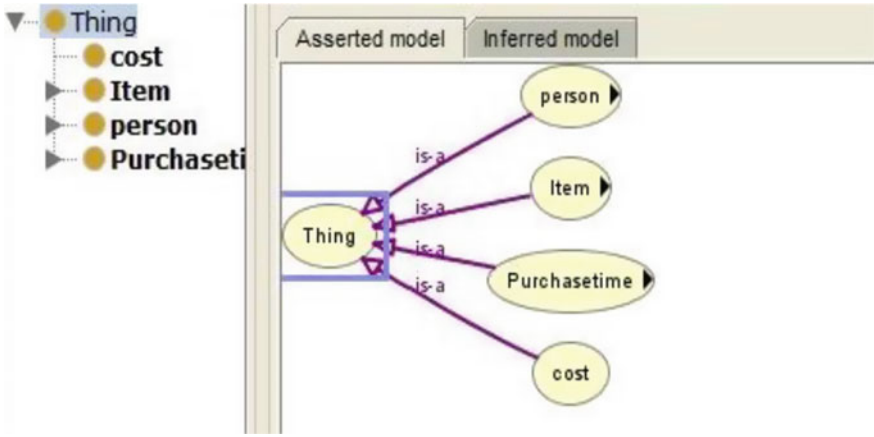


Fig. 3 Master data core elements

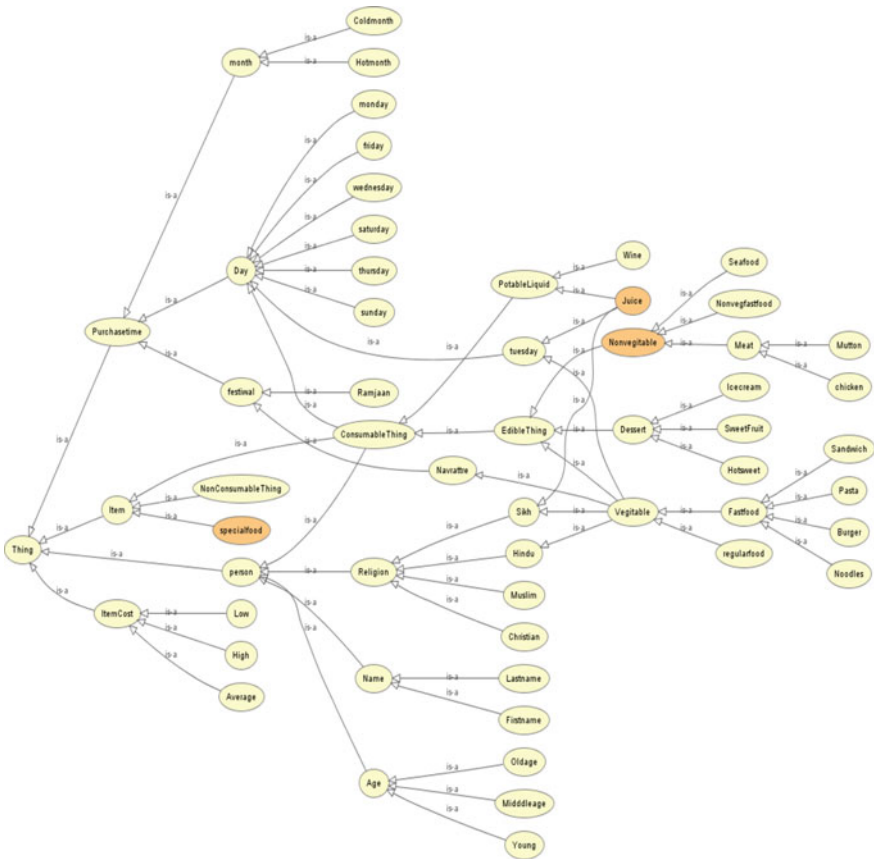


Fig. 4 Food chain master data management solution using an ontology

Table 1 Considered master data elements details

Master data element	Details
Person	Name of owner, customer, supplier, age, and religion
Item	Product name, specifications
Cost	Price
Purchase time	Day, festival
Location	Outlet address and location details

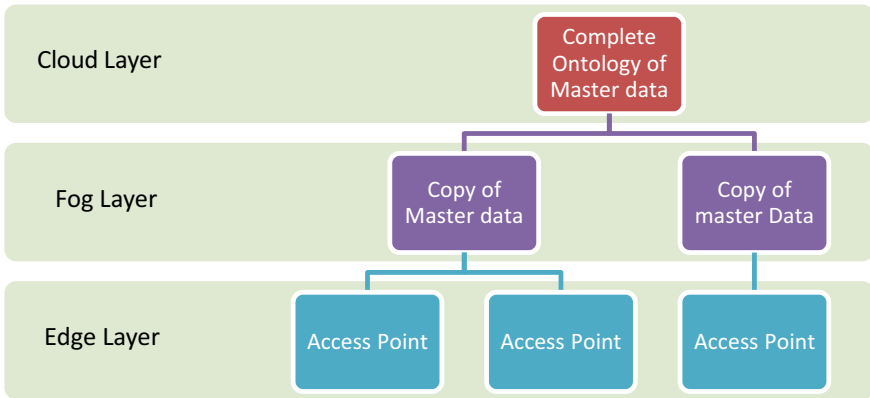


Fig. 5 FAMDM technique to handle SME master data

```

SPARQL query:
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?subject ?object
WHERE { ?subject rdfs:subClassOf ?object }
    
```

Fig. 6 Sample query to access the sematic master data

4 Results and Discussion

To validate FAMDM, a simulation environment was created using a Webserver, load balancer, and ontology builder. One cloud port and 2 fog ports were created using a

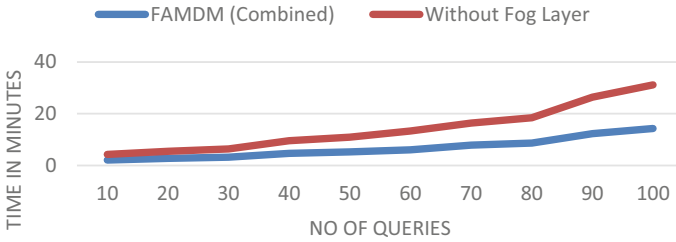


Fig. 7 Response time based on number of queries

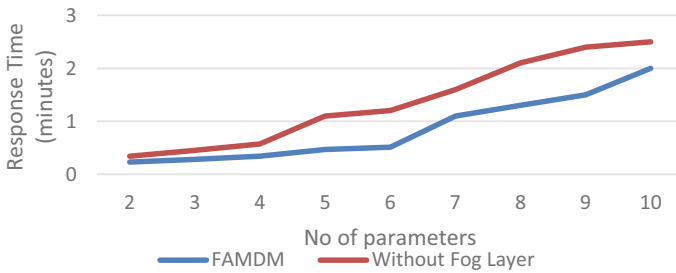


Fig. 8 Response time based on number of parameters in queries

Web server. Sample data of fast food joint were considered. Three different access points were created to access the master data. To analyze the storage requirement and response time queries with different parameters and selectors were executed on FAMDM.

A comparison of FAMDM with a cloud-based approach based on response time is shown in Fig. 7. FAMDM technique has less response time in comparison with cloud-based techniques. Two queries were executed with 2 to 10 parameters on both techniques and response time were analyzed. Figure 8 shows the impact of the number of parameters in a query on response time. As per the Figs. 7 and 8, FAMDM has less response time in comparison with cloud-based approach. FAMDM approach requires additional space for maintaining different copies of the master data.

5 Conclusion and Future Scope

This paper presented a master data management technique using an ontology and fog computing-based architecture for small and medium-sized enterprises. Due to the high cost of existing MDM solutions, SMEs cannot afford them. The presented FAMDM approach handles master data using ontology and provides fog-based data access. FAMDM technique has three layers; cloud, fog, and edge layer. Master data are placed at the cloud layer, and all fog layers have copies of master data. Each user

can access the master data fog layer, and the cloud layer is responsible to provide a consistent view of data. The proposed FAMDM technique was validated using a food joint data set. According to performed analysis, the response time of the FAMDM technique is better than the cloud-based technique. Additional space is required to store multiple copies of master data at each fog layer. In future, this technique can be extended to use a better ontology access method. Data security and privacy can also be handled in the extended version of FAMDM.

References

1. S. Singh, J. Singh, SSMDM: An approach of big data for semantically master data management, in *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)* (IEEE, 2015), pp. 586–590
2. V. Siner, Master data management in the EU. comparative analysis of access to base registries (2020)
3. S. Singh, J. Singh, Management of SME's semi structured data using semantic technique, in *Applied Big Data Analytics in Operations Management* (IGI Global, 2017), pp. 133–164
4. J. Kokemuller, A. Weisbecker, Master data management: Products and research, in *ICIQ* (Citeseer, 2009), pp. 8–18
5. N. Qodarsih, S.B. Yudhoatmojo, A.N. Hidayanto, Master data management maturity assessment: A case study in the supreme court of the republic of Indonesia, in *2018 6th International Conference on Cyber and IT Service Management (CITSM)* (IEEE, 2018), pp. 1–7
6. E. Akhmetshin, I. Ilyina, V. Kulibanova, T. Teor, Special aspects of master data-based integrated management of region reputation in modern it environment, in *IOP Conference Series: Materials Science and Engineering*, vol. 497 (IOP Publishing, 2019), p. 012022
7. I. Athanasiadou, *Evaluating the Maturity of Companies in Supplier Master Data Management: The Design of a Maturity Model* (2019)
8. L.K. Fernando, P.S. Haddela, Hybrid framework for master data management, in *2017 Seventeenth International Conference on Advances in ICT for Emerging Regions (ICTer)* (IEEE, 2017), pp. 1–7
9. S. Khillari, Impact of coronavirus on master data management market—growth, trends and forecast report, 2026 (2020)
10. K. Murthy, P.M. Deshpande, A. Dey, R. Halasipuram, M. Mohania, P. Deepak, J. Reed, S. Schumacher, Exploiting evidence from unstructured data to enhance master data management. *Proc. VLDB Endowment* **5**(12), 1862–1873 (2012)
11. M. Spruit, K. Pietzka, MD3M: The master data management maturity model. *Comput. Hum. Behav.* **51**, 1068–1076 (2015)
12. D. Subotic, V. Jovanovic, P. Poscic, Data warehouse and master data management evolution—a meta-data-vault approach. *Issues Inform. Syst.* **15**(2) (2014)
13. C. Zhao, L. Ren, Z. Zhang, Z. Meng, Master data management for manufacturing big data: a method of evaluation for data network. *World Wide Web* **23**(2), 1407–1421 (2020)
14. B. Ganesan, S. Parkala, N.R. Singh, S. Bhatia, G. Mishra, M.A. Pasha, H. Patel, S. Naganna, Link prediction using graph neural networks for master data management. arXiv preprint [arXiv: 2003.04732](https://arxiv.org/abs/2003.04732) (2020)
15. S. Singh, J. Singh, Location driven edge assisted device and solutions for intelligent transportation, in *Fog, Edge, and Pervasive Computing in Intelligent IoT Driven Applications* (2020), pp. 123–147

A Smart Mobile Application for Stock Market Analysis, Prediction, and Alerting Users



Rutvi Boda, Saroj Kumar Panigrahy , and Somya Ranjan Sahoo

Abstract In this paper, an efficient Android mobile application for stock market analysis and alert is presented which can be used by the users for accurate and efficient trading in stock market. Stock market prediction is a method using which it is tried to determine or predict the future values of any company's stocks, or any other financial instruments traded in a financial exchange. The prediction of stock price is significant in any financial field. All investments should be backed by a sturdy research, and the second thing is that time is another essential factor because the stock market is totally the place where you snooze, you lose is true. The aim of this paper is to study the various algorithms such as long short-term memory neural networks and deep learning and reach a solution which gives an accurate prediction so that even the beginners can start trading in the stock market. And once reaching the optimal solution, this paper explores an Android application which can send notifications to the user at the time which is optimal for his investment so that she does not miss out on any opportunities coming her way. In short, utilizing technologies like machine learning in Jupyter Notebook, Google cloud, Android Studio based on Java, and Alpha Vantage application programming interface, the study covers the development of an Android application and finding the optimal algorithm to achieve our objective.

Keywords Mobile application · Stock market · Prediction · Forecasting · LSTM · Deep learning

1 Introduction

The advent of mobile computing has opened up applications of this technology in several areas of the enterprise. The ability to check for information on the fly or execute transactions and display analytics for decision-making has made mobile-based applications a powerful tool in all sectors of the industry. Fundamentally, quantitative merchants with a ton of cash from financial exchanges purchase stocks

R. Boda · S. K. Panigrahy (✉) · S. R. Sahoo
VIT-AP University, Amaravati, Andhra Pradesh 522237, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022
P. Verma et al. (eds.), *Advances in Data Computing, Communication and Security*,
Lecture Notes on Data Engineering and Communications Technologies 106,
https://doi.org/10.1007/978-981-16-8403-6_34

379

subsidiaries and values at a modest cost and later sell them at exorbitant cost. The stock market prediction is an old thing and yet is being considered by various organizations. Stock value forecast has been at center for quite a long time since it can return critical benefits. There are two methods used for forecasting stock prices—fundamental analysis and technical analysis. Investors use these two types of analyzes before they invest in a stock market. The fundamental analysis deals with the looking at the inherent value of stocks, performance of the industry, economy, and political climate and to decide whether to invest or not. The technical analysis deals with the study of the statistics generated by market activity based on volumes and past prices. In recent years, expanding noticeable quality of artificial intelligence (AI) in different businesses have enlightened numerous merchants to apply AI and machine learning (ML) techniques in this field and have resulted in very promising results.

Stock market follows the arbitrary walk also known as random walk, i.e., that the best prediction that we can have about tomorrow's values is today's value. Unquestionably, the forecasting stock files are troublesome because of the market unpredictability which needs precise figure models. The stock market lists are exceptionally fluctuating, and it impacts the investor's confidence. Stock prices are viewed as vigorous and prone to snappy alterations due to the basic environment of the commercial domain and to some degree due to the blend of known boundaries like previous day's closing price and the obscure elements or unknown factors like rumors or election results. Researchers have been trying to predict the stock market using AI and ML techniques. The focus varies due to three main factors: (a) The targeting price change which can be near-term which is less than a minute, short-term which is tomorrow to a few days later, and long-term which is months later; (b) The set of stocks which can be limited to less than 10 particular stocks, to stock in particular industry, to generally all stocks; (c) The prediction which is used which can range from a global news and economy trend, to particular characteristics of the company, to purely time series data of the stock price [1].

This paper will be comparing different algorithms in which the stored live and historical stock prices are compared, and that data will be treated as a training dataset for the program. The primary motivation behind the forecast is to lessen vulnerabilities related to investment decision-making. The following are the objectives of the paper.

- To compare different algorithms and find out which is best suited.
- Using the algorithm to predict the future stock price of any company.
- To design an application which takes inputs from the user like email id, company of interest, range in which they are interested to purchase, etc., and upload these details online.
- A program that runs continuously in the backend and compares the stock price and the range given.
- To send an alert email to the user when the stock price of the company they are interested in falls in the given range.

The rest of this paper is organized as follows. Section 2 describes the background and related works. Section 3 explains the proposed system and working methodology.

Section 4 describes the software and hardware details of the mobile application system. Section 5 discusses the results obtained when the mobile app was deployed. Finally, Sect. 6 concludes the paper with a scope for future development.

2 Background and Related Works

Yadav presents about one of the most all-round sectors of the Indian financial system which is the stock market [2]. The author gives the definition of stock market, how transactions take place, volatility of the market, and other such technical terms which provides a good foundation to the study behind this paper. The author has based his study on the stock market on factors like inflation rate, economic growth, corporate earnings, volume, returns generated, and so on. Similar background study is provided by Bala where the author gives information about all investments available out there in the world [3]. Even though we know the definition of stock market and the technical terms too, we cannot proceed unless we know the importance of investment. The study by Thomas points to how important this can be for the general public as it is not a cup of tea for everyone to do a complete technical analysis to invest in the market [4]. The paper talks about the principles of technical analysis which gives a basis to the development of the study.

Rajput and Bobde have studied the methods for stock market movement prediction and determined the factors affecting price only, human sentiment classification, using latent Dirichlet allocation (LDA)-based method, joint sentiment-topic (JST)-based method, and support vector machine (SVM) [5]. Kute and Tamhankar have proposed an algorithmic approach to stock market analysis which gives probable solutions [6]. Hegazy et al. have proposed a machine learning model for stock market price prediction. They have integrated particle swarm optimization (PSO) and least square SVM (LS-SVM) to optimize the prediction of daily stock prices [7]. Khan et al. have used social media and news to predict stock market using machine learning classifiers [8]. The authors have used algorithms on financial news data and social media to determine the effect of this data on expectation accuracy of stock market for coming ten days. Authors have researched on pertinency of recurrent neural networks in LSTM networks, exploring the model summary and model structure and giving important results like average returns and so on [9, 10]. Yang et al. have studied stock market prediction using neural networks and proposed a framework for the model and the algorithm too [11]. All these papers have been the background study for the entire paper giving a proper foundation for this study.

3 Mobile Application for Stock Market Analysis and Alert

This section describes the proposed system, working methodology, and standards details.

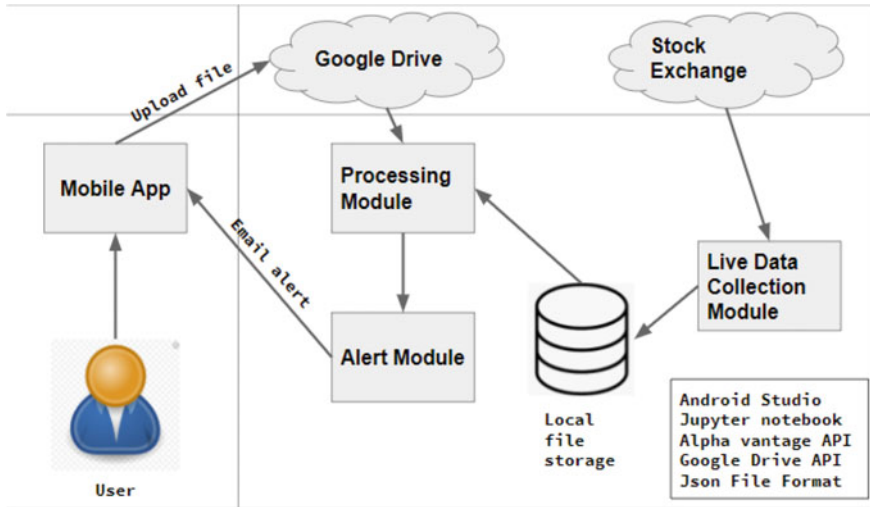


Fig. 1 System architecture of stock market analysis and alerting users using Android mobile application

3.1 Proposed System

This system consists of one main section which includes the software. The software includes a mobile application, live stock exchange data collection module, alert module, and processing module using machine learning. The hardware part is in the form of a mobile phone and is used for deploying and testing the app. The architecture of the proposed system is depicted in Fig. 1.

3.2 Working Methodology

The system has two sections—an Android mobile application and Python program. The application starts by asking the user to login using her credentials, i.e., her phone number and password. If the user is not already registered, then she is asked to sign in using her phone number. Once the user registers herself, she is directed to the login page where she is asked to login using the same credentials, she used to register herself. Once she logs in using her login credentials, she is directed to home page where she is asked to choose the company whose stocks, she is interested in buying; she is asked to enter her email address where she would receive her alert email and her lower and upper price limits, i.e., the cost range she is willing to pay for each stock. Once she enters these details, she can click on the confirm input button. Once the button is clicked, a JSON format file is made with all these details and she is then directed to the verification page where she is asked to verify all the details, she

entered in the home page. If there are any changes, she could go back to the home page by clicking the back button where she could make changes before submitting her entry. Once she verifies all the details entered are correct, she can click the submit button which then uploads the file created in JSON format to google drive. Once the JSON file is uploaded successfully, the user gets an SMS notification saying the file has been uploaded and the request has been generated. The user can then click the logout button and she would be logged out and directed to the login page.

The Python program written using Jupyter Notebook [12] is divided into four sections. First one is where it collects the live data from New York Stock Exchange through Alpha Vantage application programming interface (API) [13]. Alpha Vantage API offers free stock APIs in JSON and comma separated value (CSV) formats for real-time and historical equity, forex, cryptocurrency data, and over 50 technical indicators. The second part uses different types of plots to show the daily returns of company stocks, shows the risk involved in investing in company stocks, shows the correlation plot of daily returns, comparison of daily returns of one company with other companies, shows the graph with adjacent closing price and moving average of each company over the past 10, 20 and 50 days, shows the graph of total vol being traded for each company, historical view of the closing prices of each company, and uses different algorithms to predict the future values of stocks. The third part compares and checks if the current price of the selected company's stock lies in the range specified by the user. And the last section is the notification section where an email alert is sent to the user's registered email id when the stock price lies in the specified range.

4 Mobile Application Implementation Details

This section describes the software and hardware details for the implementation of the proposed system.

4.1 Software Details

The main software in this application includes an Android app and Python programming. In addition to this, a machine learning algorithm is used for stock market analysis.

Android Application. The Android application is built using Android Studio which is the official integrated development environment (IDE) for development of Android applications [14]. It allows a developer to create applications which are compatible with various mobile devices.

Alpha Vantage API. Alpha Vantage API is a method to obtain historical and real-time data for several markets like stocks, forex, and cryptocurrencies [13]. We can

access the data directly in Python or any other programming language. Also, the data can be manipulated or stored for later use. Several time frames are available ranging from 1-min bars up to monthly durations. Also, there are more than 50 technical indicators available as well as performance data for 10 US equity sectors.

Google Cloud Platform. It is a set of cloud computing services by Google and provides infrastructure as a service, platform as a service, and serverless computing environment [15]. After the application is created, it needs authentication of the app so that it can access the Google Drive (the storage service by Google).

Simple Mail Transfer Protocol (SMTP). It is a communication protocol for sending an email to one or more recipients.

Python Programming. As shown in Fig. 1, various Python programs are being used in this system. The detail of each program is as follows.

Collecting Live Data: The live data of stock prices are collected using the Alpha Vantage API [13]. The data are stored in the JSON format. We create another JSON file which contains only the symbol and current price of all the company stocks because that is what we will be using to send the notification to the user. The first file will be used to predict future values and display different graphs of comparison. After the other file is created, we call the next program that would check if the price was in the range or not.

Stock Analysis: This program plots many different types of graphs like calculating and plotting the adjacent closing price, moving average over the past 10, 20, and 50 days. Calculating the daily returns of each company and plotting them. Calculating the risk in investing for each company and plotting its graph, etc. Closing price is the last price at which the stock is traded. Figure 2 shows the historical view of the closing price of each company. Figure 3 shows the total day-wise volume traded of each company. Figure 4 shows the adjacent closing and moving average of 10, 20, and 50 days for each company.

Stock Price Prediction: The main function of this program is to predict the future stock prices of all companies using LSTM and deep learning. Figures 5 and 6 show the output graph of LSTM and deep learning methods, respectively.

Drive Access: This program is used to access the details file uploaded from the Android application to Google Drive.

Comparison: This program is called every time a JSON file is with the current price. From the drive access program, it takes the details file that has all the user details. Each time when this program is called it compares the current price with the price range given by the user. If the current price lies in the range, it calls the *notify* program. Once the *notify* program is called it will not be called again till a certain amount of time, say n minutes. This program will have a 45 s sleep time because once the JSON file is created, there is a 45 s sleep time given.



Fig. 2 Historical view of closing prices of different stocks of each company

Notify: From the details file that we get from the drive access program, it takes the email id that the user entered where he wants to receive the email notifications. Every time this program is called it sends an email notification saying “price in given range. Probably the best time to purchase stock,” so that the user knows that the stock price of the company she was interested in is available in the range she is interested in.

5 Results and Discussion

The application was successfully able to take input from the user, convert it into JSON format and save it as a file, upload it to google drive, and send an SMS notification to the user once the file is uploaded to the drive successfully. Figure 7 shows the different activities of the mobile application.

The programs were successfully able to collect live data, predict future values, access the file made through the Android app from google drive, compare the details in the file to the current prices of the chosen company stocks, and send an alert email at a certain interval when the price lies in the specified range which is shown in Fig. 8.

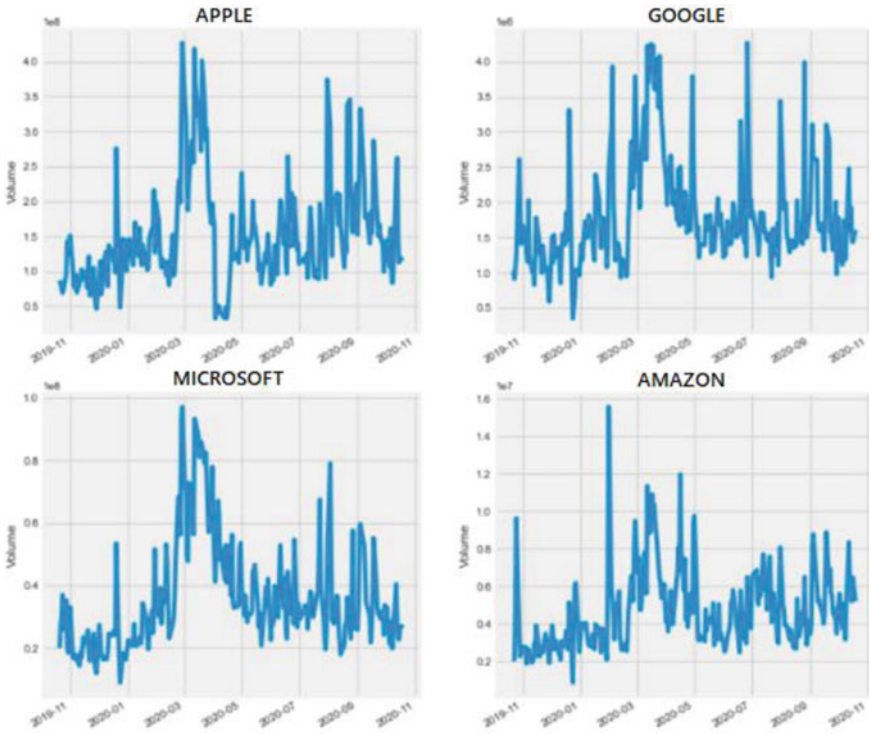


Fig. 3 Total day-wise volume traded of each company

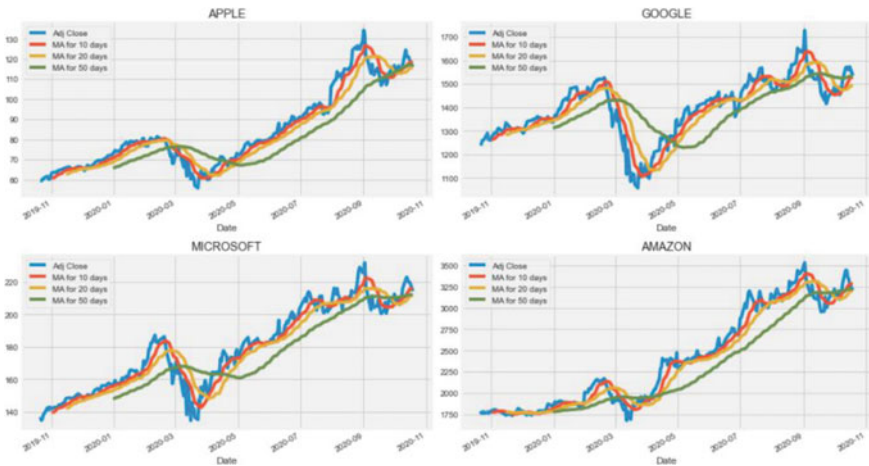


Fig. 4 Adjacent closing and moving average of 10, 20, and 50 days for each company

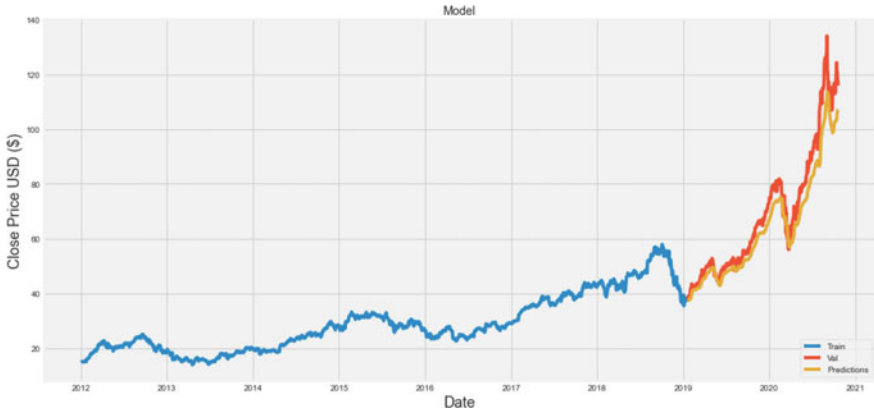


Fig. 5 Stock price prediction using LSTM

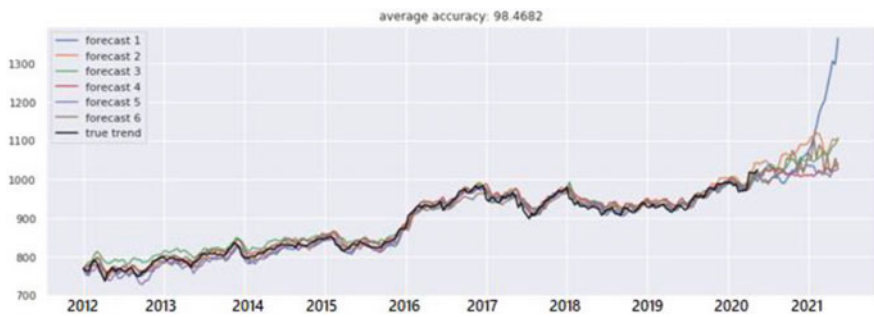


Fig. 6 Stock price prediction using deep learning

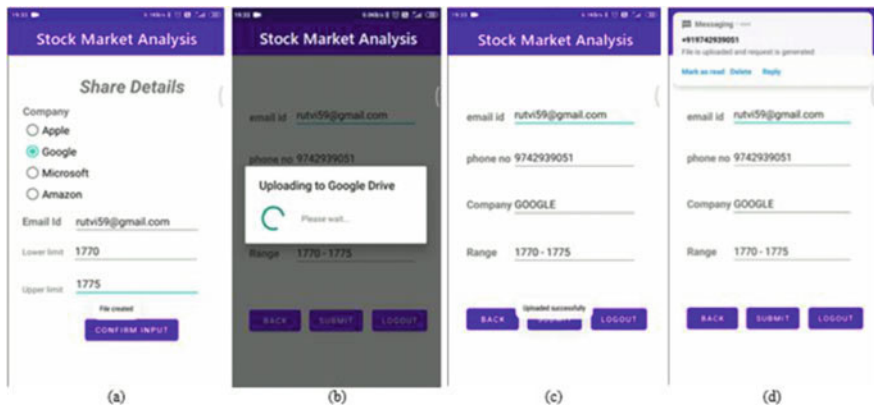


Fig. 7 Different activities: **a** the JSON file has successfully been created and is ready to be uploaded, **b** the file is being uploaded to Google Drive, **c** the file has successfully been uploaded to the Google Drive, **d** the user has been notified by SMS after successfully uploaded

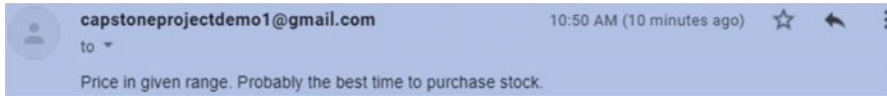


Fig. 8 Email alert when the stock price comes in the range provided by the user

6 Conclusion and Future Work

Lots of profit can be earned through the stock market if a person knows when to invest in a company's stock and when to sell those stocks. The app currently accepts only one user at a time, has only 4 non-Indian company stocks, and notifies a user only when the stock price falls in the given range. As a part of future scope, the app will support Indian company stocks, multiuser, stocks of companies from different countries, and we will be able to send more alerts like get notified when there is a continuous n times increase/decrease in a particular company stock price. Though, these insights represented in Figs. 2, 3, and 4 are generic on the stock data, more feature-based analytics, and their visualization shall be conducted and presented. Currently, the app notifies through email, but as part of future scope, we can have different means through which the alerts can be sent based on the users' preferences like SMS or call.

References

1. Moses Charikar and Chris Ré: CS229: Machine Learning, <http://cs229.stanford.edu>, last accessed 2021/05/09.
2. S. Yadav, Stock market volatility—a study of Indian stock market. *Glob. J. Res. Anal.* **6**(4), 629–632 (2017)
3. A. Bala, Indian stock market—review of literature. *TRANS Asian Res. J.* **2**(7), 67–79 (2013)
4. E. Asha, Thomas: a study on technical analysis and its usefulness in Indian stock market. *Mirror* **3**(2), 159–165 (2014)
5. V. Rajput, S. Bobde, Stock market forecasting techniques: literature survey. *Int. J. Comput. Sci. Mobile Comput.* **5**(6), 500–506 (2016)
6. S. Kute, S. Tamhankar, A survey on stock market prediction techniques. *Int. J. Sci. Res. (IJSR)* **4**(4), 303–306 (2015)
7. O. Hegazy, O.S. Soliman, Mustafa Abdul Salam: a machine learning model for stock market prediction. *Int. J. Comput. Sci. Telecommun.* **4**(12), 17–23 (2013)
8. W. Khan, M.A. Ghazanfar, M.A. Azam, A. Karami, K.H. Alyoubi, A.S. Alfakeeh, Stock market prediction using machine learning classifiers and social media, news. *J. Ambient Intell. Humanized Comput.* (2020)
9. A. Moghar, M. Hamiche, Stock market prediction using LSTM recurrent neural network. *Procedia Comput. Sci.* **170**, 1168–1173 (2020)
10. D.M.Q. Nelson, A.C.M. Pereira, R.A. de Oliveira, Stock market's price movement prediction with LSTM neural networks, in *International Joint Conference on Neural Networks (IJCNN)* (IEEE, Anchorage, 2017), pp. 1419–1426
11. C. Yang, J. Zhai, G. Tao, *Deep Learning for Price Movement Prediction using Convolutional Neural Network and Long Short-Term Memory*, *Mathematical Problems in Engineering*, vol 2020 (Hindawi, 2020)

12. Jupyter Notebook, <https://jupyter.org/>. Last accessed 2021/05/09
13. Alpha Vantage: Free Stock APIs in JSON & Excel, <https://www.alphavantage.co/>. Last accessed 2021/05/09
14. D. Griffiths, D. Griffiths, *Head First Android Development: A Brain-Friendly Guide*, 2nd edn. (O'Reilly Media, 2017)
15. Google Cloud: Cloud Computing Services, <https://cloud.google.com>. Last accessed 2021/05/09

Impact of Blockchain Technology on the Development of E-Businesses



Jatin Sharma and Hamed Taherdoost

Abstract Blockchain technology is one of the greatest discoveries of technology that has been developed as the driving force in many businesses in recent years. One of the main areas that is increasingly using blockchain technology to facilitate its processes is e-business. Businesses soon recognized an opportunity to employ blockchain technology in different operations by investing time and money in the technology. Since both blockchain and e-business involve transitions, blockchain technology is likely to make major differences in the developments of e-businesses. This paper will focus on understanding the impact of blockchain technology on the development of e-businesses. In addition, since reviewing all areas of e-business may be influenced by blockchain technology, this paper will narrow down the topic to supply chain management. Moreover, it points to future expectations of blockchain technology and how it will impact various sectors.

Keywords Blockchain · E-business · Blockchain technology · Business development · Blockchain supply chain · E-business challenges

1 Introduction

Blockchain is a database in which information is recorded in a manner that is hard to hack, impossible to change, and difficult to cheat. Today, there are some areas of business, which have started to utilize the advantages of this blockchain technology including e-businesses [1].

The growth of the Internet and over-reliance of many individuals on it has provided a new requirement for businesses to rethink about their core processes. Offering online presence services is not necessarily a competitive advantage today; however,

J. Sharma (✉) · H. Taherdoost
University Canada West, Vancouver, Canada
e-mail: Jatin.sharma3555@myucw.ca

H. Taherdoost
e-mail: hamed.taherdoost@ucanwest.ca

it has turned into a must. E-businesses involve transactions between parties and the transfer of valuable data. Thus, e-business owners have always been seeking solutions to make the data transfer and money transactions safer and faster to earn trusts of parties. Based on the fundamentals of blockchain, its employment seems to be a promising solution to address this challenge of e-businesses.

Blockchain technology provides secure data transfer, and enables to handle users' activities ranging from filling online forms and setting purchase orders to processing a payment [2]. Therefore, it would benefit both buyers and sellers by offering convenient and smart solutions to considerably alleviate financial security concerns and cyber threats.

1.1 E-Business

Electronic business can be defined as conducting business activities online. However, it is not just limited to sale and purchase, the scope of e-business is broader that it includes activities such as recruiting, procurement, and supply chain management, customer relationship management, delivering the product, sending, and receiving payments [3].

Both start-ups and traditional businesses focus on establishing a reliable and strong e-business because of being cheaper, and having access to a wider range of customers. There are several models of e-business including business to consumers, business to business, government to business, consumer to consumer, government to consumer, and business to government.

1.2 Blockchain Technology

The history of blockchain is linked with Bitcoin. Since Bitcoin was offered to open source communities, the need to have a system that could record the transactions was risen as well [4]. After one decade since blockchain technology was first introduced, still people used to think that blockchain and bitcoin are the same. In 2014, people realized that blockchain technology can be used for other operations as well [5], and started to invest their time and money to explore applications of blockchain in other operations. However, based on statistics [6] on the use of blockchain technology by various sectors in business, blockchain technology has been used mostly in digital currency such as Bitcoin and Ethereum. After that blockchain technology is used for data sharing purposes because the technology promotes authenticity [7]. Thus, organizations use blockchain technology because they know the information is reliable.

2 Focus Areas of Blockchain Technology in Canada

Canada is one of the leading countries in the blockchain sector, and it has gained the third position after Malta and China in terms of the number of blockchain companies. According to a survey, 51% of companies that participated are currently investing in blockchain technology [8]. More than 400 blockchain related companies are operating in Canada, with concentrations in Ontario (52%); British Columbia (29%); Quebec (9%); Alberta (8%); and Nova Scotia, Newfoundland and Labrador, and Prince Edward Island (2%).

Blockchain technology seems attractive in Canada because of the lower energy price and high-speed Internet in Canada. Thus, the electricity for computers that are needed for mining will be provided at a more reasonable price. Besides, the cold temperature in most parts of Canada allows maintaining the heat as a result of over usage of energy. Besides, Canadian government supports this approach by setting supportive regulations and tax incentives for blockchain companies [8].

Quebec is recognized as the main mining center of Canada, and the Central Bank of Canada first interacted with payments Canada to launch a blockchain project back in 2017. Based on statistics on the blockchain companies in different sectors in Canada [9], it can be concluded that companies are more interested in the financial sector and blockchain consultancy. Sectors such as mass media, supply chain management, infrastructure, and cloud data are given equal importance. Besides, sectors such as energy, law firms, and government are still in the initial stages and are expected to grow in upcoming years.

3 Blockchain in E-Businesses

There are various applications of e-business that can benefit from blockchain technology to be developed in terms of effectiveness and efficiency. One of the most critical success factors for e-business is the supply chain that has been significantly influenced by blockchain [10].

Blockchain technology can be used in supply chain management to deliver products quickly through smart contracts that decrease stop points of products [11], increase coordination among parties by making the verification of both parties essential for change, help to better decision-making with the help of distributed ledger technology, reduce errors through avoiding duplication of data or missing data, increase transparency by providing access for both parties, prevent fraud through reducing the risk of hacking in a chain, and save costs by determining the optimum level of orders and eliminating many intermediaries.

Businesses need to monitor and track products through the supply chain; however, it is difficult for e-businesses to keep real-time track of the supply chain. Applications that are commonly used to manage the supply chain are the potential for the risk of hacking and fraudulent practices but blockchain technology reduces the risks

by eliminating the middleman and assuring authenticity. By providing data security, companies present a better sense of security for customers that encourages customers' trust and leads to the growth of business in turn. Many e-businesses keep the list of customers' personal information including their phone number, credit card details, and address. Blockchain technology aids to store this private information in a decentralized manner. To clarify, data is stored in blocks that are connected in a chain with a randomly generated unique identifier. Therefore, the security will be increased remarkably since it is impossible to hack all blocks in a chain. Another key impact of blockchain technology for e-business is transparency. Since businesses cannot simply manipulate the published reviews and information about the business on the Internet because of reliance on chains, it portrays the true identity of a business. In addition, blockchain technology reduces transaction costs for both businesses and customers since businesses can cut out the costs of middlemen [12].

4 Traditional Supply Chain Versus Blockchain Supply Chain

In a traditional supply chain including retailers, suppliers, and banks, the retailer commonly places an order with the supplier, and the supplier asks for funds [13]. In this process, one party may be unaware of the transactions occurring.

On the other hand, in a blockchain supply chain, all three parties are given access to the blockchain and can verify the information. Thus, it promotes transparency and builds trust [14]. There are a lot of Internet-based businesses which employ blockchain technology in their processes such as IBM, Circle, Coin base, etc. [15]. Since 2018, Walmart has been working with IBM to add blockchain technology to their supply chain. In September 2019, Walmart implemented blockchain technology in the veggies department. Earlier it took seven days to track the source of product but after applying the technology, it just took 2.2 s. The process is lengthy because, in the food supply chain, all the entries are added manually. Whereas in blockchain technology, the processes are fully digitalized which allows quick addition of information. It also cuts down many steps in the supply chain by establishing direct control. In 2019, Walmart asked every supplier of green vegetables to upload data in blockchain [16].

5 Future of Blockchain Technology

Considering the growth rate of blockchain technology, it seems that fruitful results will be witnessed as the technology is growing to cover most sectors in the economy and more individuals and businesses will get access to the technology.

Companies like Unilever are investing in blockchain technology to attain a deforestation-free supply chain, reduce carbon emissions, and fight against the global warming issue [17]. Another expected application of blockchain technology in the future will be in charitable organizations to fund people in different parts of the world in case of any emergency need or natural disaster. Government agencies are also expected to use blockchain technology in the future to reduce corruption and bureaucracy through transparency. The upcoming years will be full of innovations of blockchain technology in different sectors ranging from major economic sectors to artificial intelligence and virtual reality.

6 Conclusion

The usage of blockchain technology in e-business provides numerous benefits from its developing stages. It ensures data protection by reducing the risk of hacking, helps to build trust among consumers, and increases effectiveness. Moreover, it helps to save money for both businesses and customers. The blockchain can be applied to many applications of E-business but the financial sector and supply chain management is more highlighted. One considerable key area in e-businesses to apply blockchain technology in supply chain management that can reap businesses to a lot of benefits such as increased coordination, quick delivery time, better decisions, improved transparency, cost-saving, and reduction in errors and frauds. Thus, many companies such as Walmart and Unilever are shifting to a blockchain supply chain from a traditional supply chain. Regarding the rapid growth in blockchain technology, its future looks promising and with positive impacts for almost every sector of the economy.

References

1. S. Daley, 31 blockchain companies paving the way for the future, Built In (2019). Available at: <https://builtin.com/blockchain/blockchain-companies-roundup>. Accessed May 2021
2. M. Nofer, P. Gomber, O. Hinz et al., Blockchain Bus. Inform. Syst. Eng. **59**(3), 183–187 (2017). Available at: <https://link.springer.com/article/10.1007/s12599-017-0467-3>
3. A. Brzozowska, D. Bubel, E-business as a new trend in the economy. *Procedia Comput. Sci.* **65**, 1095–1104 (2015)
4. N. Reiff, Blockchain: one of history's greatest inventions? Investopedia (2021). Available at: <https://www.investopedia.com/tech/blockchain-one-historys-greatest-inventions>. Accessed May 2021
5. B. Marr, A very brief history of blockchain technology everyone should read. *Forbes* (2018). Available at: <https://www.forbes.com/sites/bernardmarr/2018/02/16/a-very-brief-history-of-blockchain-technology-everyone-should-read/?sh=2896f6e97bc4>. Accessed May 2021
6. Statista, Global use cases for blockchain technology 2020 (2020). Available at: <https://www.statista.com/statistics/878732/worldwide-use-cases-blockchain-technology>. Accessed May 2021
7. A. Takyar, Blockchain in payments|blockchain use cases|LeewayHertz (2020). Available at: <https://www.leewayhertz.com/blockchain-in-payments>. Accessed May 2021

8. B. Weinberg, Why Canada is at the forefront of the blockchain sector? OpenLedger Insights (2020). Available at: <https://openledger.info/insights/blockchain-in-canada/>
9. Statista, Focus areas of Canadian blockchain companies 2019 (2020). Available at: <https://www.statista.com/statistics/1070723/canada-blockchain-focus-area-category>. Accessed May 2021
10. H. Treiblmaier, C. Sillaber, The impact of blockchain on e-commerce: a framework for salient research topics. *Electron. Commer. Res. Appl.* **48**, 101054 (2021)
11. M. Denizhenko, Smart contracts for smart choices: blockchain's automation will transform business decisions. Kaspersky.com (2020). Available at: <https://www.kaspersky.com/blog/secure-futures-magazine/permissioned-blockchain/35661>. Accessed May 2021
12. H. Saakian, How blockchain has helped the business of e-commerce. *Asia blockchain review—gateway to blockchain in Asia* (2020). Available at: <https://www.asiablockchainreview.com/how-blockchain-has-helped-the-business-of-e-commerce/>. Accessed May 2021
13. S. Jabbar, H. Lloyd, M. Hammoudeh et al., Blockchain-enabled supply chain: analysis, challenges, and future directions. *Multimedia Syst.* **27**(4), 787–806 (2021)
14. V. Gaur, A. Gaiha, Building a transparent supply chain. *Harvard Bus. Rev.* (2020). Available at: <https://hbr.org/2020/05/building-a-transparent-supply-chain>. Accessed May 2021
15. Blockchains, What companies are using blockchain technology? (2020) Available at: <https://101blockchains.com/companies-using-blockchain-technology>. Accessed May 2021
16. R. Miller, Walmart is betting on the blockchain to improve food safety. *TechCrunch* (2018). Available at: <https://techcrunch.com/2018/09/24/walmart-is-betting-on-the-blockchain-to-improve-food-safety>. Accessed May 2021
17. S. Tran, Unilever to Use Blockchain for Transparency and Traceability to Achieve Deforestation-Free Supply Chain by 2023 (2020). Available at: <https://blockchain.news/news/unilever-blockchain-transparency-traceability-deforestation-supply-chain>. Accessed May 2021

Automatic Audio and Video Summarization Using TextRank Algorithm and Convolutional Neural Networks



Kriti Saini and Mayank Dave

Abstract There is a surge in data generation in last few years due to digitalization of daily activities, ubiquitous use of smartphone-based applications using Internet that include audio, text and video data. Summarization is the process of generating short and useful data from these large amounts of varied data. In this paper, an application is developed to automatically summarize the extensive and wordy individual speaker audios as well as videos into crisp and concise format. We extended the existing techniques of extractive document summarization using natural language processing (NLP) and neural networks to automatic audio and video summarization. The summarization is done by two approaches which are novel in terms of their implementation technique. One uses TextRank algorithm from *gensim*, a powerful Python library used for performing NLP tasks. The other approach uses convolutional neural network (CNN)-based model that works on the word embedding of the sentences stated in the audio and video. We have used the BBC News Summary dataset for building the CNN model. For evaluating the automatically generated summary, we used ROUGE metric. The results illustrate that the NLP-based approach gave better results for both audio and video summarization.

Keywords Extractive summarization · TextRank algorithm · Convolutional neural network (CNN) · Word embedding · ROUGE metric

1 Introduction

In this digital era, multimedia has become primary form of information interchange by replacing the conventional methods after introducing more impactful means of content. The useful and relevant details present in the audio and video lectures can be retrieved by compacting them in an efficient way. The use of machine learning

K. Saini (✉) · M. Dave
National Institute of Technology, Kurukshetra, Haryana, India

M. Dave
e-mail: mdave@nitkr.ac.in

techniques for text summarization can be extended for summarizing these audios and videos, thus eliminating the redundant and unnecessary information from the original ones thereby reducing human efforts in reading, listening and viewing data.

Development of libraries in programming languages such as Python has provided an easy environment for performing NLP tasks like document summarization. An alternate approach implements a convolutional neural network (CNN)-based model that uses Google's pre-trained word2vec mapping to obtain word embedding of the text in documents. The document summarization can be categorized as extractive and abstractive on the basis of output. The extractive technique simply chooses significant sentences from the document and constructs a smaller summary with exact sentences, whereas abstractive technique is a newer approach of summarization that generates a summary containing new sentences while keeping the essential and crux meaning of the document intact. This work outstretches the techniques of extractive single-document summarization to audio and video summarization. There are various metrics for evaluating the accuracy of automatic text summarization systems. We used ROUGE, a collection of metrics which are used for comparing system-generated summary against a group of model or reference summaries. This work is completely different from the existing models for audio and video summarization as it works on a dataset that is primarily used for extractive document summarization, and therefore, we were able to get better results on an unconventional dataset.

The entire paper is categorized as follows: Sect. 2 elaborates the study done and brief description of the related works. Section 3 explains the approaches used to solve the task. Section 4 includes implementation details of the system. Section 5 discusses the results, and Sect. 6 elucidates the conclusion.

2 Related Works

Prominent research and work has been done in the domain of summarization during the last few decades. The research paper [1] emphasized upon the proficiency of CNNs in solving computer vision problems. However, many recent publications have shown that the CNNs are also capable of performing NLP tasks [2] efficiently. There are many NLP algorithms like TextRank, LexRank [3], latent semantic analysis [4] which are used for carrying out the task of document summarization. Supervised learning algorithms have also been used for text summarization that can be found in [5, 6]. Recursive neural networks (RNNs) have been used for sentence ranking in [7] with incredible results. However, this approach has many drawbacks which limit its use. One of the main limitations is the use of hand-crafted features as input to RNN. The usage of hierarchical CNN for generating summary by extracting sentences based on the sentence features can be found in [8].

Much work has already been done in deep learning-based effective automatic speech recognition (ASR) systems with minimal human intervention. This paper [9] presented a speech to text conversion system using neural networks and algorithms to perform feature extraction, text prediction and model improvisation. Different

approaches have been used for carrying out video summarization. One such recent framework works by doing key frame extraction and video skimming [10]. Some researchers solve the problem by considering it a sequence-to-sequence learning problem and built an attentive encoder-decoder networks for video summarization [11].

3 Proposed Approach

In this work, the system-generated summary refers to the summary of the content present in the audio and video in textual form. We used ROUGE-1 and ROUGE-2 as evaluation metrics in order to evaluate the accuracy of the summary generated using NLP and CNN framework. Figure 1 describes the entire framework implemented in this paper.

3.1 Audio and Video to Text

In order to achieve automatic speech recognition, we used ffmpeg module to detect silent parts of video. This enables to split the video into sub-clips, each consisting of one complete sentence. Then, audio is extracted from these clips and converted into text using PyPI and SpeechRecognition packages that work with Google Speech API.

3.2 NLP-Based Text Summarization

Our work uses the TextRank algorithm from the genism package. TextRank runs a graph-based ranking algorithm used for recommendation purposes. It works on

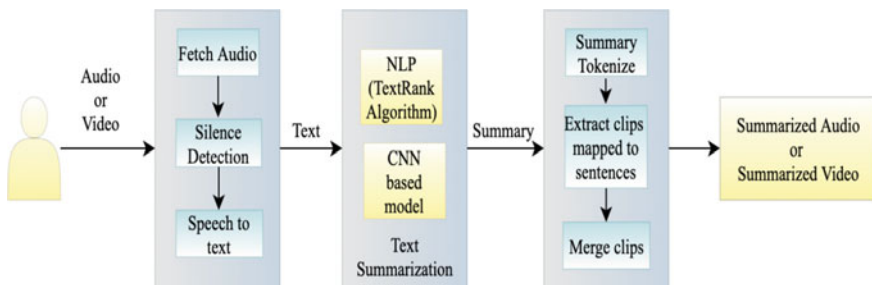


Fig. 1 Workflow of the audio and video summarization application

the ranking of sentences and builds a graph related to the text. In the graph, each sentence is considered as vertex, and each vertex is linked to the other vertex. These vertices cast a vote for another vertex. The importance of each vertex is defined by the number of votes. Therefore, more important sentence has higher ranking. The summary is constructed by including the top-ranked sentences only.

3.3 CNN-Based Text Summarization

In this approach, the sequential model is constructed with convolutional, maxpooling, dropout and fully connected layers. The pickle files containing the salience scores and word embedding of the sentences present in the documents are used as training data. The data in these files are converted into a three-dimensional tensor T where the first dimension represents sentences, second is for words, and third depicts the word embedding.

The model uses one-dimensional convolution layer that learns by applying F filters ($F = 400$) with a window size of n where $n \in N$. Here, n represents the number of consecutive words taken from the text to apply f filters. Hence, the kernel size used in this layer is $(n \times F)$ maintaining strides as 1. The activation function used for this layer is ReLU (R). This function can be applied to get a feature in the manner

$$a_{i,f,j} = R(\{w_f, [s_{i,j}, s_{i,j+1}, \dots, s_{i,j+n-1}]\}_F + b_{f,j}). \quad (1)$$

where w_f is the window with filter f , $s_{i,j}$ is the j th word in i th sentence, and $b_{f,j}$ is the bias term.

We use two-dimensional maxpooling layer with F filters. This helps to down sample and retain only the important features of the data. The maximum value is calculated over the window of size $(F, 1)$. The equation to depict the operation is

$$c_{i,f} = \max((a_{i,f,j})_{j \in [1, l-n+1]}), \quad (2)$$

where l represents the length of longest sentence.

A flatten layer is added to flatten the multi-dimensional tensor output from the previous maxpooling layer to produce a one-dimensional tensor by unstacking all the values in the n -dimensional tensor. Then, a dropout layer is added to prevent overfitting of the model. The masking vector used has the probability $p = 0.5$ as mentioned in [12]. The final layer of this model is a fully connected dense layer with sigmoid as the activation function (S).

The predicted output y'_i is calculated by the salience score of the system-generated summary. We try to estimate this close to the salience score of the human-generated summary. The output is given by the equation

$$y'_i = S(w_v \cdot (c_i \odot v) + b_v). \quad (3)$$

4 Implementation Details

The system summary was generated by TextRank algorithm in NLP-based implementation and using regression process for sentence ranking in CNN architecture.

4.1 *Audio and Video to Text Conversion*

The audio and video were uploaded using Python Pydub and MoviePy libraries, respectively. The Pydub library was used to split audio on silence by detecting pause between two consecutive sentences, whereas for video, the start and end times of silence were stored in a text file with the help of ffmpeg module. Sometimes, the pause detection method does not accurately form precise sentences. This impacts the evaluation score of the entire approach. The start and end timestamps stored for input video were used to obtain sub-clips of the video. From these sub-clips, audio was extracted individually and was stored in the memory as audio chunks. The chunk corresponding to every sentence was then loaded from the memory and uses Python SpeechRecognition library for conversion to text followed by a full stop. All the textual data are then concatenated to form the document and stored as a text file.

4.2 *NLP-based Text Summarization*

In this work, we needed an extractive summary of the document so that we can easily extract only those part of audio or video that matches completely with the sentences. Therefore, we used genism library that already had TextRank algorithm capable for extractive summarization. This is an unsupervised algorithm for summary generation that takes original text as input and gives summarized text as output. Natural language Toolkit (NLTK), a commonly used Python module for NLP is used to tokenize the summarized text. NLTK accurately analyzes the text and divides it into sentences. It makes use of a parser tree to analyze sentences and their structure.

4.3 *CNN-based Text Summarization*

CNN model is capable of efficiently classifying sentences as described in [13]. A simple convolutional neural network architecture that was used contained multiple feature maps for a sentence, and every feature map parallels a convolutional layer succeeded by a maxpooling layer. The detailed implementation is presented in this section.

Data Preprocessing For our preprocessing stage, we split the documents into sentences and obtain their word embedding using word2vec that was trained by using the Google News dataset comprising 100 billion words. This lead to an embedding of 3 million words at $k = 300$ latent dimensions. As a sentence was processed, its salience score was computed using the custom implementation for ROUGE class. The salience scores for each sentence, as well as its word embedding, were stored in pickle file formats. After the text was split into sentences, ROUGE scores were calculated using the pynrouge Python module.

Training As shown in Fig. 2, the pickle files containing the word embedding created in the preprocessing phase were loaded into the n-dimensional array. Some data cleaning operations were performed on the loaded data. The input and the output values were segregated from the loaded data. As the input was present at even indices in the array and output at odd indices, slicing was carried out. The dimensions of the input data were adjusted before converting into a tensor with a shape required by the convolutional layer of the model.

Testing To know the efficacy of the trained model, we tested it using 20% of the dataset. The text documents were embedded using word2vec and given as input to the model. The summary was constructed by iteratively adding only those sentences that were least similar to the sentences already present in the summary until the desired length summary was created. ROUGE-1 and ROUGE-2 scores of the predicted summary were calculated corresponding to the human-generated summary. The full testing pipeline is illustrated in Fig. 2.

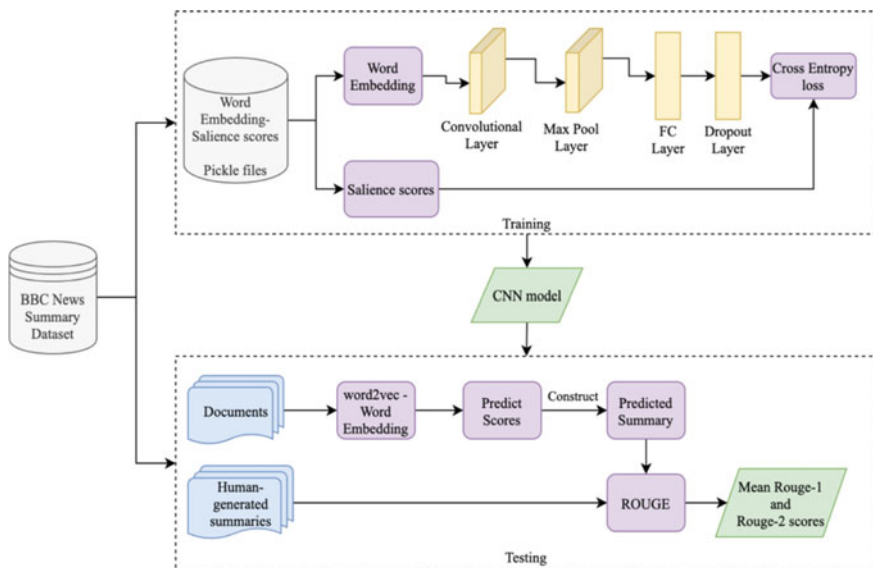


Fig. 2 Full pipeline for CNN-based implementation

4.4 Extract and Merge Audio and Video

The audio chunk whose content exactly matches with the sentence in the summary is fetched and merged to form the compact summarized audio. For building the summarized video, the start and end timestamps of the sentences that make up the summary were used to extract the sub-clips and were merged.

5 Results

In this paper, we have used BBC News Summary dataset [14, 15] for training the CNN model. Table 1 presents the comparison of ROUGE-N scores of our system with other systems based on DUC 2002 dataset for single-document summarization [12]. The ROUGE-N scores for our implementation are listed in Table 2. These are the performance results we achieved after running our system on a particular audio and video. It was observed that NLP-based implementation has out shown the CNN-based implementation in terms of audio and video summarization. The results of CNN-based implementation can be improved by adjusting the model hyper-parameters. The overall results are impacted by the efficiency of the audio and video to text conversion algorithm. The translated text depends on the clarity, fluency and how precisely words are spoken in the sentence. Also, formation of a complete meaningful sentence depends on the accuracy of silence detection libraries. All these factors contribute in determining the accuracy of system-generated summary.

Table 1 ROUGE scores comparison for single-document summarization

System	Dataset	ROUGE-1	ROUGE-2
Cssna.v2	DUC 2002	48.05	22.83
Wpdv-xtr.v1	DUC 2002	47.75	22.27
ULeth 131m	DUC 2002	46.51	20.39
Kul.2002	DUC 2002	46.38	21.25
Ntt.duc02	DUC 2002	46.02	21.27
Context-based	DUC 2002	46.43	20.70
CNN-based	BBC news summary	51.27	38.71

Table 2 Performance results on a particular audio and video

Summarization	Approach	ROUGE-1	ROUGE-2
Audio	NLP-based	0.5801	0.4031
	CNN-based	0.5761	0.3846
Video	NLP-based	0.6106	0.4031
	CNN-based	0.5424	0.4000

6 Conclusion

In this paper, we accomplished audio and video summarization by extending the work done in single-document summarization. The implementation of the work was done following two different approaches: NLP-based and CNN-based summarization. The former approach took comparatively lesser development time than the latter. For training the CNN model, BBC News Summary dataset consisting of news articles from various domains was utilized. Textual data with extractive summaries about any discipline could also have served as relevant dataset. This work used ROUGE-N scores as the evaluation metric to have the same baseline for comparison with the other systems. On comparing the accuracy of system-generated summary with the human-made summary for a particular audio and video, the NLP-based implementation achieved better performance results.

For future steps, the work can be improved by designing more optimized deep neural network structure. Also, many other Python summarization libraries can be explored for generating summary. This paper covers the extractive summarization domain only, so work can be extended for abstractive summarization of documents too. This will enable generating abstractive audio and video summaries which will be more meaningful than the extractive summaries.

Acknowledgements We would like to express our sincere thanks to Yogita Sangwan for helpful discussions on text summarization techniques. We also want to extend our appreciation to Rajesh Swami for discussions on evaluation metric.

References

1. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, in *Proceedings of the IEEE*, vol. 86, no. 11 (1998), pp. 2278–2324
2. R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu P. Kuksa, Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **12**, 2493–2537 (2011)
3. G. Erkan, D.R. Radev, Lexrank: graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.* 457–479 (2004)
4. Y. Gong, X. Liu, Generic text summarization using relevance measure and latent semantic analysis, in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (ACM, 2001), pp. 19–25
5. C. Li, X. Qian, Y. Liu, Using supervised bigram-based ilp for extractive summarization, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (ACL, 2013), pp. 1004–1013
6. Y. Zhang, M.J. Er, R. Zhao, Multi-document extractive summarization using window-based sentence representation, in *2015 IEEE Symposium Series on Computational Intelligence* (IEEE, 2015), pp. 404–410
7. R. Nallapati, F. Zhai, B. Zhou, SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents, in *The Thirty-First AAAI Conference on Artificial Intelligence* (2016)

8. M. Denil, A. Demiraj, N. Kalchbrenner, P. Blunsom, N. de Freitas, Modelling, visualising and summarising documents with a single convolutional neural network. In: arXiv preprint [arXiv:1406.3830](https://arxiv.org/abs/1406.3830) (2014)
9. A. Pardha Saradhi, A. Sai Kiran, A. Dileep Kumar, B. Srinivas, M.V. Nageswara Rao, Design and implementation of speech to text conversion on raspberry Pi, in *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, Issue-6. ISSN: 2278-3075, April 2019
10. S. Jadon, M. Jasim, Unsupervised video summarization framework using keyframe extraction and video skimming. Xiv preprint. [arXiv:1910.04792v2](https://arxiv.org/abs/1910.04792v2) (2020)
11. Z. Ji, K. Xiong, Y. Pang, X. Li, Video summarization with attention-based encoder-decoder networks. arXiv preprint [arXiv:1708.09545v2](https://arxiv.org/abs/1708.09545v2) (2018)
12. Y. Zhang et al., Extractive document summarization based on convolutional neural networks, in *IECON 2016—42nd Annual Conference of the IEEE Industrial Electronics Society* (2016), pp. 918–922
13. Y. Kim, Convolutional neural networks for sentence classification. arXiv preprint [arXiv:1408.5882](https://arxiv.org/abs/1408.5882) (2014)
14. P. Sharif, BBC News Summary, kaggle.com (2021). <https://www.kaggle.com/pariza/bbc-news-summary>. Accessed 26 Feb 2021
15. D. Greene, P. Cunningham, Practical solutions to the problem of diagonal dominance in kernel document clustering, in *Proceedings of the ICML 2006* (2006)

Dealing with Class Imbalance in Sentiment Analysis Using Deep Learning and SMOTE



Shweta Kedas, Arun Kumar, and Puneet Kumar Jain

Abstract In textual data, sentiments or opinions expressing polarities (positive or negative) often form the basis for human decision-making. Therefore, sentiment analysis has always been an important area of research in the field of artificial intelligence. Recently, deep learning models have been used for the sentiment analysis. However, the class imbalance of the dataset adversely affects the performance of these models. To address this issue, the paper presents a method to resample the dataset using synthetic minority over-sampling technique (SMOTE). The proposed method is applied to three different datasets of customer reviews, each of which exhibits different class imbalance ratios. To show the impact of the method, the modern recurrent neural network (RNN)-based architectures (LSTM, GRU, and Bi-directional) are trained with both the originally imbalanced datasets and the SMOTE balanced datasets and comprehensively analyzed the performance of these approaches. The obtained results show that the models trained with balanced datasets outperform other methods across most models with significant improvements over the original dataset.

Keywords Sentiment analysis · Class imbalance · SMOTE · Deep Learning · Recurrent neural network

1 Introduction

The procedure of capturing patterns and understanding people's opinions from data collected through questionnaires had first taken place in the early twentieth century. This work laid the foundation for a branch of computational linguistics called sentiment analysis. In the literature, there are two main approaches for the sentiment

S. Kedas · A. Kumar (✉) · P. K. Jain

Department of Computer Science and Engineering, National Institute of Technology Rourkela, Rourkela, Odisha 769008, India
e-mail: kumararun@nitrkl.ac.in

P. K. Jain

e-mail: jainp@nitrkl.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022
P. Verma et al. (eds.), *Advances in Data Computing, Communication and Security*,
Lecture Notes on Data Engineering and Communications Technologies 106,
https://doi.org/10.1007/978-981-16-8403-6_37

407

analysis including lexicon-based approach and machine learning-based approach [1–5]. Lexicon-based approaches are used to identify the words' polarity, and then frequency-based and location-based approaches have been used to classify the text [6]. Machine learning methods are being used extensively for sentiment analysis due to the availability of adequate dataset and advancement of learning techniques [7–9]. In recent times, deep learning-based machine learning methods have shown to outperform the classical machine learning methods [3, 4]. However, performance of the machine learning models significantly gets affected due to the class imbalance in the dataset [10]. A dataset is said to be imbalanced if imbalance ratio (IR) > 1.5 [11]. IR is the ratio of cardinality of the majority to the minority class. Imbalance in class impacts the accuracy of the analysis method because when a machine learning method will be trained with the imbalance dataset, the trained model will be biased toward the majority class [7]. Therefore, machine learning models have to adopt approaches to deal with class imbalance.

Cost-sensitive learning methods, data-level preprocessing methods, and algorithm-level methods are three major approaches to deal with the class imbalance issue [7, 11–13]. Prustoa et al. [14] used random undersampling (RUS) for Twitter sentiment classification using machine learning techniques. On the other hand, Ah-Pine et al. [6] implemented oversampling techniques SMOTE [15], Borderline-SMOTE, and ADASYN methods on imbalanced Twitter datasets using decision trees and logistic regression. To balance the image dataset, Dablain et al. proposed a deep learning and SMOTE-based approach and termed it as DeepSMOTE [16]. The proposed approach consists of three major components: (i) an encoder/decoder framework, (ii) SMOTE-based oversampling, and (iii) a dedicated loss function that is enhanced with a penalty term.

In the literature, various studies have addressed class imbalance issue; although, impact of these approaches on deep learning approach for sentiment analysis has to be explored. Therefore, the main objective of the presented work is to use the SMOTE algorithm and to analyze the impact of it on the sentiment analysis performed using the deep learning models. The original class imbalanced dataset and dataset balanced using SMOTE is applied to the long short-term memory (LSTM) [17] and gated recurrent units (GRU) [18] models for sentiment classification.

The rest of the paper is structured as follows: Sect. 2 presents the methods and material including the dataset used for experiment, SMOTE algorithm, and deep learning models. Section 3 presents and discusses the results of the experiments. Finally, Sect. 4 concludes the work with a discussion on the potential directions for future research.

2 Methods and Materials

The block diagram of the proposed work is depicted in Fig. 1. Three main components of the block diagram, dataset, SMOTE algorithm, and RNN models, are described in the following subsections.

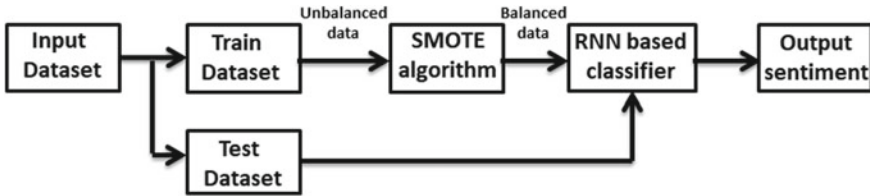
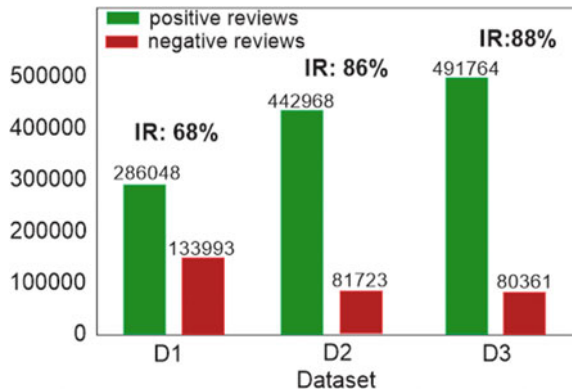


Fig. 1 Block diagram of the SMOTE based sentiment classification

Fig. 2 Sentiment distribution of the reviews for each dataset



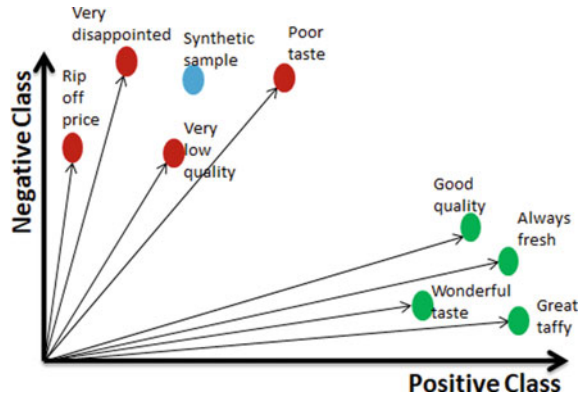
2.1 Dataset

The dataset is collected from Amazon which consists of reviews belong to the three categories—software, fine foods, and appliances and abbreviated as D1, D2, and D3, respectively, in this study [19]. Total number of reviews in D1, D2, and D3 is 459436, 568454, and 602777, respectively. Each review consists of “ProductId,” “UserId,” “ProfileName,” and “overall.” Duplicate entries are removed from the dataset based on the combination of the columns “ProductId,” “UserId,” and “ProfileName.” The column “overall” describes the rating given by the users with 5 being excellent and 1 being bad. Since the work deals with binary sentiment classification, the ratings 1 and 2 are mapped to 0 (negative), 4 and 5 are mapped to 1 (positive), and rating 3 is not considered as it exhibits neither positive nor negative sentiment. Figure 2 depicts the number of positive and negative reviews, and the IR in each dataset post data cleaning.

2.2 SMOTE Algorithm

SMOTE is an oversampling technique used to create synthetic samples of the minority class [15]. It is an iterative approach which considers the *k*-nearest neighbor (default

Fig. 3 Illustration of SMOTE-generated synthetic samples using word vectors belonging to different classes



$k = 5$) samples belonging to the minority class, and uses random interpolation to compute synthetic samples. This algorithm focuses on the feature space and creates synthetic data points as a convex combination of the chosen minority class data points, and its nearest neighbors.

Figure 3 illustrates an example of SMOTE-generated synthetic sample in two-dimensional feature vector space. The easier it is to generate SMOTE-based synthetic samples for a given class, clearer the boundary between the word vectors belonging to positive and negative classes. The time taken for SMOTE algorithm to produce synthetic samples for dataset D1, D2, and D3 is, respectively, 30 min 8 s, 8 min 37 s, and 11 min 20s, respectively. The time taken for dataset D1 is much higher than the other datasets. This might be due to the less imbalance ratio, implying the overlapping of positive and negative class data points in vector space. While in datasets D2 and D3, SMOTE algorithm consumed less time which implies that the positive and negative class samples are clearly differentiated and grouped.

2.3 Deep Learning Models

In this work, three RNN-based models including LSTM, GRU, Bi-directional LSTM (BiLSTM), and Bi-directional GRU (BiGRU) models are implemented using Keras library [20].

Long Short-Term Memory (LSTM) [17] is a type of RNN which consists of input gate, output gate, and forget gate, which is inspired from logic gates to control the information flow within the LSTM units and memory cells.

Gated Recurrent Units (GRU) [18] is similar to LSTM except it consists of only two gates, reset and update gates, and memory cells. The gates of GRU slightly restrict the flow of information by updating to learn necessary information based on the importance of the sequential data and skipping unnecessary information.

Table 1 Hyperparameters of the models

Parametre	Value
Input length	100
Batch size	128
Dropout	0.3
Loss function	Logarithmic loss
Optimizer	RMS Prop

Bi-directional RNNs (Bi-RNN) [21]: The Bi-RNN working technique is similar to the look-back and look-ahead ability of hidden Markov models [22]. The Bi-RNNs consist of two hidden layers, one each for forward and backward passes.

These models are trained with the original imbalanced data and SMOTE balanced data. Before training, the text sequences are cleaned by removing punctuation and HTML tags using regular expressions. Next, the natural language processing techniques: stemming and lemmatization are performed on the text. Then, the Keras Embedding layer is used to encode the input using Keras Tokenizer API. The advantages of using the embedding layer are as follows:

1. The embedding layer derives word embeddings from the input document, and this layer can be saved and used in other models.
2. The embedding layer is flexible and can be utilized to load pre-trained word vectors.
3. Being implemented as a part of the neural network model, the embedding can comprehend from the model itself.

The hyperparameters of the proposed models are provided in Table 1. To measure the proposed models' performance, the logarithmic loss function (binary cross-entropy) [23] is used. The models are optimized using "RMSProp" with an initial learning rate of 0.001, since RMSProp converges faster compared to other optimizers [24]. To overcome the overfitting of the models', dropout and early stopping regularization methods have been implemented [25]. In this work, a dropout layer is added between the RNN layer and a fully connected dense layer with a 0.3 probability. To overcome the issue of overfitting, early stopping [26] regularization method is used that halts the training before the model can overfit. In early stopping, to decide the stopping point for model training, we choose "validation loss" with mode set to "minimum" for early stopping, as lowering the loss value improves the model's performance.

3 Results and Discussion

To evaluate the models' performance, we use the metrics F1-score and area under curve (AUC) of receiver operating characteristic (ROC) curve [12]. The accuracy along with F1-scores and AUC trained on the three datasets are in Table 2. Also, the

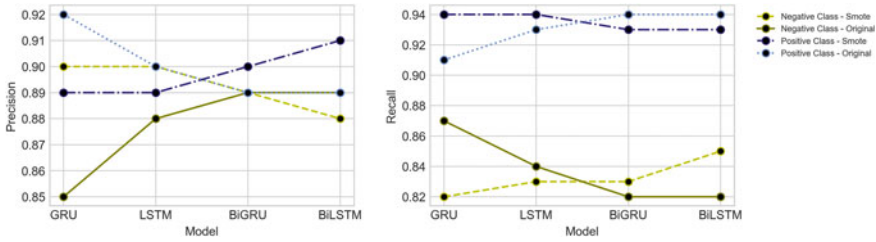


Fig. 4 Precision and recall values of models trained with dataset D1

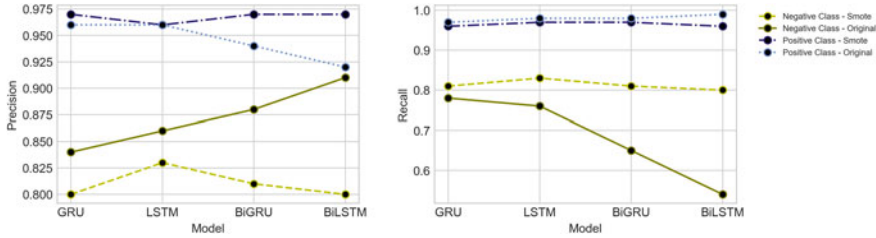


Fig. 5 Precision and recall values of models trained with dataset D2

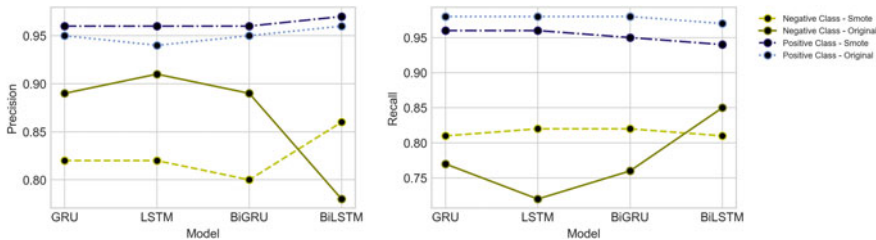


Fig. 6 Precision and recall values of models trained with dataset D3

precision and recall values for each class for every dataset are compared. F1-score, a measure of test accuracy, which gives equal weightage to precision and recall, makes it an appropriate performance metric. ROC metric is only used for binary classification problems. The ROC curve is a plot of sensitivity (Y -axis) and inverted specificity (X -axis) for different threshold values, ranging from 0 to 1. The area under curve summarizes the model's ability to classify. High AUC for a given binary classification model implies that the model can correctly distinguish between the positive and negative classes. Figures 4, 5, and 6 visualize the precision and recall scores of each class for datasets D1, D2, and D3, respectively.

From Table 2, it can be observed that for the dataset D1, F1-score and AUC of all the models are slightly high in case of the SMOTE-generated dataset. Precision and recall values for each class are balanced in case of both original and SMOTE-generated datasets of D1. Overall, it can be concluded that models trained with

Table 2 Results of our models across datasets and data distribution

Data distribution	Model	D1			D2			D3		
		Accuracy (%)	F1 score (%)	AUC	Accuracy (%)	F1 score (%)	AUC	Accuracy (%)	F1 score (%)	AUC
Original	GRU	92.57	89.30	0.9571	96.77	94.11	0.9698	93.41	93.26	0.9694
	LSTM	93.84	89.35	0.9558	95.83	94.44	0.9691	94.43	93.45	0.9711
	BiGRU	93.52	89.13	0.9556	94.69	92.85	0.9612	94.59	92.84	0.9660
	Bi-LSTM	93.54	89.20	0.9562	92.80	91.59	0.9551	96.13	92.79	0.9677
SMOTE	GRU	91.03	89.33	0.9578	96.00	94.60	0.9686	96.27	94.01	0.9743
	LSTM	92.21	89.57	0.9584	97.44	94.42	0.9688	95.85	93.60	0.9730
	BiGRU	93.17	89.39	0.9568	95.87	94.30	0.9685	96.45	93.86	0.9723
	Bi-LSTM	91.93	89.68	0.9585	97.21	94.61	0.9702	95.68	93.96	0.9731

The bold font is signifying the maximum value in the respective column

SMOTE-generated dataset performed better than the original dataset. The precision and recall scores of models trained with dataset D1 are shown in Fig. 4.

The model results for dataset D2 from Table 2 show that all the three metrics; accuracy, F1, and AUC values are slightly high for SMOTE dataset. From Fig. 5, it can be observed that there is a drastic decrease in the negative class recall values trained with the original dataset D2, while the SMOTE dataset of D2 provides balanced results of precision and recall values for all the models. This shows that the high values of F1 scores in the case of models trained with the original dataset of D2 are due to the high values of precision and recall scores of only the majority (positive) class.

Similar, accuracy, F1-score, and AUC of all the four models are slightly high in the SMOTE dataset of D3 than the original dataset. As shown in Fig. 6, precision and recall values of models trained with original dataset D3 are less for all the models in the case of minority class. These precision and recall values are high and balanced for all the models trained with SMOTE dataset of D3.

The similar values of AUC for every model in their respective dataset column in Table 2 correspond to the conclusion that the models trained with both balanced and imbalanced datasets can correctly distinguish the positive and negative class samples. Across the datasets, the lower values of F1-score and AUC for D1 compared to D2 and D3 are because of fewer number data present in D1. Also, as mentioned in Sect. 2.2, the data in D1 is not distinguished, leading to the AUC values around 0.95 while the AUC scores are between 0.96 and 0.97 for D2 and D3.

Among the models and across the datasets, the LSTM models, specifically BiLSTM models, performed comparatively better. For dataset D1, BiLSTM model performs the best among the others with F1-score 89.68% and 0.9585 AUC. BiLSTM model performs the best among the others, even for dataset D2 with an F1 score of 94.61% and 0.9702 AUC. In dataset D3, GRU performs the best with an F1 score of 94.01% and 0.9743 AUC, with BiLSTM being second best with an F1 score of 93.96% and AUC 0.9731. For every model, the SMOTE-balanced datasets have given better results than the original imbalanced datasets in all the above cases.

4 Conclusion and Future Work

In this work, deep learning models for binary sentiment classification have been presented with the focus on dealing with the class imbalance of the datasets. The SMOTE method is applied to observe how the data augmentation of text data can affect the performance of the deep learning models. Deep learning models including LSTM, GRU, BiLSTM, and BiGRU are trained with the original imbalanced and SMOTE-balanced datasets and presented a comprehensive comparison of the results.

Three datasets with different imbalance ratios are considered to analyze the performance of the models with varying numbers of positive and negative class samples. Unlike in the original datasets, the models trained with the oversampled datasets (SMOTE algorithm) showed a good trade-off between precision and recall scores,

with the overall performance comparatively better. The results show that BiLSTMs performed better than other models due to their performance metrics values slightly more significant than that of the other models, across all the three datasets, with balanced and high precision and recall scores for both majority and minority sentiment classes. Thus, it can be concluded that SMOTE increases the models' performance in case of class imbalance.

In future work, the extensions of SMOTE can be experimented. Future work also aim to use pre-trained word embeddings like Glove and Word2Vec, and ensemble deep learning models like convolution neural networks (CNNs) and RNN to find if there can be any significant difference in the models' performances.

References

1. Y.M. Aye, S.S. Aung, Sentiment analysis for reviews of restaurants in Myanmar text, in *18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)* (2017), pp. 321–326
2. J. Barry, Sentiment analysis of online reviews using bag-of-words and LSTM approaches, in *25th Irish Conference on Artificial Intelligence and Cognitive Science* (2017), pp. 272–274
3. H. Feng, R. Lin, Sentiment classification of food reviews (2016). <https://arxiv.org/abs/1609.01933>
4. M. Heikal, M. Torki, N. El-Makky, Sentiment analysis of arabic tweets using deep learning. *Proc. Comput. Sci.* **142**, 114–122 (2018)
5. M.V. Mäntylä, D. Graziotin, M. Kuutila, The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. *Comput. Sci. Rev.* **27**, 16–32 (2018)
6. J. Ah-Pine, E. Soriano-Morales, A study of synthetic oversampling for Twitter imbalanced sentiment analysis. *DMNLP@PKDD/ECML* **1646**, 17–24 (2016)
7. A. Fernández, S. García, M. Galar, R. Prati, B. Krawczyk, F. Herrera, *Learning from Imbalanced Data Sets* (Springer, 2018)
8. B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing* (2002), pp. 79–86
9. P. Turney, Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *Comput. Res. Reposit.* 417–424 (2002)
10. J. Johnson, T. Khoshgoftaar, Survey on deep learning with class imbalance. *J. Big Data* **6**, 27 (2019)
11. M. Lango, Tackling the problem of class imbalance in multi-class sentiment classification: an experimental study. *Found. Comput. Decis. Sci.* **44**(2), 151–178 (2019)
12. J. Brownlee, Imbalanced classification with python: better metrics, balance skewed classes, cost-sensitive learning, in *Machine Learning Mastery* (2020). <https://books.google.be/books?id=jaXJDwAAQBAJ>
13. B. Krawczyk, Learning from imbalanced data: open challenges and future directions. *Prog. Artif. Intell.* **5**, 221–232 (2016). <https://doi.org/10.1007/s13748-016-0094-0>
14. J. Prusa, T.M. Khoshgoftaar, D.J. Dittman, A. Napolitano, Using random undersampling to alleviate class imbalance on tweet sentiment data, in *16th IEEE International Conference on Information Reuse and Integration*, pp. 197–202 (2015)
15. N. Chawla, K. Bowyer, L. Hall, W.P. Kegelmeyer (2002) SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res. (JAIR)* **16**, 321–357
16. D. Dablain, N. Krawczyk, N.V. Chawla, DeepSMOTE: fusing deep learning and SMOTE for imbalanced data (2021). [arXiv:2105.02340v1](https://arxiv.org/abs/2105.02340v1)

17. S. Hochreiter, J. Schmidhuber, Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
18. K. Cho, B. Van Merriënboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: encoder-decoder approaches, in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation* (2014), pp. 103–111
19. J. Ni, J. Li, J. McAuley, Justifying recommendations using distantly-labeled reviews and fine-grained aspects, in *EMNLP*, pp. 188–197 (2019)
20. F. Chollet, Keras (2015). <https://keras.io>
21. M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks. *IEEE Trans. Sig. Process.* **45**(11), 2673–2681 (1997)
22. A. Zhang, Z.C. Lipton, M. Li, A.J. Smola, Dive into Deep Learning (2020). <https://d2l.ai>
23. Machine Learning Glossary (2017). https://ml-cheatsheet.readthedocs.io/en/latest/loss_functions.html
24. T. Tieleman, G. Hinton (2012). Lecture 6.5—rmsprop: divide the gradient by a running average of its recent magnitude
25. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014)
26. L. Prechelt, Early Stopping—But When? *Neural Netw. Tricks Trade* 55–69 (1998)

Classification of Machine Learning Algorithms



Hamed Taherdoost

Abstract Machine learning (ML) is to make logical patterns out of various types of input data including images, texts, numbers and any other types of data. Data derived from research will be processed through machine learning algorithms and leads to a prediction that is mainly considered as the output of the machine learning algorithm. In this paper, the most popular learning algorithms have been reviewed and their specific features are discussed to help select the most appropriate algorithm through comparison in different research projects. However, there is not just one efficient method to apply to all data sets, and the appropriate algorithm may differ based on factors in a study.

Keywords Machine learning · Machine learning algorithms · Classification · Supervised algorithms · Unsupervised learning · Semi-Supervised learning

1 Introduction

Human has developed machines to accomplish tasks that are not easily processed by human brain. Intelligent tools and machines help businesses to be more productive. The initial concept of intelligent machines was shaped in the mid-twentieth century when Alan Mathison Turing first thought about the possibility of employing machines to process data. Then, artificial intelligence that is considered as a branch of computer science was developed at a rapid pace. Machine learning that its learning capacity is not dependent on programming was initially based on the simulation of human intelligence. Today, ML that is generally defined as learning from different levels and classes of data [1] through employing different algorithms to make accurate predictions [2] is broadly used in various fields aiming to solve problems that are mainly based on big datasets. ML can solve complex projects through processing complicated and big data inputs to predict potential threats and profitable opportunities for businesses that are widely reliant on data to proceed. This review discusses and

H. Taherdoost (✉)
University Canada West, Vancouver, Canada
e-mail: hamed.taherdoost@ucanwest.ca

compares popular and commonly used supervised learning, unsupervised learning, and semi-supervised learning algorithms.

2 Machine Learning

Since ML employs different types of algorithms based on AI to proceed with the learning phase from data [3], it is mainly recognized as one of the subsets of AI. However, they are used interchangeably in many cases. A more developed viewpoint to ML was revealed when Frank Rosenblatt used his inspirations from the human nervous system called “perception” to create a procedure for alphabet recognition. Then, various ML algorithms were developed to solve different problems since the results of employing ML algorithms are highly reliable to make critical decisions in different industries [4]. Despite many developments that have been witnessed to employ ML algorithms for real-life issues and challenges, there are also some difficulties in solving real problems through algorithms. This review presents commonly used machine learning algorithms that make them more accurate, and reliable to use.

3 Machine Learning Algorithms

In a general classification, machine learning algorithms are divided into supervised, unsupervised, and semi-supervised algorithms. However, more detailed classifications are also provided based on different types of learning.

3.1 *Supervised Algorithms*

Supervised learning makes predictions based on learning from labeled data through observing variables. Supervised machine learning algorithms work with the help of two data classes including train data and test data and are commonly used for accurate predictions ranging from speech recognition, spam detection, bioinformatics to database marketing purposes. Output of the training data leads to accurate pattern recognition and will be applied for classification and prediction [5]; however, although its computation is simple, it may be cost and time-consuming to determine datasets.

Popular supervised learning algorithms that will be discussed in this section are decision tree, Naïve Bayes, support vector machine (SVM), logistic regression, and random forest. All these algorithms have their specific advantages and deficiencies, and should be selected based on specific conditions of the problem.

Decision tree is recognized because of alternating results that it provides from different decision series, and is mostly employed for planning and defining strategies.

Decision tree is a simple tree-like model that is made of branches and nodes to process input values and make different decisions [5]. Decision tree has demonstrated to work effectively in predicting accurate and reliable student performance, predicting the production of petroleum [6], and segmenting credit status in banking sector [7]. Besides, it is widely used because of being simple to use, quick to run, and clear to understand; however, it may show lower accuracy when there is too much data [8].

Naïve Bayes that was initially developed based on the Bayes theorem by Thomas Baye [9] calculates probability of each attribute independent from others through the probability of each class with the highest probability [10]. NB has been effectively used in clustering and classifying big data on the basis of conditional probability. One of the main functions of Naïve Bayes is text classification in industry.

Support vector machine (SVM) is another supervised learning algorithm that is mainly based on the classifying datasets, calculating margins, and minimizing the classification error. SVM maximizes the difference between margin and data classes through reliance on hyperplane. SVM has demonstrated the highest accuracy in predicting oil usage. Besides, since it is an effective algorithm to analyze image data, it has demonstrated 92.92% accuracy rate in classifying sugar beets based on image data.

Logistic regression that analyzes the relation between a single or multiple predictors is the most appropriate algorithm in case that data needs multiple classification [11]. This algorithm that works based on the chance of occurrence is the best choice when dependent variable is binary. The likelihood of a certain class or event will be determined using either 0 or 1 values through data processing [10].

Random forest is another supervised learning algorithm that is made of combining prediction trees [12] in which there is equal distribution in all trees, and trees are based on a vector selected randomly. Random forest is not the best choice for overfitting data [8] and more trees in numbers can lead to induction error coverage.

3.2 Unsupervised Learning

Unsupervised learning algorithm is not excessively dependent on historical data to make predictions. Instead, it refers to previous extracted patterns from data to feed the new dataset. Unsupervised learning algorithms are mainly employed for clustering purposes and reducing features in a dataset. Human error is minimized considerably in unsupervised learning. Popular supervised learning algorithms that will be discussed in this section are K-means clustering and principal component analysis.

K-means clustering that can organize different clusters from large amounts of data and subsequently placing them in related groups is an unsupervised learning algorithm that was developed by two scientists from Yale University [13]. Different clusters will be detected and formed through this method, and the average value will be set as the core of the cluster [14].

Principal component analysis (PCA) is another popular unsupervised learning algorithm that was developed in 1901 [15]. PCA seeks to reduce the size of dataset aiming to alleviate the process of computing. For example, it turns a 2D dataset to 1D to facilitate computing data.

3.3 *Semi-Supervised Learning*

Combining the features of supervised and unsupervised learning has led to semi-supervised learning that can work in case of having unlabeled datasets [16]. Semi-supervised learning is easier to understand, stable, and highly efficient; however, its results may not be fully accurate, and stable.

Popular semi-supervised learning that is discussed in this section are generative model, self-training, and transductive support vector machine (TSVM).

Generative model is a popular and applicable method that is frequently used to recognize mixed components in an unlabeled dataset.

Self-training is also a semi-supervised learning algorithm which classifies dataset using a portion of labeled data. Unlabeled datasets will be fed to the system and a blend of labeled data and prediction will be achieved. Since the process of classification will be repeated automatically in times, it is called self-training method.

Another semi-supervised learning algorithm is transductive support vector machine (TSVM) that is mainly an extended and modified form of support vector machine [17]. This method considers both labeled and unlabeled data to process and maximize their margin value. Despite advantages of using this method to analyze both labeled and unlabeled data, it has demonstrated deficiencies in solving NP-hard problems.

4 Discussion and Conclusion

Machine learning algorithms are employed to find an appropriate pattern using the dataset and eventually leading to an output result. ML algorithms differ in features and applications based on the input that they can process and the method of analysis, each algorithm comes with its specific advantages, disadvantages to employ.

The requirement of deciding among existing algorithms is having enough information about datasets and variables. Naive Bayes performs more efficiently in case that input data is independent and large. SVM is a promising model to employ when the size of the data is medium and it is not very complicated. Linear regression, logistic regression, and SVM are appropriate to use when the relationship between dependent and independent variables is linear. On the other hand, in case that the relationship between the dependent and independent variable is unclear, and the dataset is small in size, k-NN seems to be a smarter choice.

Besides, accuracy of the output, processing time, complexity of the model, and nature of data are also detrimental factors in selecting an algorithm. Logistic regression results in clear and easy to understand outputs, and thus, is applicable in case that data is simple, small, and uncomplicated. SVM is also a very accurate model since it handles overfitting data; however, it may be time-consuming in processing when data is large. Naive Bayes is ideal in case that computing probability is simple but it is highly reliant on independent variables. Decision tree is easy to understand and handling missing value; however, it is sensitive to data that needs more training time. Random forest is highly efficient for imbalanced datasets and applicable for missing data but it works based on uncorrelated predictions.

Comparing specific features and applications of different ML algorithms leads to the fact that there is no definite algorithm to solve a problem, and the solution may vary based on determinative factors. Therefore, each algorithm has its advantages and disadvantages to consider based on the topic of each project. Thus, there is still plenty of future research opportunities for scientists and scholars to seize in the machine learning science.

References

1. S. Amornsamankul, B. Pimpunchat, W. Triampo et al., A comparison of machine learning algorithms and their applications. *Int. J. Simul. Syst. Sci. Technol.* **8**, 1–17 (2019)
2. K. Das, R.N. Behera, A survey on machine learning: concept, algorithms and applications. *Int. J. Innov. Res. Comput. Commun. Eng.* **5**(2), 1301–1309 (2017)
3. V. Sze, Y. Chen, T. Yang et al., Efficient processing of deep neural networks: a tutorial and survey. *Proceed. IEEE* **105**(12), 2295–2329 (2017)
4. M. Sokolova, G. Lapalme, Performance measures in classification of human communications. *Advances in Artificial Intelligence* (Springer, Berlin, 2007)
5. S. Kotsiantis, Supervised machine learning: a review of classification techniques. *Informatica (Slovenia)* **31**, 249–268 (2007)
6. X. Li, C.W. Chan, H.H. Nguyen, Application of the neural decision tree approach for prediction of petroleum production. *J. Petrol. Sci. Eng.* **104**, 11–16 (2013)
7. F. Butaru, Q. Chen, B. Clark et al., Risk and risk management in the credit card industry. *J. Bank. Finance* **72**, 218–239 (2016)
8. Z. Omary, F. Mtenzi, Machine learning approach to identifying the dataset threshold for the performance estimators in supervised learning. *Int. J. Infonom.* **3**(3), 314–325 (2010)
9. S. Raschka, *Naive Bayes and Text Classification I Introduction and Theory* (2014). Available at: <https://arxiv.org/pdf/1410.5329>. Accessed Aug 2021
10. M.C. Belavagi, B. Muniyal, Performance evaluation of supervised machine learning algorithms for intrusion detection. *Procedia Comput. Sci.* **89**, 117–123 (2016)
11. P. Reed, Y. Wu, Logistic regression for risk factor modelling in stuttering research. *J. Fluency Disord.* **38**(2), 88–101 (2013)
12. L. Breiman, Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)
13. J.A. Hartigan, M.A. Wong, Algorithm AS 136: A K-means clustering algorithm. *J. Royal Stat. Soc. Series C (Appl. Stat.)* **28**(1), 100–108 (1979)
14. S. Shalev-Shwartz, Y. Singer, N. Srebro et al., Pegasos: primal estimated sub-gradient solver for SVM. *Math. Program.* **127**(1), 3–30 (2010)
15. K. Pearson, LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **2**(11), 559–572 (1901)

16. X. Zhu, A.B. Goldberg, *Introduction to Semi-Supervised Learning* (Morgan & Claypool Publishers, 2009)
17. N. Kasabov, S. Pang, Transductive support vector machines and applications in bioinformatics for promoter recognition, in *Proceedings of the 2003 International Conference on Neural Networks and Signal Processing*, vol. 1 (2004), pp. 1–6

Performance Analysis of Object Classification System for Traffic Objects Using Various SVM Kernels



Madhura M. Bhosale, Tanuja Satish Dhope (Shendkar), Akshay P. Velapure, and Dina Simunic

Abstract In the area of autonomous vehicle, object classification is an important task. Different classification algorithms are available. In this work, we have focused on different kernels of support vector machine (SVM) to analyze the performance of traffic object classification system. We have performed experimentation on open-source database had been carried out on the basis of performance metrics such as recall, precision, F1, and accuracy. Our classification system provides maximum accuracy 69% with the help of RBF kernel.

Keywords Linear SVM · Machine learning · Classifier

1 Introduction

Object classification plays important role in the field of autonomous vehicle. Different classification algorithms are present in market such as SVM, KNN, random forest, and decision tree. In this paper, we have focused on performance of different kernels of SVM for the classification system of traffic objects. Non-parametric supervised learning model is a support vector machine (SVM) [1]. Figure 1 shows the decision function for a linearly separable problem, with three samples on the margin boundaries, called ‘support vectors’. SVM is used for outlier’s detection, regression, and

M. M. Bhosale (✉)

Department of Electronics and Telecommunication Engineering, JSPM’s Rajarshi Shahu College of Engineering, Pune 411033, India

Tanuja Satish Dhope (Shendkar)

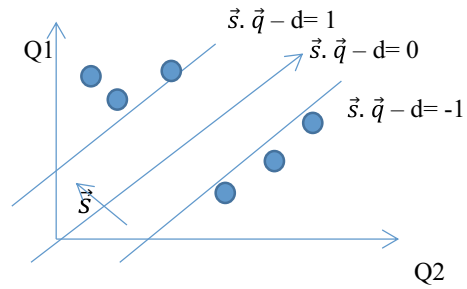
Professor, Department of Electronics and Communication, College of Engineering, Bharati Vidyapeeth (Deemed to be University), Dhankawadi, Pune 411043, Maharashtra, India
e-mail: tsdhope@bvuoep.edu.in

A. P. Velapure

Pune, India

D. Simunic

Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia
e-mail: Dina.Simunic@fer.hr

Fig. 1 Graph of SVM

classification. Multiple categorical data are handled with SVM. Separation of two classes based on feature set is the main aim of SVM. As graph shown in Figure 1 have three lines, one is marginal $\vec{s} \cdot \vec{q} - d = 0$, and other two lines $\vec{s} \cdot \vec{q} - d = 1$ and $\vec{s} \cdot \vec{q} - d = -1$ indicate the position of closest data points of both the classes.

2 Literature Review

There is a huge research has been going on object classification as specified in [2–5]. This paper investigates the application of support vector machines (SVMs) in texture classification [6]. In [7], it gives you idea about multiclass classification using SVM. In [8], it gives idea about non-linear kernel of the SVM and provides you survey about various SVM concept and some real-time applications using SVM. In [9], paper reviews various concepts of random forest and SVM for remote sensing image classification and provides you comparative analysis using different parameters. In [10], SVM with k -fold algorithm is used to predict and evaluate degradation of concrete strength in a complicated marine environment. After some experimentation, average relative error is reduced from 34.8 to 27.6%, and median-relative error declines from 24.7 to 20.8%. In [11], it introduces least square support vector machine for regression into reliability analysis to overcome the shortcomings present in support vector regression. In [10], the influence of vehicle, traffic, people, and traffic management on driverless vehicle and constructs a highway safety evaluation model based on support vector machine. In [12], this paper presents performance of SVM for recognizing road images. Images divided into four classes turn left, right, forward, stop. Accuracy of this model is 70.77%. In [13–19], it compares independent component analysis, Fourier transform principal component analysis, and independent component analysis for data preprocessing with the help of machine learning method of SVM.

3 Methodology

The proposed methodology, overall algorithm has been splinted into two parts; one is training, and other is testing.

1. **Model Training:** Overall training model has been splinted into two steps. As shown in Figure 2, a. first step is preprocessing and feature extraction of images [20]. First, we resize the image and convert RGB to gray. We have used histogram of gradient HOG [20] features. The feature set was used to train machine learning models using different kernels of SVM.

A SVM Parameters:

- (i) **C Parameter:** This parameter is regularization parameter. The C parameter is trade off correct classification of training examples against maximization of decision function margin. Classifying all training data points correctly there need to be smaller margin accepted which is only possible for the larger values of C [21].
- (ii) **Gamma Parameter:** the gamma parameter defines how far the influence of a single training example approaches, high values meaning ‘close’ and with low values meaning ‘far’.
- (iii) **Kernel:** Selection of the kernel is totally depends on whether your data points distribution. Figure 3 gives you clear idea about types of the kernels.

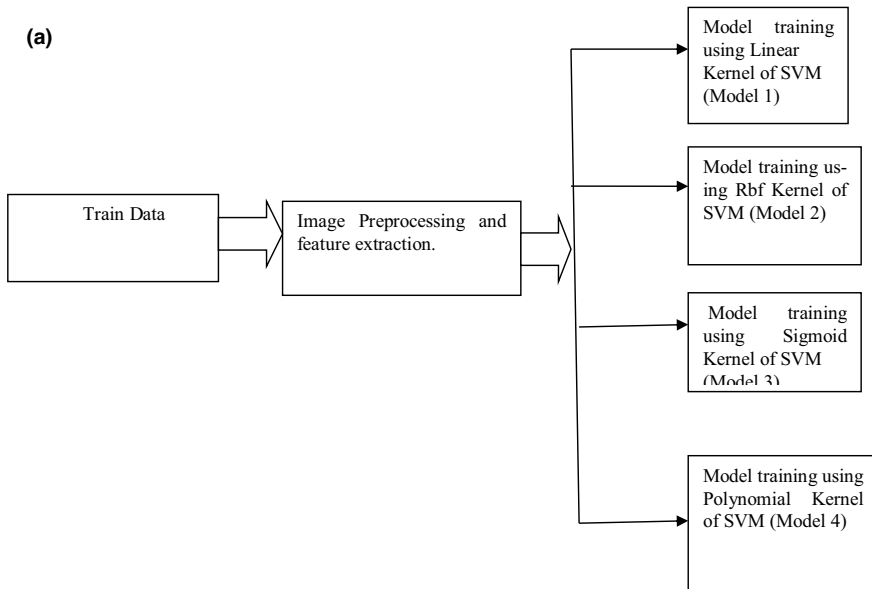


Fig. 2 a Block diagram of training of machine learning model using different SVM. **b** Block diagram of testing system kernels

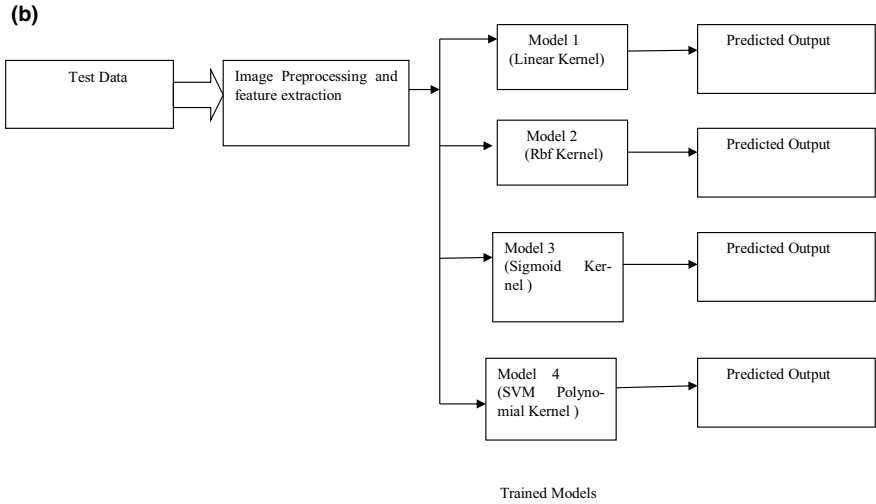


Fig. 2 (continued)

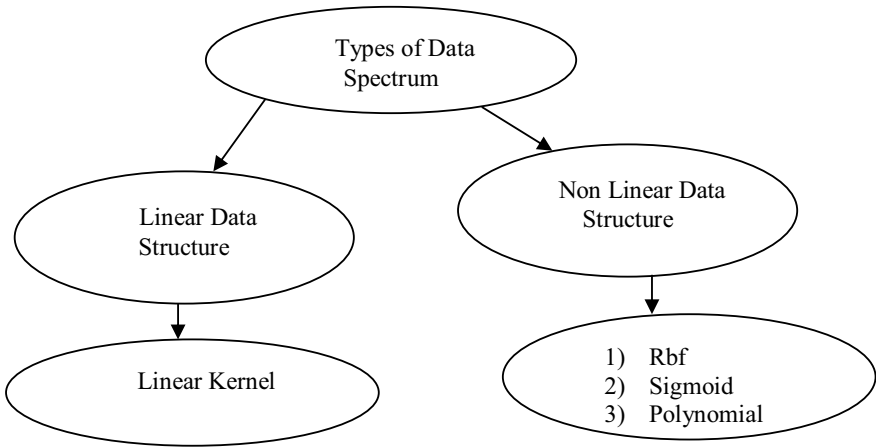


Fig. 3 Types of kernels

(a) **Linear kernel:** In linear classifier as shown in Fig. 1, n dimensional points are separated with $(n - 1)$ dimensional hyperplane [22]. The best hyperplane is one which separate two classes with maximum marginal value. Suppose we have n number of points $(\vec{q}_1, y_1) \dots (\vec{q}_n, y_n)$. Here, $y(j)$ is 1 or -1 representing the class of point $q(j)$. Now, we are interested in exploring maximum margin hyperplane separating class of point $q(j)$ having $y(j) = 1$ from class of $y(j) = -1$. Distance between hyperplanes and nearest $q(j)$ point should be maximum. Equation of the hyperplane is as follows

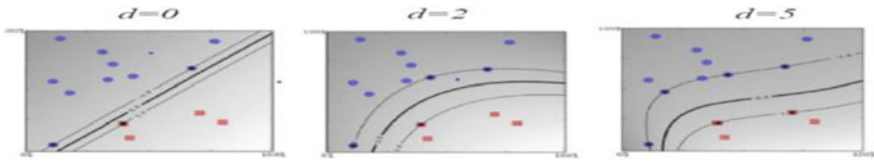


Fig. 4 Response of the kernel functions on different degrees [8]

$$\vec{s} \cdot \vec{q} - d = 0 \tag{1}$$

where \vec{s} = normal vector to the hyperplane, $d/\|\vec{q}\|$ = offset of hyperplanes from origin along \vec{s} .

- (b) **Polynomial kernel:** SVM represents the similarity of the training samples in the feature space over polynomial of the original variables [23]. For the degree d, polynomial function is represented by following equation (Fig. 4):

$$K(x, y) = (xy + C) \tag{2}$$

- (c) **Radial Basis Function:** This function is used Euclidean distance. In SVM, radial basis function is used to define the Gaussian radial basis function [22, 24]. Euclidean distance is calculated by following equation.

$$\Phi(x, c) = \Phi(\|x - c\|) \tag{3}$$

In radial basis function, there is free parameter σ present which help to calculate Euclidean distance between two landmarks. Figure 5 shows effect of σ on Rbf kernel.

This is most popular kernel method which is given by formula.

$$K(x, x') = \exp(-\gamma \|x - x'\|^2) \tag{4}$$

where γ given by,

$$\gamma = 1/2\sigma^2 \tag{5}$$

- (d) **Sigmoid kernel:** When the numbers of features are too large and non-linear, then sigmoid kernel is used. This kernel is basically inspired by neural network.

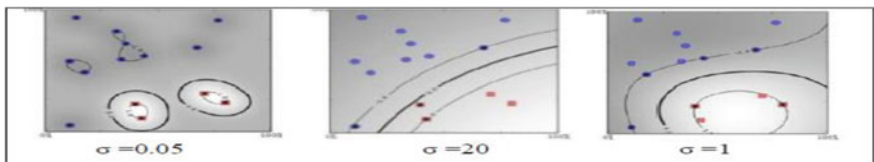


Fig. 5 Variation in the Gaussian RBF kernel with variation in σ [8]

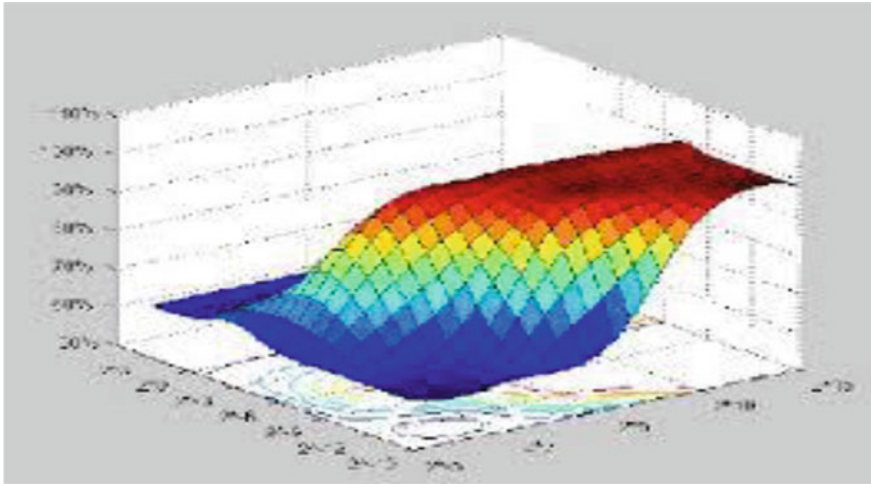


Fig. 6 Sigmoid function plot using C-SVM [8]

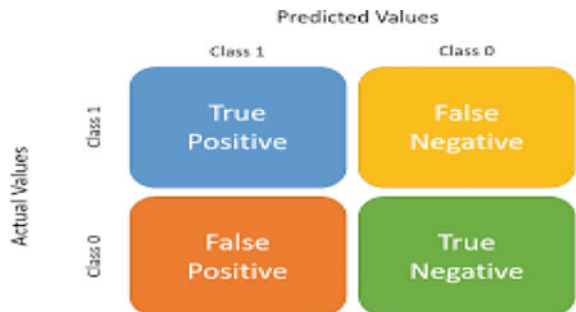
Sigmoid kernel function is given by (Fig. 6),

$$K(x, y) = \tanh(\alpha x^T y + c) \tag{6}$$

2. Second part of algorithm is testing. In this, we do same operation on testing data which we already performed on training data. And provide that data as an input to trained model and predict the output as shown in Fig. 2b.

B **Confusion matrix:** The confusion matrix is used to extract more information about model performance. The confusion matrix helps us visualize whether the model is “confused” in discriminating between the classes. The labels of the two rows and columns are *negative* and *positive* to reflect the two class labels as shown in Fig. 7.

Fig. 7 Confusion matrix [19]



4 Results

For our work, we used open-source database acquired from [25]. The image database consists of two classes of objects truck and car. We implemented this work using Python 3.8 programming language. We extracted total 16,384 HOG features from 782 images. First, we split dataset into two parts, 70% kept for training and 30% used for testing.

Figures 8, 9, 10, and 11 show classification results obtained from various models using linear, polynomial, RBF, and sigmoid kernels of SVM, respectively. Machine learning model using SVM linear kernel true positive, false negative, false positive,

```
[Status] Loaded features of shape (782, 16384)
[Status] Loaded labels of shape (782,)
Enter 1 for linear kernel
Enter 2 for polynomial kernel
Enter 3 for RBF kernel
Enter 4 for sigmoid kernel
Enter your choice: 1
You have selected SVM Linear kernel
Confusion matrix:
[[75 51]
 [44 65]]
Classification report
precision      recall  f1-score   support
   car                0.63      0.60      0.61       126
  truck                0.56      0.60      0.58       109
   accuracy                0.60      0.60      0.60       235
  macro avg                0.60      0.60      0.60       235
 weighted avg                0.60      0.60      0.60       235
```

Fig. 8 Results of linear kernel of SVM

```
[Status] Loaded features of shape (782, 16384)
[Status] Loaded labels of shape (782,)
Enter 1 for linear kernel
Enter 2 for polynomial kernel
Enter 3 for RBF kernel
Enter 4 for sigmoid kernel
Enter your choice: 2
You have selected SVM Polynomial kernel
Confusion matrix:
[[94 32]
 [46 63]]
Classification report
precision      recall  f1-score   support
   car                0.67      0.75      0.71       126
  truck                0.66      0.58      0.62       109
   accuracy                0.67      0.66      0.67       235
  macro avg                0.67      0.66      0.67       235
 weighted avg                0.67      0.67      0.67       235
```

Fig. 9 Results of polynomial kernel of SVM

```
Enter 1 for linear kernel
Enter 2 for polynomial kernel
Enter 3 for RBF kernel
Enter 4 for sigmoid kernel
Enter your choice: 3
You have selected SVM RBF kernel
Confusion matrix:
[[81 45]
 [27 62]]
Classification report
precision      recall  f1-score   support
   car                0.75      0.64      0.69       126
  truck                0.65      0.75      0.69       109
   accuracy                0.70      0.70      0.69       235
  macro avg                0.70      0.69      0.69       235
 weighted avg                0.70      0.69      0.69       235
```

Fig. 10 Results of RBF kernel of SVM

```
[Status] Loaded Features of shape (782, 16384)
[Status] Loaded labels of shape (782, )
Enter 1 for linear kernel
Enter 2 for polynomial kernel
Enter 3 for RBF kernel
Enter 4 for Sigmoid kernel
Enter your choice: 4
You have selected SVM Sigmoid kernel
Confusion matrix:
[[68 58]
 [52 97]]
Classification report
precision recall f1-score support
car 0.57 0.54 0.55 126
truck 0.50 0.52 0.51 109
accuracy 0.53 0.53 0.53 235
macro avg 0.53 0.53 0.53 235
weighted avg 0.53 0.53 0.53 235
```

Fig. 11 Results of sigmoid kernel of SVM

and true negatives values is 75, 51, 44, and 65, respectively. Precision values are 63% for car and 56% for truck. Recall values are 60% for car and 60% for truck. *F1*-score is 61% for car and 58% for truck, and overall accuracy of system using linear SVM kernel is 60%.

Machine learning model using polynomial kernel true positive, false negative, false positive, and true negatives values is 94, 32, 46, and 63, respectively. Precision values are 67% for car and 66% for truck. Recall values are 75% for car and 58% for truck. *F1*-score is 71% for car and 62% for truck, and overall accuracy of system using linear SVM kernel is 67%.

Machine learning model using RBF kernel true positive, false negative, false positive, and true negatives values is 81, 45, 27, and 82, respectively.

Precision values are 75% for car and 65% for truck. Recall values are 64% for car and 75% for truck. *F1*-score is 69% for car and 69% for truck, and overall accuracy of system using linear SVM kernel is 69%.

Machine learning model using sigmoid kernel true positive, false negative, false positive, and true negatives values is 68, 58, 52, and 57, respectively. Precision values are 57% for car and 50% for truck. Recall values are 54% for car and 52% for truck. *F1*-score is 55% for car and 51% for truck, and overall accuracy of system using linear SVM kernel is 60%.

5 Conclusion

In this paper, we have discussed various object classification papers in the literature review. We have presented comparative study of different kernels of SVM. From Table 1, we have observed that for RBF kernel precision, recall, and *F1* values for car are 75%, 64%, 69%, and for truck are 64%, 75%, 69%. The precision values mean true positive values and recall which means ratio of true positive and false negative is high with RBF kernel. From result, we have concluded that SVM classifier with RBF kernel provides best accuracy for this set of parameters.

Table 1 Comparison between different kernels of SVM

Kernel	Confusion matrix parameters	Car (%)	Truck (%)	Accuracy (%)
Linear SVM	Precision	63	56	60
	Recall	60	60	
	F1	61	58	
Polynomial	Precision	67	66	67
	Recall	75	58	
	F1	71	62	
RBF	Precision	75	64	69
	Recall	64	75	
	F1	69	69	
Sigmoid	Precision	57	50	53
	Recall	54	52	
	F1	55	51	

6 Future Scope

In future scope, focus needs to be given on other parameters of SVM such as C, gamma. Also, different classifiers can be utilizing for classifying multiple classes. Based on our current results and future experimentations, results will be helpful in the area of autonomous vehicle.

References

1. V. Vapnik, *Statistical learning theory* (Wiley-Interscience Publication, New York, 1998)
2. C. Papageorgiou, T. Poggio, A trainable system for object detection. *IJCV* **38**(1), 15–33 (2000)
3. P. Viola, M.J. Jones, D. Snow. Detecting pedestrians using patterns of motion and appearance, in *The 9th ICCV, Nice, France*, vol. 1 (2003), pp. 734–741
4. H. Schneiderman, T. Kanade, Object detection using the statistics of parts. *IJCV* **56**(3), 151–177 (2004)
5. H. Ling, C. Qian, W. Kang, C. Liang, H. Chen, Combination of support vector machine and K-fold cross validation to predict compressive strength of concrete in marine environment. *Constr. Build. Mater.* **206**, 355–363 (2019)
6. K.I. Kim, K. Jung, S.H. Park, H.J. Kim, Support vector machine for texture classification, in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, No. 11 (2002)
7. C.-W. Hsu, C.-J. Lin, A comparison on methods for multi-class support vector machines, in *Technical report, Department of Computer Science and Information Engineering, Nat'l Taiwan University* (2001)
8. S. Ghosh, A. Dasgupta, A. Swetapadma, A study of support vector machine based on linear and non-linear pattern classification, in *International Conference on Intelligent Sustainable Systems (ICISS 2019)*
9. M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, S. Homayouni, Support vector machine vs. random forest for remote sensing image classification: a meta analysis and systematic review, in *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2020)
10. W. Gang, Safety evaluation model for driverless car using support vector machine. *J. Intell. Fuzzy Syst.* **37**, 433–440 (2019)

11. Z. Guo, L. Zhang, D. Zhang, Rotation invariant texture classification using LBP variance (LBPV) with global matching. *Pattern Recogn.* **43**(3), 706–719 (2010)
12. I. Thammachantuek, S. Kosolsombat, M. Ketcham, Support vector machine for road image recognition in autonomous car, in *18th International Symposium on Communications and Information Technologies (ISCIT 2018)*
13. H. Qian, Y. Ou, X. Wu, X. Meng, Y. Xu, Support vector machine for behaviour based driver identification system. *Hindawi Publishing Corporation, Journal of Robotics* (2010)
14. S. Gupta, M. Sameer, N. Mohan, Detection of epileptic seizures using convolutional neural network. *Int. Conf. Emerg. Smart Comput. Inform. (ESCI)* **2021**, 786–790 (2021)
15. A. Mahajan, K. Somaraj, M. Sameer, Adopting artificial intelligence powered ConvNet to detect epileptic seizures. *IEEE-EMBS Conf. Biomed. Eng. Sci. (IECBES)* **2021**, 427–432 (2020)
16. S.M. Beeraka, A. Kumar, M. Sameer et al., Accuracy enhancement of epileptic seizure detection: a deep learning approach with hardware realization of STFT. *Circuits Syst. Signal Process.* (2021)
17. M. Sameer, A.K. Gupta, C. Chakraborty, B. Gupta, “Epileptical seizure detection: performance analysis of gamma band in EEG signal using short-time Fourier transform, in *2019 22nd International Symposium on Wireless Personal Multimedia Communications (WPMC)* (2019), pp. 1–6
18. M. Sameer, B. Gupta, Beta band as a biomarker for classification between interictal and ictal states of epileptical patients, in *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)* (2020), pp. 567–570
19. <https://ailearnerhub.com/2020/05/10/what-is-the-confusion-matrix/>
20. N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 05 (2005), pp. 1063–6919
21. H. Jiang, K. Huang, R. Zhang, Field support vector regression, in *Proceedings of the International Conference on Neural Information Process* (2017)
22. W. Shang, K. Sohn, D. Almeida, H. Lee, Understanding and improving convolutional neural networks via concatenated rectified linear units, in *Proceedings of 33rd International Conference on Machine Learning* (2016), pp. 2217–2225
23. Y. Chang, C. Hsieh, K. Chang, M. Ringgaard, C. Lin, Training and testing low-degree polynomial data mappings via linear SVM. *J. Mach. Learn. Res.* **11**, 1471–1490 (2010)
24. K. Huang, H. Yang, I. King, M.R. Lyu, Maximin margin machine: learning large margin classifiers locally and globally, in *IEEE Transaction on Neural Network*, vol. 19, no. 2 (2008), pp. 260–272
25. <https://www.kaggle.com/enesumcu/car-and-truck>
26. D.K. Agarwal, R. Kumar, Spam filtering using SVM with different kernel functions. *Int. J. Comput. Appl.* **136**(5) (2016)

Analysis and Prediction of Liver Disease for the Patients in India Using Various Machine Learning Algorithms



U. Sinthuja , Vaishali Hatti, and S. Thavamani

Abstract Predicting diseases in humans used to be an extremely time-consuming and complex procedure. It is now easier to save information and photographs due to the availability of multiple workstations and computers. Machine learning is vital in the healthcare sector because the number of liver patients is expanding on a big scale, so predicting liver illness at an early stage is essential to keep the patient from suffering more. The liver is a complicated organ located on the right side of our stomach that serves several key tasks in the human body. Many machine learning techniques are used in this study to categorize liver patient datasets, including logistic regression, k-nearest neighbor, decision tree, random forest, AdaBoost, LightGBM, XGBoost, and multilayer perceptron. These techniques are employed in the frame model to cleanse the collected dataset by using data preprocessing methodology, and data visualization was used to visualize the null values and substitute duplicates.

Keywords Liver disease · Machine learning algorithms · Data preprocessing methodology · Data visualization technique

1 Introduction

Machine learning algorithms have advanced in recent years significantly in the healthcare industry; this is extremely significant. Diseases judgment is based on even a clinical system. Several machine learning is being used by both scholars and businesses, to help with clinical diagnosis. The carried work concentrated liver disease identification, as the initial stage taken sample data has been preprocessed from selected features followed by the classification process done via some of the following machine learning methods has been processed using Python. The literature

U. Sinthuja (✉) · S. Thavamani
Sri Ramakrishna College of Arts and Science, Coimbatore, Tamilnadu, India
e-mail: sint@techie.com

V. Hatti
Vivek Vardhini Public School, Aland, Karnataka, India

review also carried out to show up the ideas of various researchers. Finally, the findings demonstrated the accuracy of several machine learning algorithms in predicting liver illness.

1.1 The Role Machine Learning in Liver Disease Perception

Machine learning is a subset of artificial intelligence that enables computers to behave like humans and make decisions with no need for social interaction. Machine learning has made significant advances in the detection of various diseases like liver disease as a result of recent advances in artificial intelligence. Furthermore, machine learning technology allows us to make more correct estimates and improve our efficiency. Machine learning can be classified into several types, as indicated in the diagram (Fig. 1).

In this study, some of the algorithms were chosen and processed to predict liver disease. This has happened in both supervised and unsupervised learning categories of then machine learning.

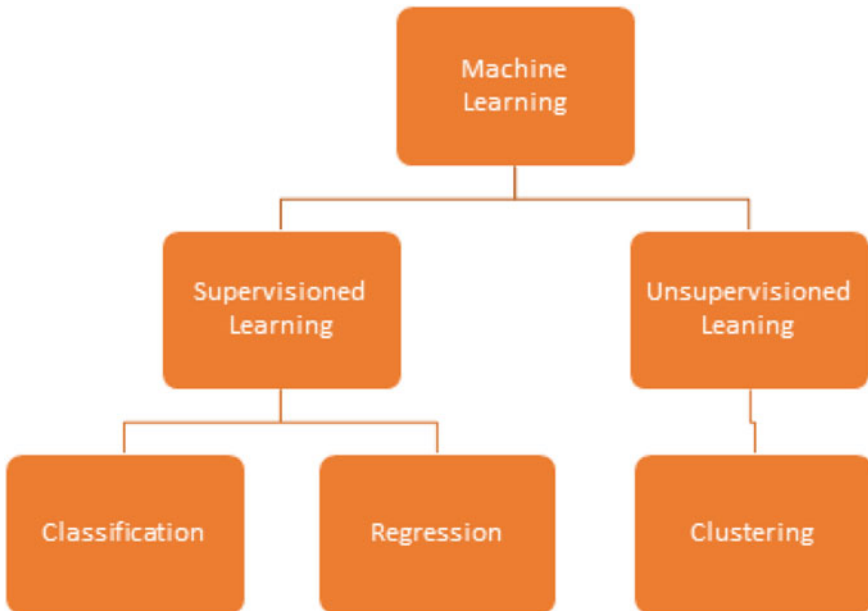


Fig. 1 Machine learning algorithm types

1.2 COVID-19 and Acute Liver Disease

Liver disease is also regarded as one of the world's most serious and deadly diseases [1]. Liver fibrosis, poor diet, cirrhosis of the liver, and HCV virus are some of the symptoms of liver disease, extreme alcohol consumption, drug usage, and hazardous and hereditary factors anomalies. There is no way to save a liver that is completely failing. There is only one method to regain it, is through liver transplant. It is extremely difficult to detect liver illness in its early stages, even when liver tissue has been moderately damaged; in these cases, many medical expert systems struggle to detect the sickness. It is critical to provide correct treatment in order to avoid this early prognosis and save the patient's life.

According to the Centers for Disease Control and Prevention, some COVID-19 patients exhibited high levels of liver enzymes. This indicates that a person's liver has been harmed, at least briefly, as a result of their sickness. Furthermore, patients with preexisting liver illness who have been diagnosed with COVID-19 have a greater mortality rate than people who do not have preexisting liver disease. It is vital to success this research right now in order to avoid future problems of the patient with COVID-19.

2 Literature Survey

The review has been done for the past one decade by the various researchers which is clearly depicted in Table 1.

3 Materials and Methods

The initial dataset for the process was acquired from the repository for liver illness. In the preprocessing stage, the data were sanitized. Classification was carried out using a couple of the abovementioned machine learning algorithms. Null values were removed from the collected dataset using data visualization. The findings have been included to demonstrate the higher accuracy of the machine learning algorithms used. This study was conducted using the block diagram shown in Fig. 2.

3.1 Data Collection

The liver patient dataset was used to create the dataset (ILPD). This is from the UCL machine learning repository, which you can find here. There are 567 instances and

Table 1 List of ML algorithms and its accuracy value for past decade to predict liver disease

S.No.	Year and Ref No.	Disease	Name of the ML algorithm	Accuracy (%)
1	2011 [2]	Liver	C4.5, NB, KNN, Backward propagation, and SVM	NB-95.07, C4.5-96.27, KNN-96.93, Backward propagation-97.47, and SVM-97.07
2	2012 [3]	Liver	Modified rotation forest	MLP-74.78 NN+CFS-73.07
3	2013 [4]	Liver	DT, NB, SVM, and ANN	DT-98.46 (which has given higher accuracy)
4	2014 [4]	Liver cancer, hepatitis, and cirrhosis	FT tree, NB	FT Tree-72.66 NB-75.54
5	2015 [5]	Liver fibrosis	DT	DT-93.7
6	2016 [6]	Liver disease disorder	C4.5, BPNN, regression, NB, SVM, and DT	C4.5 given higher accuracy
7	2017 [7]	Liver disease	Back propagation and SVM	Back propagation-73.2 and SVM-71
8	2018 [8]	Liver disease	KNN, ANN, logistic regression, and SVM	Logistic regression-73.23, KNN-72.05, SVM-75.04 and ANN-92.8
9	2019 [9]	Liver disease	C4.5 and k-means	C4.5-94.36 (highest accuracy)
10	2020 [10]	Liver disease	SVM, J48 and NB	J48-95.04 (highest accuracy)

10 attributes in this dataset. Age, gender, DB, TB, ALB, SGOT, SGPT, TP, ALP, and A/G ratio are all attributes (Table 2).

3.2 Data Preprocessing

Imputation of missing values: By viewing the data in seaborn or matplotlib, we can identify missing values and replace them with the mean (average) value, resulting in a high level of accuracy. If the number of null values (missing values) exceeds the

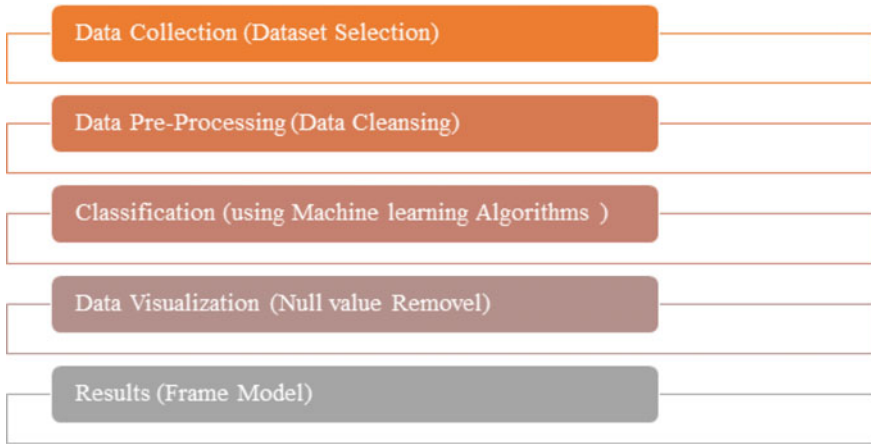


Fig. 2 Block diagram of the study model

Table 2 List of parameters and its data types

S.No.	Parameter name	Data type
1	Age	Integer
2	Gender	String
3	tot_bilirubin	Real
4	direct_bilirubin	Real
5	tot_proteins	Integer
6	Albumin	Integer
7	ag_ratio	Integer
8	sgpt	Integer
9	Sgot	Real
10	Alkphos	Real
11	is_patient	Integer

number of null values, the contained column can be removed or feature engineering used.

Label Encoding: The data now contain some string values in the gender column that should be changed to integers to improve the analysis. Label encoding, a phase in data preprocessing that focuses on transforming string values to integers, is now used (machine readable form).

Elimination of duplicate values: To increase the quality and efficiency of data, this technique involves the removal of unnecessary variables.

Figure 3 has shown that the dataset processing technique for feature selection (also known as variable selection, attribute selection, or variable subset selection) is a method of selecting the features in your data that contribute the most to the

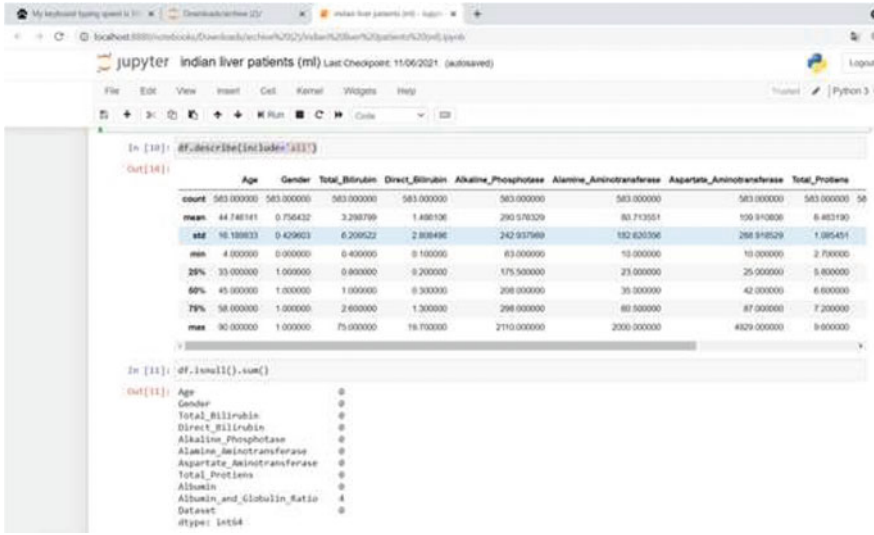


Fig. 3 Sample screen for dataset process

prediction (output). It simply means limiting the amount of input variables that do not contribute to the model. If the appropriate subset is chosen, it allows machine learning algorithms to train faster and enhances model accuracy. The data are now divided into two categories: independent (X) and dependent (y) features. In [11], NB has done research with researchers.

3.3 Classification Using Machine Learning Algorithms

In AI-based work done by the authors, the data are divided into two categories: training data and testing data, which are used to train and test the model, respectively [12–14]. The classification is done using a variety of machine learning algorithms which are discussed below.

Logistic Regression: The supervised learning classification technique logistic regression, also known as logit regression or logit model, is used to predict the likelihood of a categorical-dependent variable. There are several types of logistic regression (for example, binary logistic regression, multinomial logistic regression, and ordinal logistic regression), but because the dependent variable in the Indian liver patient dataset is a binary variable with data coded as 1 or 0, binary logistic regression is used to train the model [15].

K-Nearest Neighbor (KNN): KNN is a supervised machine learning algorithm that can perform both classifier and regression tasks using numbers (K) of neighbors (instances). Before the implementation of KNN, the given data must be preprocessed

(i.e., imputation, label encoding, and elimination of duplicate values, resampling, outlier detection, and feature selection) which means data should be balanced. Next, the data are divided as independent and dependent features.

Formulae to measure the distance between two points

Euclidean distance formula

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

But, other measures can be more suitable for a given setting that include the Manhattan, Chebyshev, and hamming distance. KNN is used to train a model to predict which class or feature a new data point belongs to, based on a value of K (it is preferable to consider the value of k as an odd number; by default, it is 5). The distance between the new data point and rest of the data points can be calculated using K -Closest numbers and the new data point is assigned to the class or feature based on the chosen point.

Decision Tree Classifier: Decision trees are a type of supervised machine learning in which data are continuously separated based on a parameter. Two entities, nodes and leaves, can be used to explain the tree. The decision nodes are where the data are split, and the leaves represent decisions or ultimate outcomes. Two mathematical equations, Gini Impurity and entropy, are used to apply decision trees for training any model [16].

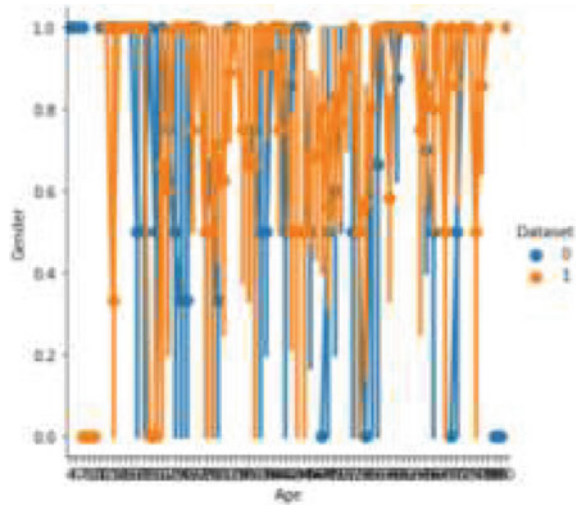
Random Forest Classifier: Random forests, also known as random decision forests, are an ensemble learning method for classification and regression that uses random samples from training data to create numerous decision trees. Random forest combines the results of numerous decision trees (each of which trains a different observation) to provide a more accurate and reliable forecast [17].

AdaBoost Ensemble Technique: AdaBoost, short for adaptive boosting, is a boosting approach used in machine learning as an ensemble method. It works on the premise of converting weak learners into strong ones. During the training of data, it creates n -number of decision trees. The record that was mistakenly classified during the initial model is given additional priority as the first decision tree is constructed. Only these records are sent to the second model as input. The procedure will continue until we designate how many base learners we wish to produce.

LightGBM: LightGBM is a high-performance gradient boosting framework based on decision tree techniques that may be used for ranking, classification, and a variety of other machine learning applications. It divides the tree leaf by leaf (instead of depth wise). It may result in overfitting, which can be reduced by setting the splitting depth.

XGBoost: XGBoost (eXtreme gradient boosting) is a gradient boosting framework-based decision tree-based ensemble ML method. To improve the algorithm, the processing is done in parallel. In [18–20], it is given importance to security in communication technologies.

Fig. 4 Sample screen for **age** and gender visualization of dataset 0 and 1



Multilayer Perceptron: It is a feedforward neural network augmentation. It is made up of three different layers.

- The input signal to be processed is received by the input layer.
- The required task, such as prediction and classification, is executed in the output layer.
- An arbitrary number of hidden layers is inserted between the input and output layers in a hidden layer.

3.4 Data Visualization

After the process of the dataset, Python will be displaying the processed data as shown in Fig. 4. Here, age and gender data have been visualized as an example.

3.5 Results and Discussion

The accuracy of each classification machine learning algorithm result has been given below after processing the dataset of the liver disease patient which has been taken as an example.

Table 3 shows that out of 11 classification-based machine learning algorithms studied, decision tree, random forest, LightGBM, and XGBoost generated the best results. According to these comparative studies, the best result supplied algorithms can be applied to improve prediction outcomes [21, 22].

Table 3 Results of proposed classification algorithms

S.No.	Proposed machine learning algorithms	Accuracy (out of 1.00)
1.	Logistic regression	0.70
2.	KNN	0.78
3.	Decision tree	1.00
4.	Random forest	1.00
5.	AdaBoost	0.88
6.	LightGBM	1.00
7.	Multilayer perceptron	0.70
8.	XGBoost	1.00

4 Conclusion

This research summarizes previous research on the identification and diagnosis of liver disease using various machine learning algorithms. This survey and study definitely found and noticed that several machine learning algorithms are used to provide superior accuracy in detecting and predicting liver illness. Different algorithms perform differently in different scenarios, but the dataset and feature selection are also crucial in obtaining superior prediction outcomes. Various procedures, such as imputation of missing values with the mean value and conversion of categorical data to numerical data, are used to clean the data. To forecast the existence or absence of liver disease, a variety of algorithms are used. In the future, based on the prediction results of the liver disease can lead to finding the way for treatment by the clinician, further assistance can also be done in future.

References

1. D.S.F.S.M. Abdel-Hamid, Incidence and risk factors for hepatitis C infection in a cohort of women in rural Egypt, vol. 102 (2008), pp. 921–928
2. B.V.M.S.P.B., N.B.V. Ramana, A critical study of selected classification algorithms for liver disease diagnosis. *Int. J. Database Manag. Syst.* (2011), pp. 101–114
3. B.V.M.P.B., N.B.V. Ramana, Liver classification using modified rotation forest. *Int. J. Eng. Res. Develop.* 17–24 (2012)
4. Y.G.S. Kumar, Prediction of different types of liver diseases using rule based classification model. *Technol. Health Care* **5**, 417–432 (2013)
5. H.O.S.G.A., K.M.A. Ayeldeen, Prediction of liver fibrosis stages by machine learning model: a decision tree approach, in *Third World Conference on Complex Systems (WCCS)* (2015), pp. 1–6
6. D.R.J.P. Sindhuja, A survey on classification techniques in data mining for analyzing liver disease disorder. *Int. J. Comput. Sci. Mobile Comput.* **5**, 483–488 (2016)
7. S.L.J., D.R. Sontakke, Diagnosis of liver diseases using machine learning, in *International Conference on Emerging Trends & Innovation in ICT (ICEI)* (2017)
8. J.J.C.M.J.M., E.I. Jacob, Diagnosis of liver disease using machine learning techniques. *Int. Res. J. Eng. Technol.* **4** (2018)

9. M.V., A.L.G.S. Sivakumar, Chronic liver disease prediction analysis based on the impact of life quality attributes. *Int. J. Recent Technol. Eng. (IJRTE)* **7**(6S5) (2019)
10. V.S.R., D.K. Durai, Liver disease prediction using machine learning. *Int. J. Adv. Res. Ideas Innov. Technol.* **5**(2) (2019)
11. M.S., N.M.S. Gupta, ROC analysis of EEG subbands for epileptic seizure detection using Naïve Bayes classifier. *J. Mob. Multimedia* **17**(1–3) (2021)
12. M. G. B. Sameer, “Detection of epileptical seizures based on alpha band statistical features,” *Wireless Pers Commun*, pp. 909–925, 2020.
13. M.S., N.M.S. Gupta, International conference on emerging smart computing and informatics (ESCI), in *Detection of Epileptic Seizures using Convolutional Neural Network* (2021), pp. 786–790
14. K.S., M.S.A. Mahajan, Adopting artificial intelligence powered convnet to detect epileptic seizures, in *IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)* (2020), pp. 427–432
15. A.K.G.C.C., B.G.E.M. Sameer, Epileptic Seizure detection: performance analysis of gamma band in EEG signal using short-time Fourier transform, in *22nd International Symposium on Wireless Personal Multimedia Communications* (2019)
16. M.S., B. Gupta, Beta band as a biomarker for classification between interictal and ictal states of epileptical patients, in *7th International Conference on Signal Processing and Integrated Networks (SPIN)* (2020)
17. A.K.G.C.C., B.G.M. Sameer, ROC analysis for detection of epileptical seizures using Haralick features of Gamma band, in *National Conference on Communications (NCC)* (2020)
18. U. Sinthuja, Efficient employment of the MQTT And S-MQTT protocol in IOT network. *Int. J. Adv. Sci. Technol.* 858–864 (2019)
19. S. et al., Evaluating systems and tools for vulnerability study on multi-broker MQTT. *GEDRAG & Organisatie Rev.* **33**(4) (2020)
20. S.K.A.S.M., A Beeraka, Accuracy enhancement of epileptic seizure detection: a deep learning approach with hardware realization of STFT. *Circuits Syst. Signal Process.* (2021)
21. M.G.B. Sameer, Time–frequency statistical features of delta band for detection of epileptic seizures. *Wireless Pers. Commun.* (2021)
22. U. Sinthuja, The hash idea for blockchain security. *INFOCOMP J. Comput. Sci.* **18** (2019)

Vision-Based Human-Following Robot



Ajay Thakran, Akshay Agarwal, Pulkit Mahajan, and Santosh Kumar

Abstract Vision-based human-following robot is a combination of a general wheel-based motion device with the concept of a human-following robot. Such robots can help humans in a plethora of ways, whether it is shopping in a supermarket using trollies or carrying luggage in an airport. We aim to create a system that allows efficient and effective tracking using a combination of algorithms, which communicate with minimal hardware devices. We have used vision-based trackers for detecting humans and following their movements, which give smooth and accurate tracking abilities.

Keywords Mobile robots · Human-following robot · Object tracking · Gesture control · Vision-based system · Line of sight motion · IoT · Obstacle avoidance

1 Introduction

In this emerging digital world, IoT promises to have a giant influence on the future. Every industry providing any type of service is approaching toward automation. Vision-based human following robot is one such spectacular implementation of IoT.

The applications of IoT are expanding to all Internet domains, and the concept is becoming a crucial part of our lives. From the more commonly used items like desktops, tablets, and mobile phones to the more complex ones like smart homes, intrusion detection systems, and people using assistive technologies like voice commands so as to provide comfort and security, IoT is making way to our everyday life at a surprisingly strong rate. So, it makes all the more sense for us to invest time, money, and effort on such a technology that we know is going to be of such importance in the upcoming years.

A. Thakran · A. Agarwal · P. Mahajan (✉) · S. Kumar
Department of Computer Engineering, NIT Kurukshetra, Kurukshetra, India

S. Kumar
e-mail: santosh.cse@nitkkr.ac.in

With this, we have created a human-following algorithm that when applied to a system on wheels, it allows that vehicle to move autonomously with minimal human interaction. All the person needs to do is move—as he/she would anyway have done—and the robot would follow the person. We have devised a system that handles use and edge cases while being cost-effective and using minimal hardware.

This paper focuses on the concept and its applications, keeping in mind some other work done along the same lines. Then, we provide our own approach to solve the problem. The subsequent sections provide information about the implementation done and the scenarios handled in this implementation.

2 Applications

The human-following robot system can be used in a wide variety of setups—for example, shopping malls, airports, warehouses, medical zones.

Keeping in mind the current situation of world, impacted by coronavirus, our concept can play a crucial role by helping the medical team in the containment zones inside the hospital by allowing hands-free transfer of equipment, and by doing this, chances of infection can be reduced. Just like how drones are being used for public surveillance, AI is being used to detect future outburst area, IoT devices can also be used to ensure compliance to quarantine and to manage patient care [1]. Specifically speaking, not only this project is beneficial for people who are carrying a lot of weight with them on airports/supermarkets, but also for disabled people. The government does a lot of things for the benefit of people with some sort of disability. Carrying the same thought forward, this project can prove to be a boon for blind/handicapped people as they do not have to carry their stuff themselves. They can put their stuff in a small cart, attach our system to it, and cart will automatically move with them. The possible applications of this concept are innumerable.

3 Related Works

There has been some work in the field which focuses on a similar strategy. A few of the papers have been summarized as follows.

The research paper presented in Ref. [2] addresses a novel architecture for person-following robots using active search. The system can be used for tracking and navigating toward that user. The research paper presented in Ref. [3] introduces a human-following robot, which has a CNN tracker which is based on stereo vision. This allows the system to detect, track, and follow the human in real time. The research paper presented in Ref. [4] shows another application of a human-following robot, which uses vision as its primary tracking medium. They provide an approach in which we must manually select the object to be chosen, and a point-based algorithm will basically virtually draw points on the human body. This is an iterative process that starts

with all the points being scattered. This project has the need for manually selecting the human to be tracked in the beginning and uses a Kinect camera and a laser to find the range of the object in the environment. The research paper presented in Ref. [5] proposes a controlled environment in which a robot agent can follow a human. This controlled environment, called intelligent space, consists of distributed independent networked devices. These devices are distributed in the controlled environment to detect and track the movements of the user and then send the processed data to central computer for deciding. The research paper presented in Ref. [6] uses a virtual spring strategy to connect a human and the robot that follows the human. If this spring is at a specific angle with the human, this strategy can be used for side by side following of the human.

Our system does not require any complicated/expensive piece of hardware. It handles occlusion efficiently and accurately with an algorithm ready to jump in whenever the line of sight breaks. This makes our research stand out—affordability, range of operation, and coverage of edge cases.

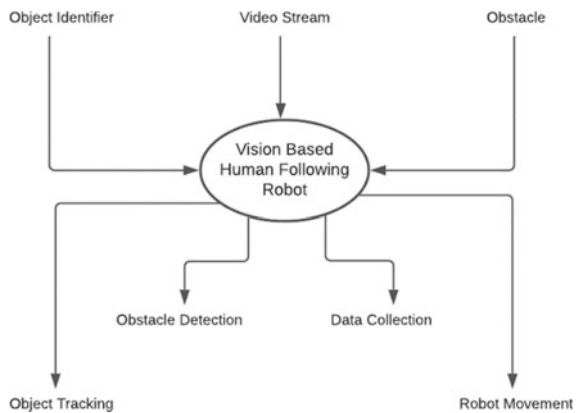
4 Proposed Approach

We plan to create a system that uses information such as the position of a person in the plane, distance between the robot and the person to drive the robot so as to follow that person (Fig. 1).

A simple approach can be used to track software using camera input and feed the information to a microcontroller to control the direction of movement. But, the biggest limitation of this system is that the human has to be in a direct line of sight of the system.

To overcome this limitation, we have added certain components (both software and hardware) such as the use of human detection algorithm [7], image comparison algorithm [8], Wi-Fi module, increased number of ultrasonic sensors and motors.

Fig. 1 Context diagram



We also added hand gesture detection system to allow the user to control the robot's tracking by hand gestures. The microcontroller gets its required high voltage from a separate lithium-ion battery. Finally, the entire apparatus is fixed on a chassis. The aim is to create a highly robust system that increases agility and reduces the possibility of breaking of line of sight between the robot and the human.

This is the high-level flow of information and data in our system. The robot will take the video stream as an input; the processor will parse it and send the movement data back to the robot to make it move. While moving, the robot might encounter any obstacle, which will be maneuvered by the obstacle detection system.

5 Initialization and Configuration

To start human following, the robot performs the following operations in sequence.

First, the host human to be tracked that stands in front of the robot. Then, a human recognition algorithm is run to detect the human in the frame, and a rectangular bounding box is drawn around the human (this image inside the box is important for the tracker). Then, using knee and back coordinates (which we got from our human recognition algorithm), we find the pixel colors near that region to get shirt and trouser colors. Robot initially takes multiple pictures of the human to be tracked for better detection and to avoid false positives. Finally, before starting, robot slowly moves backward from the human. Using ultrasonic sensors (that is pointing toward human), robot stores head coordinates of the human for some particular distances. Now, it has a table of approximate y coordinate of the human head for given distances from the human. Then, as the human starts moving, the robot follows the human at a predefined distance that it tries to maintain throughout it is in motion; i.e., if the human starts running, the robot will increase its speed and vice-versa so as to maintain that safe distance. All these activities of the proposed robot are powered by a small portable power source like a lithium-ion battery.

6 Tracking Approach

The information flow in the system works as follows:

1. A camera will capture real-time video as input and relay it as a stream of data to central processing server.
2. Server processes these incoming frames as follows:
 - 2.1. The person to be followed is selected and marked. Two pictures of the object are selected to get better tracking results.
 - 2.2. For each incoming frame, a tracker looks for that selected human in the frame and returns its position to the microcontroller.

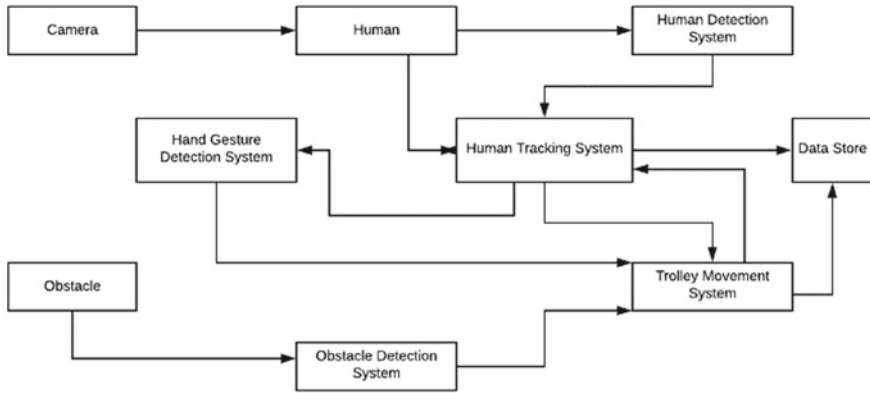


Fig. 2 Architecture diagram

2.3. Information related to the hand gesture of the person being tracked is also sent to the microcontroller (Fig. 2).

3. The microcontroller, gathering information from the front ultrasonic sensor, calculates the distance between the robot and the person being tracked. Based on this distance, the speed of the robot is decided.
4. The microcontroller, based on the received information, finally instructs the motor drivers and makes the robot move.
5. Along with all this, there is always another system in place that is constantly working. Whenever the human raises his/her hands up in the air, the gesture detection system comes to play and orders the processor to stop sending movement signals to the microcontroller. The human can perform said gesture again to restart sending movement signals to the microcontroller. This scenario is useful for whenever the human wants the robot to stop tracking, for example, if the person wants to visit the washroom. The person can do so in a quick, efficient, and hands-free manner. For this, we use the OpenPose library [9] which is the first real-time multi-person system to jointly detect human body, hand, facial, and foot key points (in total 135 key points) on single images.

7 Obstacle Avoidance

Ultrasonic sensors are used to detect obstacles in the direction of turn in order to decide when the robot should turn. For this purpose, we have attached an ultrasonic sensor on the left as well as one on the right side of the robot. This is also crucial whenever the robot has to move left or right. Imagine a situation when the robot gets a signal to move left and the left, ultrasonic sensor shows the presence of an object—the robot would not know what to do. To rectify this, we decided that the data coming from the left/right ultrasonic sensors will have precedence over other signals. Hence,

now, the robot can turn left/right only if the left/right ultrasonic sensor shows no obstacle present. Failing to do so might cause the robot to crash to the object present on the left.

If distance detected by ultrasonic sensor on front is less than minimum safe distance, robot checks if target human is present in the view, or something else is in front.

If the human is present, that means human has stopped and robot will wait for human to move. If the human is not present, this means human might have moved ahead and something is blocking the way. These options can be confirmed by checking whether the last position of human in the frame was in middle of the frame and distance from ultrasonic sensor suddenly got reduced to a value below the minimum safe distance. If this is the case, then definitely we have an obstacle blocking the line of sight. In this case, the robot will use left and right ultrasonic sensors to measure space available on the sides. Then, it makes a decision, depending on availability of space on sides, and moves around the obstacle and continues tracking the human.

8 Target Moves Out of Line of Sight

At each moment, the robot stores the last-known position of the human being tracked. Here, the last known position is the combination of the last-known distance between the human and the robot + the direction in which the human was turning. Direction of turn can be found by checking in which part of the frame the human was detected last. If last positional signal was in the left half of the frame, it would indicate that the human must have turned left, and similarly for right turn.

Hence, if a human takes a sharp turn and goes out of the frame, the robot will move according to the last-known position of the person. Here, the robot will go forward till the last-known position and will turn to the last-known direction. The ultrasonic sensors present on the front and on the sides enable the robot to turn without crashing to anyone/anything on the sides.

For the purpose of restarting tracking, we follow a three-step approach.

First, a human detection algorithm is used to detect humans in the frame which returns the positions of all the humans detected in the frame. Now, each of the frames detected in this step is compared with the original frame using an image comparison algorithm.

Second, we use the stored shirt and trouser colors of the original host human to match the same details with each person in the frame. This step acts as a confirmation to our image processing step that the selected person is indeed the original human host.

Third, the robot moves up to a certain distance near that person and performs height calculation of the person. If height is near about the expected height which was stored earlier, then robot starts following that human and tracker is reset. This step too confirms that the selected person indeed is the original human host.

If the above steps do not give positive results, then robot will perform the above processes again for the remaining individuals; i.e., this time, it will exclude the rejected person from the set.

9 Conclusion

The desired objective was to create a human-following robot using a camera and an ultrasonic sensor in a prototype model. The plan was to overcome limitations of line of sight motion and non-agile movement, along with adding additional functionalities such as hard stop, lane driving, and data collection.

References

1. S. He, Using the Internet of Things to Fight Virus Outbreaks. Technol. Netw. (2020), <https://www.technologynetworks.com/immunology/articles/using-the-internet-of-things-to-fight-virus-outbreaks-331992>
2. M. Kim, M. Arduengo, N. Waler, Y. Jiang, J.W. Hart, P. Stone, L. Sentis, *An Architecture for Person-Following using Active Target Search* (2018), <https://arxiv.org/abs/1809.08793v1>
3. B.X. Chen, R. Sahdev, J.K. Tsotsos, Integrating stereo vision with a CNN tracker for a person-following robot, in the 11th International Conference on Computer Vision Systems, Schezhen, China, July 10–13, 2017
4. M. Gupta, S. Kumar, L. Behera, V.K. Subramanian, A novel vision-based tracking algorithm for a human-following mobile robot. IEEE Trans. Syst. Man Cyber. Syst. **47**(7), 1415–1427 (2017). <https://doi.org/10.1109/TSMC.2016.2616343>
5. K. Morioka, J.-H. Lee, H. Hashimoto, Human-following mobile robot in a distributed intelligent sensor network. IEEE Trans. Ind. Electron **51**(1), 229–237 (2004). <https://doi.org/10.1109/TIE.2003.821894>
6. J. Hu, J. Wang, D.M. Ho, Design of sensing system and anticipative behavior for human following of mobile robots. IEEE Trans. Ind. Electron. **61**(4), 1916–1927 (2014). <https://doi.org/10.1109/TIE.2013.2262758>
7. Z. Wang, A.C. Bovik, Mean squared error: love it or leave it? A new look at signal fidelity measures. IEEE Signal Process. Magaz. **26**(1), 98–117 (2009). <https://doi.org/10.1109/MSP.2008.930649>
8. T. Watanabe, S. Ito, K. Yokoi, Co-occurrence histograms of oriented gradients for pedestrian detection, in *Advances in Image and Video Technology. PSIVT 2009. Lecture Notes in Computer Science*, vol 5414, ed. by T. Wada, F. Huang, S. Lin (Springer, Berlin, Heidelberg, 2009). https://doi.org/10.1007/978-3-540-92957-4_4
9. Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. Sheikh, *OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields*, <https://arxiv.org/abs/1812.08008>

MRI Cardiac Images Segmentation and Anomaly Detection Using U-Net Convolutional Neural Networks



Kriti Srikanth, Sapna Sadhwani, and Siddhaling Urolagin

Abstract Healthcare industry is increasingly adopting artificial intelligence in analyzing laboratory and radiology outputs to provide optimal treatments for patients. Medical imaging has been playing an important role in understanding several underlying conditions of patients. With rise in incidents of cardiac diseases world-wide, usage of computer vision and deep learning methods are proving to be very useful in detecting anomalies that are conventionally done using human perception. This paper aims at establishing efficacy of using convolutional neural network in detecting cardiac anomaly. The hypothesis is substantiated through the process of predicting systole and diastole volumes from MRI scan images of left ventricle and calculation of ejection fraction which is a vital parameter in assessing cardiac dysfunction. In this research project, representative dataset and normalization techniques were used to arrive at the results published. With the progression in medical imaging techniques combined with training the model with high volume of stratified data can result in very high accuracy of outcome. The model was successfully able to predict the diastole and systole volumes of any given image based on image segmentation, pixel values, slice locations, and other DICOM metadata. The cardiac anomaly was determined using calculated ejection fraction from the predicted systolic and diastolic volumes.

Keywords Artificial intelligence · Healthcare · Deep learning · Neural networks · Automated segmentation of heart · Convolutional neural network · U-Net · Ejection fraction · TensorFlow

K. Srikanth (✉) · S. Sadhwani
Department of Computer Science, Birla Institute of Technology and Science, Pilani, Dubai,
United Arab Emirates
e-mail: f20160049d@alumni.bits-pilani.ac.in

S. Sadhwani
e-mail: sapna@dubai.bits-pilani.ac.in

S. Urolagin
Department of Computer Science, APP Centre for AI Research (APPCAIR), BITS-Pilani, Goa,
India
e-mail: siddhaling@dubai.bits-pilani.ac.in

1 Introduction

Artificial intelligence that was earlier used as lagging technology for analysis is now being used in preventive and early detection of illness and monitoring of progression, proving better clinical outcomes. In this research, a model is trained to study scanned image of a heart and automatically perform segmentation [1], predict volumes, calculate ejection fraction and suggest cardiac anomaly which is comparable to experienced radiologists' assessments. Typical use case of adoption is given in Fig. 1.

The paper is organized as under: Sect. 2 describes the functional scope of the work. Section 3 details data preparation and segmentation of left ventricle image, Sect. 4 describes the predicted results and analysis using AI algorithm to estimate systole and diastole and focusses on calculation of ejection fraction to infer anomalies. The last part concludes the work and outlines future potential for research in this domain.

2 Functional Scope of Work

Cardiac volumes and ejection fraction are determined by examining and analyzing MRI scans. Cardiac functions are calculated by measuring the end-systolic volume (amount of blood in the ventricle at end of heart contraction) and end-diastole volume (amount of blood in ventricle before contraction) of the left ventricle [2]. The two volumes are further used to derive ejection fraction to find anomaly correspondingly. Medical specialist can use the calculated parameters in tandem with other comorbidity or lifestyle factors of a patient to provide appropriate treatment. The following functions were performed to support the hypothesis:

- (a) Use of convolutional neural network model to segment MRI images in digital imaging and communications in medicine (DICOM) format of heart left ventricle.
- (b) Training of images using a deep learning model.
- (c) Designing of an algorithm to predict total systole and diastole volumes.
- (d) Developing a program to derive ejection fraction from resultant volumes that would be used to identify cardiovascular anomalies.

3 Image Segmentation

Segmentation in this work refers to identification of left ventricle from MRI scan image and to predict a pixel-wise mask of the area [3]. Multiple slices of each heart based on different physical location were selected with one slice containing 30 frames (cardiac cycle). One cardiac cycle has a minimum and maximum volume called end-systolic and end-diastolic, respectively.

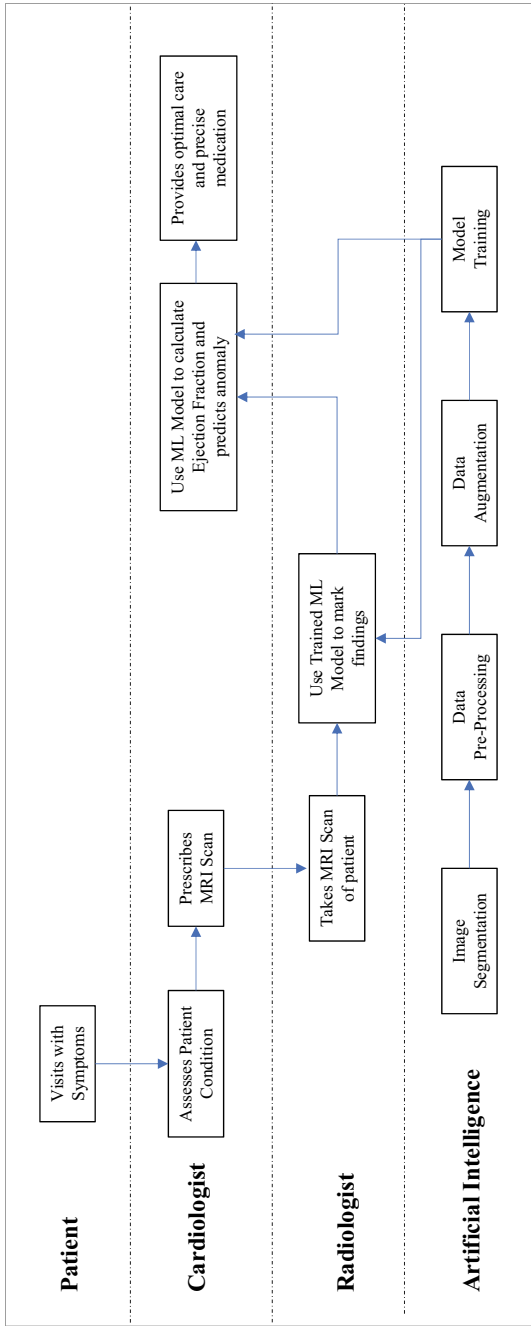


Fig. 1 Proposed model

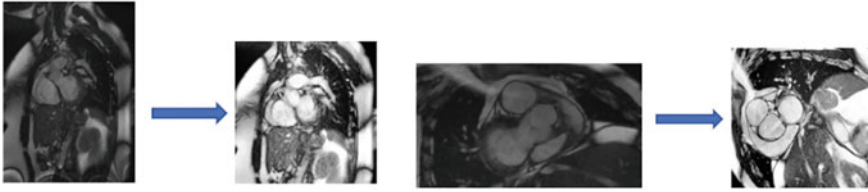


Fig. 2 Output of image preprocessing

3.1 Image Preprocessing

For data preprocessing, selected patient IDs were chosen from each dataset. As the images had variances and were at different scales, the first step was to render them uniform by scaling and cropping [3]. Difference between the slices had to be calculated to avoid overlapping using parameters like slice location slice thickness. The overlapped slices were then deleted to increase the accuracy of prediction. The meta-data extracted from the data preprocessing were stored in separate files for predicting cardiac volumes. OpenCV machine learning library was used for preprocessing and data augmentation (Fig. 2).

3.2 Data Augmentation

Augmentation refers to applying different transformations to the data which enhances the network performances. Elastic deformation, geometrical transformation, and contrast limited histogram equalization techniques were applied on images. Inter_area and Inter_nearest interpolations were used assuming antialias = False. After resizing, a pixel might get another color. Inter_nearest (nearest neighbor interpolation) is used to interpolate pixels between the source image and destination image to select the color of the nearest pixel in the area. No antialiasing interpolation will be faster when dealing with large datasets.

3.3 Deep Learning Algorithm for Segmentation

Artificial neural networks are designed based on neurons trying to simulate human brain. A basic neural network structure includes input function (weights and bias), an activation function, and output [4]. Of several neural network models, convolutional neural network (CNN) is more profound in the field of medical image segmentation and hence used in this work as well among other variants [5]. Among the several deep learning methods [6] for image processing, U-Net model, a convolutional neural

network architecture is widely used in biomedical field and is hence used in this work for training and performing image segmentation and labeled masking [7].

3.4 *Semantic Segmentation*

Image is segmented into several smaller segments, and each pixel is allocated with a label. Of various segmentation methods, semantic segmentation was chosen as the images were in gray scale and required only one object and its respective pixel-wise mask [7]. Pixel array is extracted from the image to get pixel values which scan the image horizontally from top left corner and returns tuple as rows and columns. The pixel value is accessed through row and column coordinates correspondingly. Maximum pixel values are required to get distinct object from the image.

3.5 *U-Net Implementation*

U-Net [4] can distinguish borders and objects by classifying every pixel, also the input and output have the same size. U-net is important in medical imagery as it not only classifies existence of anomaly but also localizes the area of anomaly. In this paper, the developed U-Net model delivers more efficient performance as compared to existing works in this field [7]. Various parameters and configurations were tested on the U-Net model before implementing the solution for the problem.

A few key insights and enhancements were recorded while training the model. Heavily depended on batch normalization as the segmentation nets were unstable. Small batch sizes provided more stability and the most effective augmentation technique used was elastic deformations. Logistic regression loss function was better than root mean square error. Finally, the addition of more layers to the network enhanced the performance of the model more accurately. After several trial and error, ended up with the network shown in Fig. 3. This U-Net architecture is effective compared to a few of the other implemented networks [8]. The results obtained after segmentation was impressive. Simple situations were almost ideal. Images in which the left ventricle tissues were half-covered were perfectly segmented into an overlay.

Left ventricle images that were processed using the architecture were passed through two paths viz. contraction and expansion [4]. In the contraction phase, images were downsampled repeatedly to extract dominant features with 3×3 convolution [4]. This was followed by activation function and 2×2 max pooling. In the expansion path, precise localization was achieved by upsampling and resizing the image to original dimension [4]. At every step of the expansion path, the output from the convolutional layer was concatenated with feature map of the contraction path.

Input feature map: a ; Weights: w ; Bias: c ; Output feature map: b ; Features: x , y , k :

3×3 Convolution + ReLU:

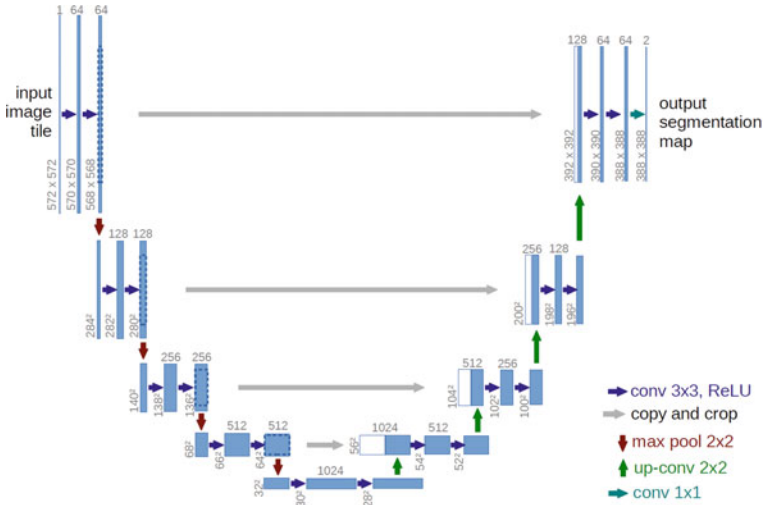


Fig. 3 U-Net architecture description [4]

$$b_{x,y,l} = \text{ReLU} \left(\sum_{\substack{i \in \{-1, 0, 1\} \\ j \in \{-1, 0, 1\} \\ k \in \{1, \dots, k\}}} w_{i,j,k} \cdot a_{x+i,y+j,k} c_l \right) \quad (1)$$

2 × 2 Convolution + ReLU:

$$b_{2x+i,2y+j,l} = \text{ReLU} \left(\sum_{\substack{i \in \{0, 1\} \\ j \in \{0, 1\} \\ k \in \{1, \dots, k\}}} w_{i,j,k} \cdot a_{x+i,y+j,k} \right) \quad (2)$$

Layer activation function ReLU was used to transform the summated weights and bias from node to activation layer for output.

$$y = \max(0, z) \quad (3)$$

Table 1 Convolutions performed using U-Net model

U-Net layers	Size-filters/strides	U-Net layers	Size-filters/strides
Input image	Image size (cropped)	Concatenation 1	Max pooling 1, crop conv
Convolutional layer 1	$3 \times 3; 32$	Dropout 2	50%
Max pooling 1	$2 \times 2; 2$	Convolutional layer 8	$3 \times 3; 128$
Convolutional layer 2	$3 \times 3; 64$	Deconvolutional layer 3	$128 \times 2 \times 2; 2$
Max pooling 2	$2 \times 2; 2$	Convolutional layer 9	$3 \times 3; 128$
Convolutional layer 3	$3 \times 3; 128$	Concatenation 2	Convolutional layer 9, max pooling 2
Max pooling 3	$2 \times 2; 2$	Dropout 3	50%
Convolutional layer 4	$3 \times 3; 128$	Convolutional layer 10	$3 \times 3; 128$
Max pooling 4	$2 \times 2; 2$	Deconvolutional layer 4	$128 \times 2 \times 2; 2$
Convolutional layer 5	$3 \times 3; 256$	Convolutional layer 11	$3 \times 3; 128$
Max pooling 5	$2 \times 2; 2$	Concatenation 3	Convolutional layer 11, max pooling 1
Dropout 1	50%	Dropout 4	50%
Deconvolutional layer 1	$128 \times 2 \times 2; 2$	Convolutional layer 11	$3 \times 3; 64$
Convolutional layer 6	$3 \times 3; 128$	Deconvolutional layer 5	$64 \times 2 \times 2; 2$
Deconvolutional layer 2	$128 \times 2 \times 2; 2$	Convolutional layer 12	$3 \times 3; 64$
Convolutional layer 7	$3 \times 3; 128$	Convolutional layer 13	$1 \times 1; 1$
Crop convolutional-crop conv	Convolutional layer 7	Output	Segmented mask image

$$R(z) \begin{cases} z, & z > 0 \\ 0, & z \leq 0 \end{cases} \quad (4)$$

The probabilistic loss was calculated using binary cross entropy class which is critical for measuring the performance of the model.

$$-(y \log(p) + (1 - y) \log(1 - p)) \quad (5)$$

$$-\sum_{c=1}^M y_{o,c} \log(p_{o,c}) \quad (6)$$

The outstanding results of the model can be inferred from comparing the work with previously achieved neural networks image segmentation. This model has achieved to yield more accurate overlay and segmentation of complex images. The researchers [9] have used similar convolutional neural networks techniques for segmenting the cardiac left ventricle image. Convolutional neural networks do produce higher accuracy in simpler picture, but not very accurate for more difficult images. The process carried out is quite traditional and many features might have been compromised during the extraction process. This is where U-Net architecture comes into the equation. The AI model used in this paper has more enhanced FCN, U-Net architecture model for segmentation. Not only must the feature map be converted into a vector, but also have to rebuild the image from this vector. This process is difficult because converting a vector to an image is far more difficult than converting it from image to a vector. The entire concept of U-Net is built to tackle this difficulty.

In this U-net model, 13 convolutional layers were used with less strike strokes, which means the model will extract more features and has higher accuracy. From trial and error method, it was made sure that the increased number of neural network layers did not result in overfitting or underfitting of the network. While using CNN [10] on images with high complexity, it could be really hard to identify the left ventricle and could lead to improper segmentation. But when this U-Net model was used on test images, it predicted segments with higher accuracy than any other model which implemented only CNN. The model developed will extract every feature from an image for segmentation and has accuracy up to 94% for high complexity left ventricle images.

4 Results

4.1 *Experimental Setup*

The work involved processing of actual images in an operable environment with limited sample size and demonstration of output of the models. The dataset was sourced from Kaggle, second annual data science bowl which consisted of hundreds of cardiac left ventricle MRI images. The actual number of images varied for each case depending on the number of slices, recorded perspectives, and also different number of frames provided in time sequences. The primary angle chosen for predicting cardiac volumes was the short axis view [11] which is perpendicular to the left ventricle. The end-systolic volumes, end-diastolic volumes, and ejection fractions are calculated from this plane of view. The train dataset contained images with its respective end-systolic and end-diastolic volumes. The training set consisted data of 500 patients. Each patient's heart had multiple slices (15–25) which varied

throughout. Also, each slice had 30 frames which contained the images of heart at different positions during the cardiac cycle.

Datasets were selected for model training, testing, and validation. This was followed by extraction of data from images, training of the model for segmentation and masking. Finally, systole and diastole volumes were predicted from segmented images and ejection fraction calculated. TensorFlow 1.13.1 was used as the framework and Keras 2.2.4 APIs, and machine learning modules were used to build the convolutional layers, train the U-Net model, model saving, serialization, continuity in case of error and prevention of overtraining and reduction of learning rate. Pydicom 1.2.2 and OpenCV (cv2) 4.0.0.21 were used for preprocessing the images and extracting metadata from the DICOM images.

4.2 Volume Prediction Using Pixel Classification

The penultimate stage was to predict the systole and diastole volume for a set of new images. The cardiac volumes were calculated using the preprocessed data and metadata. Volumes were predicted based on the pixel values, slice location, and few other parameters. Pixel values and frustum volume formula were used to predict the volumes. In case of frustum, the frame with highest volume as diastole and the frame with the lowest volume as systole. The first step in the procedure was to load the trained U-Net model. The patient images were then preprocessed, and the overlays (segmented masks) were predicted. Each frame location was extracted from the DICOM images, and the distance between two slices (image position) was computed.

The pixel values obtained from the metadata were interpolated, and the number of pixels in each frame was calculated from transparent overlays using slice thickness, slice location, frame pixel series, and frame confidence series. Frustum value was computed by calculating the distance using pixel series. Finally, the volumes were predicted by calculating the pixel sum for each frame. Figure 4 represents the output of the steps that prepare the images for volume prediction and predicted overlays.

Predicted overlay image will have two-pixel regions, the background and foreground in grayscale format. The segmented mask will have pixel value up to 255. Threshold value is required to identify the pixels. As it has two regions, the threshold

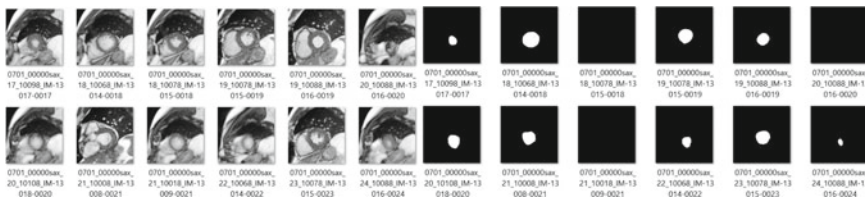


Fig. 4 Extracted segmented masks from patient images to predict cardiac volume

value must be below or above to satisfy the requirements. Area of the segmented mask is calculated by number of pixels in the frame. Frame with the highest pixel value is considered as diastolic volume, and the frame with lowest pixel value is considered as systolic volume. Every patient id will have 10–12 slices and each slice has 30 frames. The largest and smallest frames were taken from each slice as cardiac volumes. After computing all the frames in every slice, the maximum and minimum volumes were taken from each series which will represent the cardiac volumes for a given heart (Table 2).

4.3 Analysis

The volumes of systolic and diastolic given in the test data were used to improvise the prediction by calibration. To detect and minimize systematic inaccuracy, gradient boosting regressor was used along with few additional features and estimations. Experimentation with numerous features showed lots of potential for predicting many mistakes like percentage pixels and the frequency of missed slices. There were a variety of targets like estimated volume, true volume, and RMSE error on which regression was performed. Given all of the characteristics, the gradient boosting regressor predicted the estimation error. After calculation of errors, it was easier to modify the process which led to a more accurate prediction of volumes. The mean absolute errors resulted in 1 ml improvement in the average after calibration.

The results were presented as mean values with standard deviations. Biases between automated systole/diastole estimations and reference measures were evaluated for relevance and to compare the accuracy. Root mean square error was calculated to measure the spread of residuals and data. According to the Bellenger remodeling data [12], Bland Altman margins of acceptance for left ventricle end-systolic volume is 19 and 24 ml for end-diastolic volume. The total ejection fraction RMS error is about 6.5%. The calculated RMSE resulted in this model is 11 ml and 17 ml for end-systolic and end-diastolic volumes, respectively. Two estimations were evaluated against the reference values one by one. The correlation between the predicted volumes and the clinical volumes of test values was verified using linear regression and Bland Altman analysis. The automated volume predictions were quite accurate to the standard reference which demonstrated high correlation with $r = 0.9437$, $p < 0.023$ and minimal bias of 1.1%.

4.4 Calculation of Ejection Fraction

Ejection fraction is the amount of blood pumped out from the left ventricle (LV) every time the heart contracts. Ejection fraction (EF) is usually measured only in the LV and it is represented in percentage [13]. The standard ejection fraction (LVEF) varies from 55 to 70%. For example, if the ejection fraction is 65%, it means that

Table 2 Predicted volume of systole and diastole

Slice	Slice_thickness	Slice_location	Time	Slice_dist	Diastole	Diastole_vol	Systole	Systole_vol
14_8.2438520959741	8	8.243852096	93,007.744	9.999997772	0.107052613	107.0525889	0.370418785	37.04187023
15_18.243849867519	8	18.24384987	93,111.807	10.00000198	0.904768996	90.47691752	0.133736787	13.37368134
16_28.243851848743	8	28.24385185	93,142.861	9.999997772	0.173119907	173.1198689	0.903623013	90.36228116
17_38.243849620287	8	38.24384962	93,213.514	9.999997772	0.542193761	54.21936399	0.903832809	90.38326074
18_48.243847391832	8	48.24384739	93,245.628	10.00000171	0.760540279	76.05404088	0.712418301	71.24184225
19_58.24384910239	8	58.2438491	93,316.448	10.00001676	0.96832361	96.83252329	0.963656903	96.36585185
20_68.243865865636	8	68.24386587	93,358.639	10.00000135	0.412641115	41.26411701	0.423311563	42.33115634
21_78.24386721232601	8	78.24386721	93,431.503	9.999994958	0.253292608	25.32924801	0.221673525	22.16734136
22_88.243862170167	8	88.24386217	93,511.139	10.00000097	0.334274178	33.42742107	0.903171145	90.31712322
23_98.243863136136	8	98.24386314	93,543.177	10.00000171	0.866787803	86.67879514	0.711635603	71.16356028
24_108.24386484669	8	108.2438648	93,619.364	10.00000128	0.560302645	56.03027168	0.667279916	66.72800012

Table 3 Output of ejection fraction and heart anomaly

End Systolic Volume	65.6784
End Systolic Volume	106.5789
Estimated Ejection Fraction	38.37579483
Ejection Fraction	38.83%

Result moderate dysfunction—Mild heart failure with reduced (HF-rEF)

65% of total amount pumped out by the left ventricle with every beat. One of the ways EF is calculated by an MRI scan of the heart which is the basis of this work [13]. Ejection fraction value above 55% is considered as normal.

Ejection fraction percentage was calculated using the standard mathematical formula using the inferred end-systolic and diastolic values from Eq. (8).

$$SV = EDV - ESV \quad (7)$$

$$EF(\%) = \frac{SV}{EDV} \times 100 \quad (8)$$

SV = Stroke Volume; EDV = End-diastolic Volume; ESV = End-systolic volume. Below is output result for a test patient (Table 3).

5 Conclusion

From the studies augmented with demonstrated implementation, it is evident that artificial intelligence and neural networks techniques are effective in automating the medical procedure. In summary, the first part of the work dealt with image segmentation of MRI images and training a convolutional neural network model to predict/identify the left ventricle of a heart. In the second part, the segmented mask was used to predict cardiac volumes, i.e., end-systolic and diastolic volumes and calculation of ejection fraction to indicate heart dysfunctions. The model yielded high accuracy rate of volume prediction and especially the perfection of segmenting complex left ventricle heart images [9].

Extending the research work, there is significant opportunity to develop newer deep learning neural network models based on causation capabilities to accurately predict existing or progressing anomalies with minimal model training and at real time. Further, with continuous ingestion of exogenous data, the clinicians, practitioners, and medical researchers can be assisted with suggestive precision medication to provide best possible outcomes. For example, monitoring risks based on correlation between ejection fraction and arrhythmia (irregular heart beat) or between ejection fraction and blood pressure can help prevent sudden cardiac failures among outpatients. The architecture can be extended to other areas of healthcare

use where computer vision capabilities can be used to rapidly assess conditions of vital organs such as lungs, liver, and kidney based on X-Ray, CT, ultrasound, MRI, telemedicine, and other imagery. In the situation such as COVID-19 pandemic where unprecedented number of patients have to diagnosed and treated, such models can help to scale out assessments and accelerate the diagnosis to save lives and prevent community spread.

References

1. M. Avendi, A. Kheradvar, H. Jafarkhani, Fully automatic segmentation of heart chambers in cardiac MRI using deep learning. *J Cardiovasc Magn Reson* **18**, P351 (2016)
2. M. Hadhoud, M. Eladawy, A. Seddik, F. Montecvecchi, U. Morbiducci, Left Ventricle Segmentation in Cardiac MRI Images. *Am. J. Biomed. Eng.* (2012)
3. Y. Lu, P. Radau, K. Connelly, A. Dick, G. Wright, Automatic image-driven segmentation of left ventricle in cardiac cine MRI. MIDAS J—Cardiac MR Left Ventricle Segmentation Challenge
4. O. Ronneberger, P. Fischer, T. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*
5. S. Ghosh, N. Das, I. Das, U. Maulik, *Understanding Deep Learning Techniques for Image Segmentation* (2019)
6. M.H. Hesamian, W. Jia, X. He et al., Deep learning techniques for medical image segmentation: achievements and challenges. *J. Digit. Imaging* **32**, 582–596 (2019)
7. B. Seo, D. Mariano, J. Beckfield, V. Madenur, Y. Hu, T. Reina, M. Bobar, M. Nguyen, I. Altintas, *Cardiac MRI Image Segmentation for Left Ventricle and Right Ventricle using Deep Learning* (2019)
8. Palit, et al., Biomedical image segmentation using fully convolutional networks on true north, in *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*, Karlstad (2018)
9. M. Nasr-Esfahani, M. Mohrekesh, M. Akbari, S.M.R. Soroushmehr, E. Nasr Esfahani, N. Karimi, S. Samavi, K. Najarian, *Left Ventricle Segmentation in Cardiac MR Images Using Fully Convolutional Network* (2018)
10. S.M. Ibrahim, M.S. Ibrahim, I. Naseem, Heart Segmentation. From MRI Scans Using Convolutional Neural Network, in *2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)*
11. C. Petitjean, J.N. Dacher, A review of segmentation methods in short axis cardiac MR images. *Med. Image Anal.* **15**(2), 169184 (2011). <https://doi.org/10.1016/j.media.2010.12.00>
12. N.G. Bellenger, L.C. Davies, J.M. Francis, A.J. Coats, D.J. Pennell, Reduction in sample size for studies of remodeling in heart failure by the use of cardiovascular magnetic resonance. *J Cardiovasc Magn Reson.* **2**, 271–278 (2000)
13. K.B. Waghlikar, et al., Extraction of ejection fraction from echocardiography notes for constructing a cohort of patients having heart failure with reduced ejection fraction (HFrEF). *J. Med. Syst.* **42**(11), 209 (2018). <https://doi.org/10.1007/s10916-018-1066-7>
14. F. Sultana, A. Sufian, P. Dutta, *Evolution of Image Segmentation using Deep Learning Convolutional Neural Network: A survey* (2020). [arXiv:2001.04074](https://arxiv.org/abs/2001.04074)

Communication

GSM-Based Smart Marine Tracking System for Location Monitoring and Emergency Protection



T. Kesavan and K. Lakshmi

Abstract An effective marine monitoring system is required to ensure safe marine environment to passenger, fishermen, cargo, and other type of naval vessels. Marine monitoring includes ship location monitoring for border control and emergency encountering and protection. This is done by providing an automatic location and communicating marine hardware to all shipmen which is to be placed on their deck while leaving off shore for fishing. GPS and GSM interlinked device are connected directly to the satellite which is received position in 24*7 time. Each boat has been given a unique ID through which they are differentiated. Each fisherman boats position is remotely accessed on naval surveillance control room which keeps on tracking their position and stores them in their database. Through that particular ID control room alerts or offers help to that particular boat. In-case of encounter of border crossing or emergency condition of the fisherman boats, the GPS device in the fisherman boat sends an information to the marine control room through its satellite connection. If the shipman where supposed to cross their borders, then they are warned and contacted by the communication establishment between the navy guard and the fisherman. In-case of emergency situation, the current status of the boat (location, type of emergency, condition, no. of hostiles, etc. ...) is send through that particular GPS device. This information is classified on the back end on the naval control surveillance where all data are being processed. It provides a clear and wide vision of naval observations.

Keyword Marine monitoring systems · Exclusive economic zones · GPS · GSM

T. Kesavan (✉)

Department of Electrical and Electronics Engineering, Easwari Engineering College, Chennai, India

K. Lakshmi

Department of Electrical and Electronics Engineering, Sri Krishna College of Technology, Coimbatore, India

e-mail: lakshmi.k@skct.edu.in

1 Introduction

Marine monitoring systems (MMSs) are a general term to portray frameworks that are utilized in business angling to permit natural and fisheries administrative associations to track and screen the exercises of angling vessels. They are a key piece of checking control and reconnaissance (MCS) programs at national and global dimensions [1–3]. MMS might be utilized to screen vessels in the regional waters of a nation or a subdivision of a nation, or in the exclusive economic zones (EEZ) that expand a few miles from the shores of numerous nations. MMS frameworks are utilized to improve the administration and manageability of the marine condition, through guaranteeing legitimate angling rehearses and the counteractive action of unlawful angling, and subsequently secure and upgrade the jobs of anglers [4–6]. The careful usefulness of a MMS framework and the related hardware differs with the necessities of the country of the vessel's vault, and the local or national water wherein the vessel is working [7]. Inside provincial and national MMS activities there are likewise subdivisions which apply distinctive usefulness to various vessel classifications. Classes might be size or kind of vessel or movement [8, 9].

In this paper, MMS relates explicitly to fisheries the executives frameworks. MMS portrays the particular use of checking business angling vessels. It is not to be mistaken for VTS which is portrays the particular utilization of observing marine traffic basically for security and effectiveness in ports and occupied conduits [10–12]. It is additionally not to be mistaken for explicit correspondence innovations, for example, AIS, Iridium, Inmarsat, Argos, GPRS which identify with the specialized technique on which information is transmitted. Various MMS frameworks will utilize diverse correspondence innovations relying upon the usefulness necessities forced by a national or provincial MMS activity [13–15].

The expense of MMS parts will differ as indicated by the usefulness prerequisites of the particular framework being executed. By and large, the higher the usefulness the more costly the gear and required information interface (broadcast appointment costs) [16–18]. The expense of a VMS framework along these lines shifts, and in this way, the dimension of government appropriation (assuming any) fluctuates as per national and local prerequisites. EU and US MMS frameworks require costly locally available hardware and a lot of information to be transmitted over satellite connection bringing about high broadcast appointment charges, yet in addition give an abnormal state of usefulness. In different areas where per vessel cost and tremendous armada sizes are an issue, correspondence advancements, for example, AIS are utilized which altogether diminish gear and broadcast appointment costs while conveying satisfactory essential MMS framework usefulness [19, 20].

2 Marine Tracking System

In India, many accidents have been happened in coastal areas due to poor monitoring system. Emergency condition like Tsunami cannot be intimated to the fisherman. Also, their location cannot be monitored periodically. Our project mainly aims in 24*7 monitoring of the marine vessels to provide a better safety and reliable services to them [20]. Any sort of emergency aid is provided to the vessel through our technique. We achieve this by providing a location detecting module to the end user (vessel) which is to be implanted in the hull of the ship. This module consists of Arduino board, GSM module, GPS module and a LCD which is used to display emergency messages in it. Location of vessel can found by GPS and information can send to user and control room in emergency time by GSM. Block diagram of GSM-GPS-based marine controlling system as shown in Fig. 1. The proposed system entire operation can describe in above block diagram. Ardunio controller is main heart of the proposed block diagram and it is used to control and give the instruction to all the devices based on our requirement. In this block diagram, there are five blocks are mainly operated. First process starts from GPS devices, it can to give the status of the boat location. So, location information can start the work like compare current location with programmed location.

Ardunio can desired our boat is not started; boat is started; boat circulating on safe location; boats are possible to cross unsafe location or boats are return to the country. This all statement can desired by Ardunio program depends on the signal of GSM location.

GSM is another important part of our proposed system. It is able to send and receive the signal between Ardunio and controller room. In this concept, GPS module acts as an INPUT unit. The GPS module gets the data on location of the vessel in a form of latitude and longitude string and sends that information to the Arduino Uno board. The GSM module and LCD display are used as the output device. GSM is a

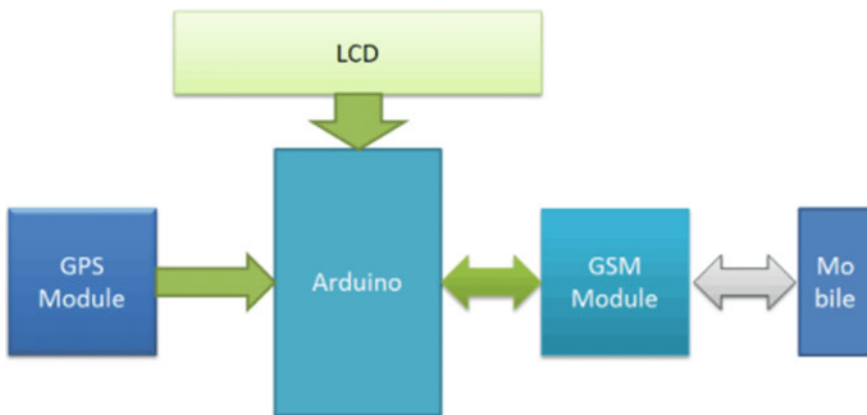


Fig. 1 Figure label IoT-based marine monitoring system using GSM and GPS

versatile correspondence modem; it is represents worldwide framework for portable correspondence. In our project, we use GSM module to send message when the interrupt is called. It is used to send the message to marine control station on the location of the vessel. It sends the message to the number which is already stored in the GSM module. Liquid crystal display may be used to display the condition of the system. In this method, LCD is shows the information about condition of the boat and it declares the results of the GPS location. In initial condition that means before start the boat in see, it displays boat is not starting from place, if boat is started it shows boat from place and it is in allocated area. In sometimes, if boated going to cross the limit means, before cross the limit, it shows the navigation of the boat and it gives the message like boat going to cross border like that (Fig. 2 and Table 1).

If Arduino get signal GSM and Audino asked to GSM to give the emergency alert to control room and ask to arrange backup from government. Another time due to bad environment like flood Tsunami control room can send information like all of written to your place due to bad environment anther GSM can receive signal from

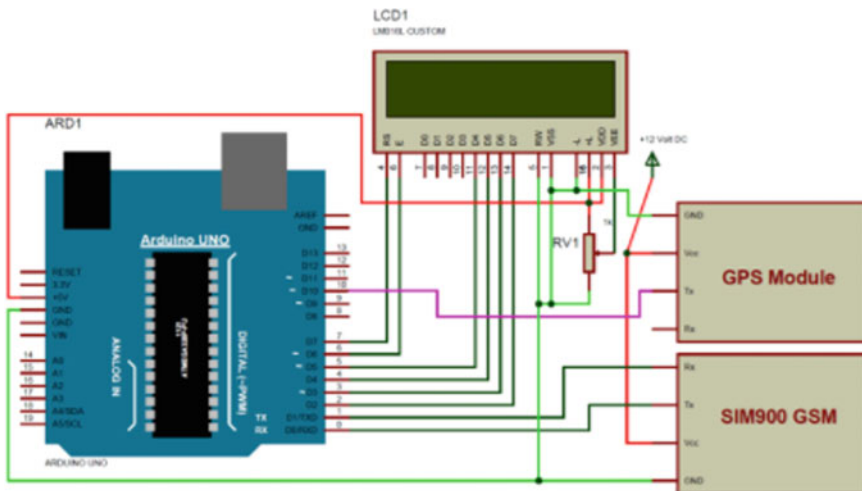


Fig. 2 Circuit connections of marine tracking system

Table 1 Table label pin connection of Arduino

Pins used	Type	Usage	Header file used
A0, A1, A2, A3, A4	Analog pins	LCD output display	Liquid Crystal.h
3, 4	Tx, Rx pins	GPS module serial communication pins	Software Serial.h
9, 10	Tx, Rx pins	GSM module serial communication pins	Software Serial.h
5 V and Gnd	Power supply	To provide power to the GSM, GPS, LCD modules	Normal access No library needed

control room and give the information to order no \$1 now Audino can stop the boat and driver return the boat immediately.

LCD is a frequently used good device to display the information presence days. In that project, LCD source status of both and message receive and send from GSM. Mobile phone is easy usage and cheap compact device. Cremation may be note down mobile phone also in proposed system.

3 Hardware and Software Description

The following hardware devices are used to implement the proposed hardware circuit.

- Arduino board
 - GPS module
 - 16 × 2 LCD
 - Power supply
 - Connecting wires
 - 10 K POT
- #### HARDWARE SETUP AND RESULTS

GSM-based proposed marine monitoring system hardware as shown in Fig. 3 (Fig. 4).

LCD may be linked with Audino number of 0 and 1. Program is uploaded in Arduino Kit by software. Security implementation and connections models of marine tracking system your Fig. 2. Global position system is an important part of proposed system aunties connect with order no boat. GPS shows location on the boat on time and in given information in emergency time. Beginning, we used to show the location of boat and its signal to order no board and some signal will be display in LCD monitor also. If GPS is connected with location gives only HTML server. Buy cost and some server Website. In this Website, exact location of particular system like Google Map. In this concept, GPS module is enough to show the location. In future, Google Map, private Website server aur government set server may be connected with our own GPS module to source the exact place. In this way, satellite communication has used to improve the accuracy of location. In our research, sin 900 GSM port has send and receive the information between transmitter and receiver. GSM mode operation is similar to the some mobile operation. Kishore Singh versus Gandhi frequency range. In particular concept, getting GPS center. GPS function is to convert the location to information and IT actus GPS Fort best and the GPS signal and GSM is send and receive the information to controller room through the order no. There are four pins are used in GSM connector. Pin number 1 2 and 4 are connected Adreno device and third pin of GSM are merged with GPS model. GPS send all the information likell started boat, crossed, and emergency message to GSM. In GSM, signals for encoder and decoder with particular frequency. In above concept, started signal is with very low frequency. If boat in center of own country ocean that signals are linked with low frequency. If emergency period or very near with border of other country ocean, that type emergency signal is bounded with very low level frequency.

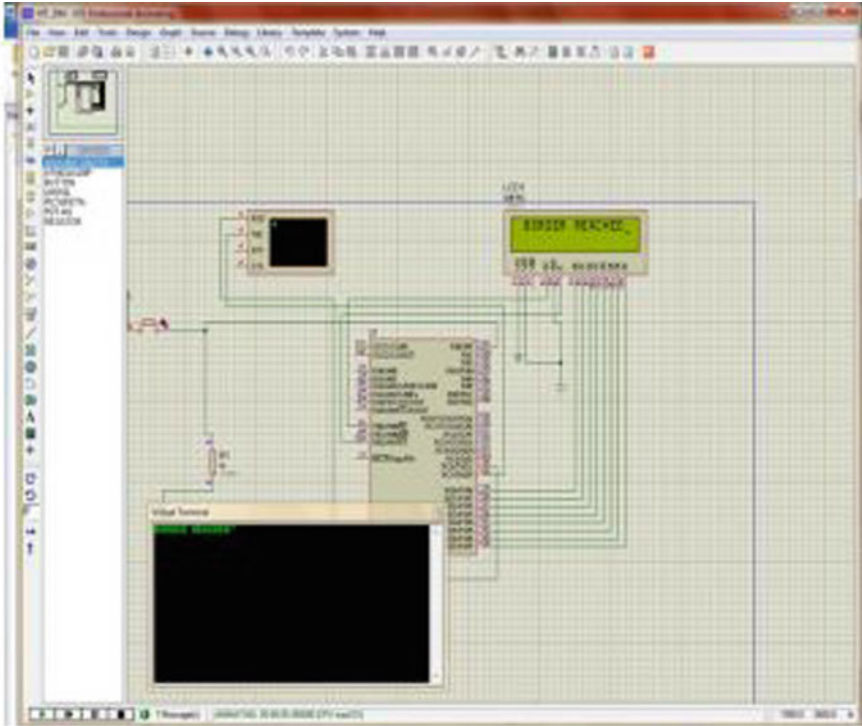


Fig. 3 Simulation of marine monitoring system



Fig. 4 Hardware setup of GPS-GSM-based marine monitoring system

So it is possible to send more number of messages and quick responses to control room. In advanced proposed method, control room replies the messages to boat on that time GSM device has receiving the information from control room and it convey the message to Arduino board. Arduino device is one of the developing devices in electronic field. In particular concept, Arduino is main part of device and other all devices are operates based on the requirements and instructions of the Arduino board. In general, condition program is developed based on our application, concept, idea, and methodology. Developing program may be uploaded to Arduino board through connecting wires. IDE software is mostly used for upload the program in Arduino kit. In normal period, Arduino can operate 5v Dc supply.

There are three external devices are tapped with Arduino port based on the uploaded program of Arduino. GPS and GSM may be function in particular project and texts are displayed in LED depends on the written program in Arduino. GPs location is minkled with Arduino. In normal location of boat that means inside the country, program shown boats are not started. If location of ocean based on the GPS position, program shown the text is boart is started and center on own country ocean. If sometimes other country oceans limits are crossed or going to cross immediately, it gives emergency alarm and message of red alert. In GSM, sending and receiving messages are control by Arduinio controller. Emergency message is send to controller based on the request of Arduinio If GSM receives any stop message or alert message from control room and GSM gives signal to Arduinio-based GSM requirements.

In proposed hardware, the GSM sends messages to the phone about the latitude and longitude position of the tracking system. This information is received as text message. Also, the location can be shown on the LCD display. The module can also be used in dual way mode to receive messages from the control room for imitation of the marine vessels about the emergency conditions like Tsunami, Earthquake, storms, etc. ... at the middle of the ocean. GPS module is used to send us the information which slightly deviates from our original location by a half a mile.

3.1 Advantages of Proposed Method

- Provides safe and secure marine journey. Enable better monitoring of vessels over all several extend off the shore.
- Emergency condition like accident, natural calamities, etc. ... can be intimated between the vessels and the control room before any sort of damages occurs.
- Remote database can be used to store the location of each and every vessel and their conditions. So, their positions can be retrieved and accessed later.
- Used to intimate the vessel whenever they cross the border.

3.2 Application of Proposed Method

- It can be used in the marine control room for efficient monitoring of the ships around their coastal coverage.
- The same technology can be used in the freight monitoring for tracking our goods.
- Vehicle can also be monitored by this system for efficient administration and management of roadways.

4 Conclusion

This solution of marine monitoring is a cost-efficient solution since it can be easily implemented in each and every vessel just by mounting it on the hull of the vessel. Thus, providing a safe and secure journey to the Indian fisherman with clean monitoring and support been provided by the Indian coastal guards and Indian Navy. The same project can be further developed with a database software to store the location of each and every ship and its location on the map. Also, further development can be made to directly connect the boats with the satellite with help of an amplifier circuit connected and coupled to the GSM and GPS module for monitoring them from very remote areas even off the globe. Interactive environment can be made so that each vessel can communicate with each other by a duplex way communication. Government of India can go ahead and implement this project on each and every ship to make their travel a safe journey.

References

1. S. Yamuna, T. Kesavan, Harmonic compensation in residential distribution system with Mppt. *Int. J. Appl. Eng. Res.* **10**(20), 15737–15741 (2015)
2. J.C. Sawhill, *Richard Cotton, Energy Conservation: Success and Failures*, Brookings Institution Press
3. T. Kesavan, K. Lakshmi, S. Sheeba Rani, R. Kavim, M. Senthilkumar, Design and study of interleaved step up DC converter with high level gain for the application of solar photovoltaic module. *Int. J. Recent Technol. Eng. (IJRTE)* **7**(6), 1426–1431 (2019). ISSN: 2277–3878
4. D.P. Kothari, *Renewable Energy Resources and Emerging Technologies*, Prentice Hall of India Pvt. Ltd
5. R. Kavim, T. Kesavan, S. Sheebarani Gnanamalar, K. Rameshkumar, Optimal charging and discharging planning for electric vehicles in energy saving system. *IEEE Conference Proceedings*, 978-1-5386-9533-3ccv, pp 976–978 (2019)
6. R. Chedid, Y. Saliba, Optimization and control of autonomous renewable energy systems. *Int. J. Energy Res.* **20**(7) (1996)
7. R. Kavim, T. Kesavan, V. Nandagopal, T. Malini, Fuzzy based Ev charging with reduced power fluctuation under renewable power consumption constraint. *Int. J. Pure Appl. Math.* **119**(18), 1691–1706 (2018)
8. K. Bansa, M. Meliss, *Renewable Energy Sources and Conversion Technology* (Tata Mc Graw Hill)

9. B. Clive, *Energy Management, Supply and Conservation* (Routledge)
10. P. Nagya, K.M. Nema, S. Rangnekar, *A Current and Future State of Art Development of Hybrid System Using PV System and Wind Energy*, vol. 13, issue 8 (2009)
11. S. Paul, R.N. Bhattacharya, *CO2 Emission from Energy use in India: a Decomposition Analysis*, vol. 32, issue 5 (2005)
12. S. Sheeba Rani, T. Kesavan, V. Gomathy, A. Anie Selva Jothi, Effectiveness of pitch control scheme in load balance of WECS. *J. Adv. Res. Dyn. Control Syst.* **10**(11), pp. 517–523 (2018)
13. R. Kavin, T. Kesavan, Compensation of harmonics in residential distribution system using virtual impedance. *Int. J. Sci. Eng. Technol.* **6**(8), 290–295 (2017)
14. R. Kavin, T. Kesavan, A smart monitoring of faults in power transformers and maintenance based on Wi-Fi. *Int. J. Eng. Res.* **6**(8), 382–387 (2017)
15. T. Kesavan, K. Lakshmi, S. Sheebarani Gnanamalar, R. Kavin, Local search optimization algorithm based monitoring and controlling of virtual power plant. *Distrib. Netw. Int. J. Pure Appl. Math.* **119**(12), 1851–1864 (2018)
16. O. Lopez, J. Alvarez, J. Doval-Gandoy, F.D. Freijedo, Multilevel multiphase space vector PWM algorithm. *IEEE Trans. Ind. Electr.* **55**(5), 244–251 (2008)
17. J.F. Moynihan, M.G. Egan, J.M.D. Murphy, Theoretical spectra of space vector modulated waveforms. *IEE Proc Electr. Power Appl.* **145**(1) (1998)
18. V.G. Agelidis, A.I. Balouktsis, M.S.A. Dahidah, A five level symmetrically defined selective harmonic elimination PWM strategy: Analysis and experimental validation. *IEEE Trans. Power Electr.* **23**(1), 19–26 (2008)
19. J.W. Chen, T.J. Liang, A novel algorithm in solving nonlinear equations for programmed PWM inverter to eliminate harmonics, in *23rd International Conference on Industrial Electronics, control and Instrumentation IEEE IECON, 97*, vol. 2, pp. 698–703 (1997)
20. K. Lakshmi, T. Kesavan, R. Kavin, S. Sheebarani Gnanamalar, M. Senthilkumar, V. Gomathy, Quick search optimization algorithm-based implementation of virtual power plant for distribution network. *Adv. Intell. Syst. Comput.* **1163**, 261–272 (2021)

Defected Ground UWB Antenna for Microwave Imaging-Based Breast Cancer Detection



Anupma Gupta, Paras Chawla, Bhawna Goyal, and Aayush Dogra

Abstract Microwave imaging (MWI) is the critically significant practice for the exposure of primary stage cancer in breast. It helps to lessen the figure of mortalities related with breast cancer. A proper sensor is a central facet of the designing of the MWI system for high-resolution images. In this article, an UWB antenna with operating frequency range of 8.4 GHz (3.1–11.5 GHz) is designed and evaluated for microwave imaging of breast tissue. Partial ground plane of tapered slot patch radiator is modified by printing a tapered slot in it, which helped to achieve good and stable impedance matching at the UWB spectrum. Backscattered signal variation is evaluated with and without the existence of malignant cell in the breast tissue. Stable and linear time domain performance of the proposed antenna makes it suitable for microwave imaging technique.

Keywords Tapered slot · UWB · Microwave imaging · Tumour detection · Backscattered signal

1 Introduction

More than 1.5 million new breast cancer cases are reported every year. It is the main reason of death for women; early diagnosis of cancer can improve the survival rate by 97% [1]. X-ray mammography examination requires painful compression of breast tissue. Furthermore, radiations and ionization caused by mammography create various health issues that also have possibility of converting healthy cells into malignant. Ultrasonic-based diagnosis fails to detect deep and solid tumours [2]. It accentuates the requirement of reliable and efficient techniques for early breast cancer detection. Microwave imaging (MWI) is a non-ionized (reduces radiation risk), reliable, comfortable, and inexpensive technology to diagnose the breast cancer

A. Gupta (✉) · P. Chawla · B. Goyal
Department of ECE, Chandigarh University, Mohali, India

A. Dogra
Ronin Institute, Montclair, NJ 07043, USA

at early stages. Ultra-wideband (UWB) technology is providing an efficient solution for microwave imaging. High data resolution, less complex system, and low power requirement are some important features of UWB technology [3].

In relations of breast cancer, there is large variation in the electric permittivity of healthy and malignant cell tissue. This variation is the key parameter for microwave imaging to identifying the cancerous cells. In microwave imaging, signal is transmitted across the breast tissue and the backscattered signal is collected in different directions through antennas. Received backscattered signal carries propagation information from various cell tissues of heterogeneous properties. By applying algorithms and signal processing techniques, signal information is mapped into image to diagnose and detect the location of tumour [4, 5].

UWB antennas for MWI must possess some stringent requirements for efficient and high-resolution images. Directional radiation pattern with high efficiency, wide impedance covering bandwidth from 3.1 GHz to 10.6 GHz for better penetration, and high-fidelity factor (FF) [6]. Various design approaches such as tapered slot [7], staircase shaped corner radiator with + slot [8], vivaldi antennas [9–11], monopole [12, 13], defected ground structures [14, 15], and horn antennas [16, 17] have been presented for UWB technology. Though the aforementioned antennas occupied UWB spectrum, these structures have some limitations like low efficiency, gain, and low directivity. To enhance the performance and compatibility of antenna for microwave imaging, artificial magnetic conductors [18], electromagnetic structures [19], meta-material structures [20], and many more structures are incorporated. However, these techniques make MWI system complex and expensive.

In this paper, a low-profile planar antenna structure, covering UWB spectrum is designed for microwave imaging of unhealthy cells in breast tissue. Antenna has covered wide frequency range that it enables the penetration of radiations at various depth. Frequency domain performance of antenna is analyzed in terms of scattering parameters and electric field distribution. In time domain, transmitted and received timing pulse are analyzed for side by side and face to face simulation setup of antennas with breast tissue.

2 Antenna Design and Simulation Setup

Designed UWB antenna has overall dimensions of $40 \times 40 \times 1.6$ mm. FR4 is used as a substrate material with electric permittivity (ϵ_r) of 4.4 and loss tangent (δ) of 0.02. Antenna dimensions are initially considered at the minimum frequency of 3.25 GHz according to the design equations mentioned in [21]. Microstrip feed line of width 1.6 mm is used to excite the radiator. Initially, antenna is designed in free space to achieve the wide impedance bandwidth of more than 6 GHz. Radiating patch is designed with tapered sides along vertical axis. It causes the generation of multiple resonance modes in the antenna and support wide band resonance frequency. Further to attain stable impedance matching for the desired band, a tapered slot of length 9.68 mm is etched in the partial ground plane. The geometrical design parameters of

Table 1 Geometrical parameters of the proposed antenna

UWB design parameters	Value (mm)	UWB design parameters	Value (mm)	UWB design parameters	Value (mm)
L_p	33.19	L_s	40	W_{g2}	11.92
L_g	19.65	W_s	40	W_p	11.6
W_{g1}	13.6	L_f	10	–	–

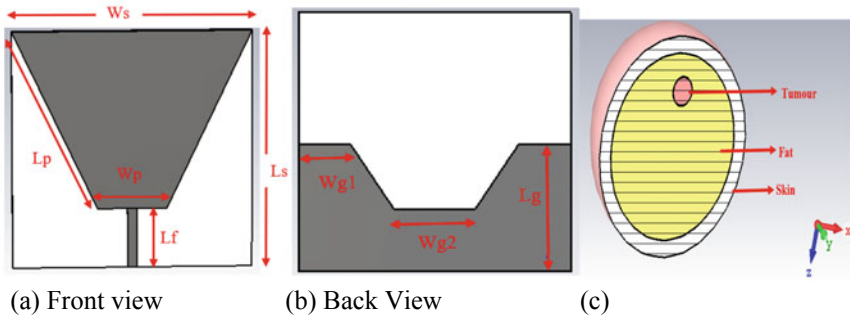


Fig. 1 a, b Antenna geometry and c breast tissue

UWB antenna are listed in Table 1. Design topology of the UWB antenna is shown in Fig. 1a, b.

The definitive aim of designed antenna is to sense the unhealthy cells inside the breast by recognizing the variation in backscattered signals of antenna in the presence and absence of malignant cells. A spherical shaped breast tissue phantom of 30 mm radius is designed with equivalent electrical properties as mentioned in [22]. It consist of skin layer ($\epsilon_r = 38$, thickness 2 mm, and conductivity = 1.49 S/m) and fat layer ($\epsilon_r = 5.141$, thickness = 28 mm, and conductivity = 0.141 S/m). Spherical tumour tissue of $\epsilon_r = 67$, conductivity of 47 S/m, and radius of 4 mm is inserted inside the breast phantom at a depth of 5 mm. Designed phantom model along with tumour is shown in Fig. 1c.

Simulation setup for the tumour detection is represented in Fig. 2. To collect the backscattered signal, two antenna structures are placed in side by side and face to face orientation at a distance of 5 mm between antenna and tissue.

3 Result and Analysis

Designed antenna performance is first evaluated in free space. Reflection coefficient plot ($|S_{11}|$) and the voltage standing wave ratio (VSWR) plot are shown in Fig. 3. Antenna has occupied the wide frequency spectrum from 3 GHz to 11.5 GHz. VSWR value for the entire bandwidth is below 2, which shows the impedance matching of

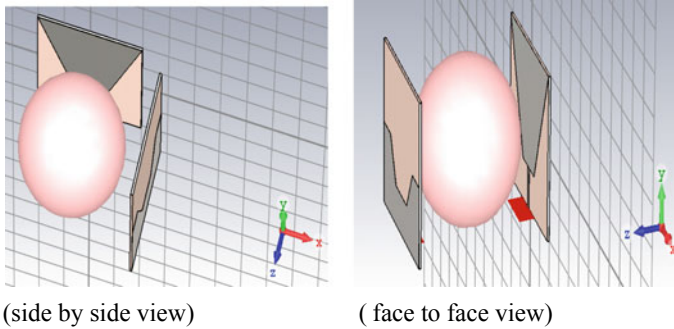


Fig. 2 Simulation setup for cancer detection

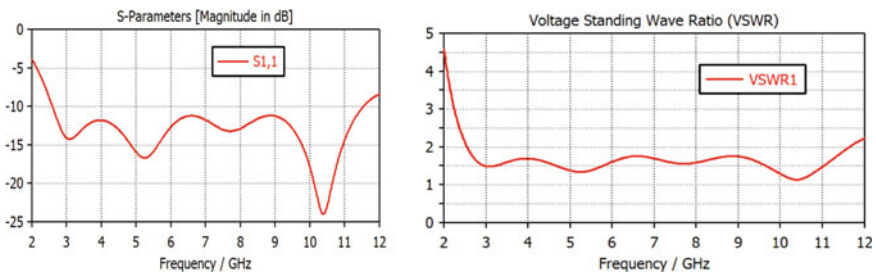


Fig. 3 Antenna $|S_{11}|$ parameter and VSWR plots

antenna. Radiation characteristics of antenna are shown in Fig. 4. Over the UWB frequency range, antenna has 80% of radiation efficiency. Maximum efficiency of 92% is achieved at 3.5 GHz. Radiated power at 3.5 GHz is oriented towards +Z direction, and for higher bands, concentrated towards X and Y directions which is required for the transmission of maximum power in the body tissue. Antenna performance is compared with the existing structures in Table 2.

4 Antenna Performance Analysis for Microwave Imaging

It is studied that healthy tissue has low water content and low relative permittivity as compared to the tumour tissue. Thus, more power will be scattered and reflected at the interface of different tissues. Backscattered signal of the antenna is plotted in Fig. 5. Antenna has covered the bandwidth from 3.1 GHz to 11.5 GHz; which is required for penetration of electromagnetic radiations at varying depth. When antenna is placed parallel to the breast tissue (without tumour), significant variation in the backscattered signal ($|S_{11}|$) can be observed in Fig. 5.

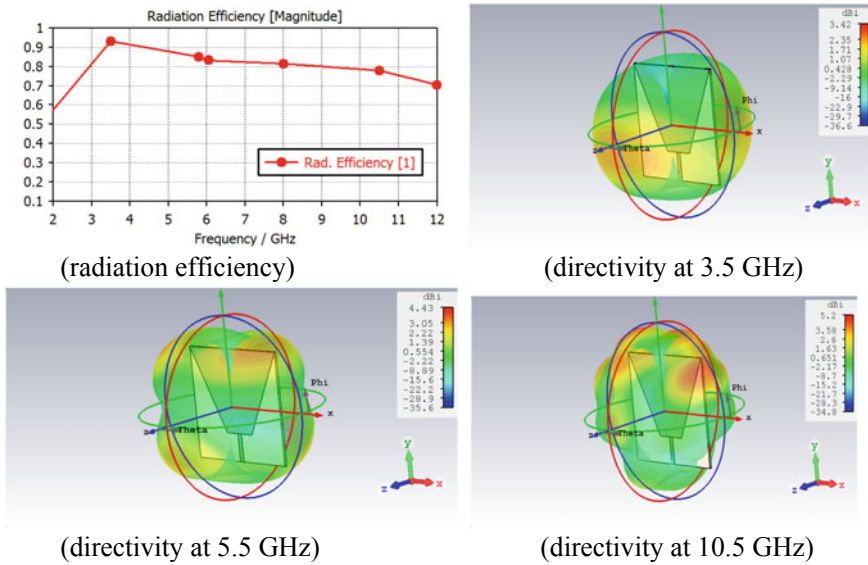


Fig. 4 Radiation performance

Table 2 Antenna performance comparison with existing structures

References	Size (mm ³)	Bandwidth	Maximum gain	Discussion
[8]	29 × 27 × 1.6	6.5 GHz	Not specified	Lacking frequency domain analysis
[15]	34 × 36 × 1.6	2.3 GHz	Not specified	Suffers with low bandwidth that reduces accuracy
[17]	40 × 50 mm ²	1.8 GHz	Not specified	Scattering parameter variation is not analyzed
This work	40 × 40 × 1.57	8.5 GHz		Wide bandwidth with linear phase distribution

Similarly, after inserting the tumour cell in the breast tissue, variation in antenna scattering parameter is observed. For both the evaluation setup, it can be found that only frequency ranges from 7.0 GHz to 7.5 GHz and 10.0 GHz to 10.5 GHz is not much altered by the presence of tumour. |S₁₁| of the remaining frequency of the entire operating band has shown significant variation and makes antenna suitable to trace the existence of tumour cell. Transmission coefficient |S₂₁| of the antenna for side by side and face to face configuration of the two matching antennas is plotted in Fig. 6. To consider the farfield effect of the two antennas over the whole bandwidth, |S₂₁| is analyzed. In both the operating scenario, stable |S₂₁| is achieved except at 5.0 GHz in face to face setup. Similar type of variation in transmission coefficient is also observed in [21].

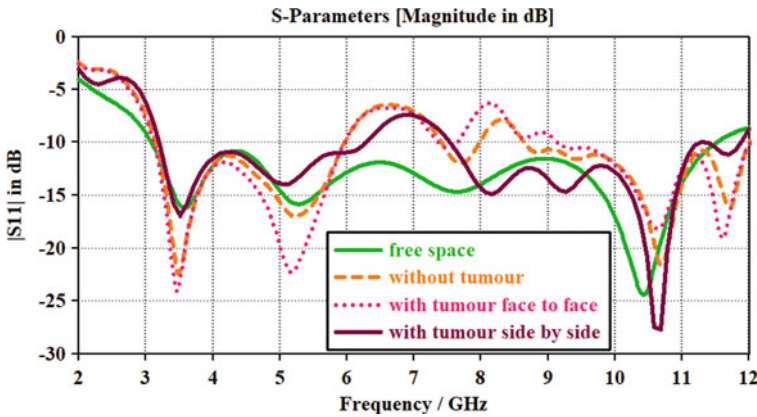


Fig. 5 Backscattered $|S_{11}|$ of the proposed antenna

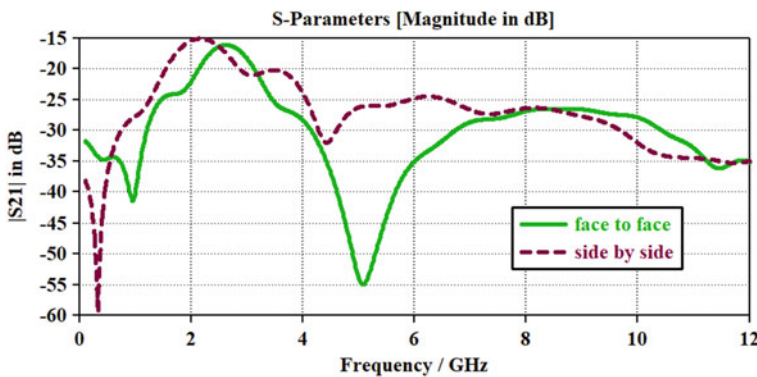


Fig. 6 Transmission coefficient ($|S_{21}|$) in side by side and face to face configuration

To justify the variation of power distribution in heterogeneous body tissues, electric field distribution is observed and shown in Fig. 7. It is clear that power absorption by the tissue layers is not uniform, and some power is absorbed by the tumour which creates deflections in the radiation patterns from different orientations. Though antenna has possessed stable frequency domain performance, it is required to study the time domain performance for MWI. The input signal and the received signal with and without tumour for the two evaluation setups (side by side and face to face) are represented in Fig. 8. Received waveform for both the setups is parallel to the input signal with small fluctuations.

Group delay in Fig. 9 represents the phase linearity of the transmitted signal over the channel. It is the negative derivative of the phase response with respect to frequency. Antenna has linear phase distribution except unconventionality at 6 GHz and 8 GHz in face to face and side by side mode, respectively. Value of group delay

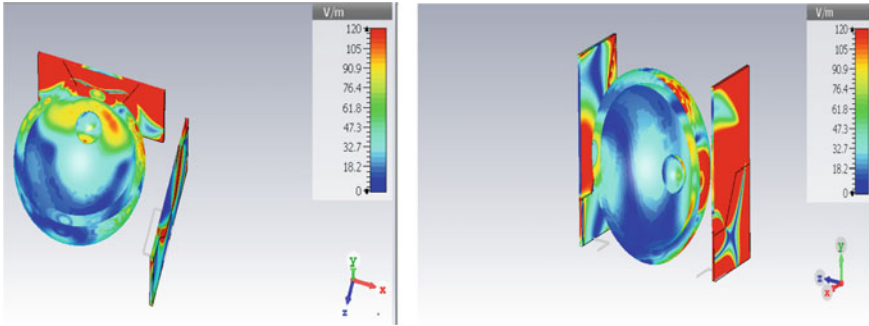


Fig. 7 Electric field distribution for side by side and face to face configuration

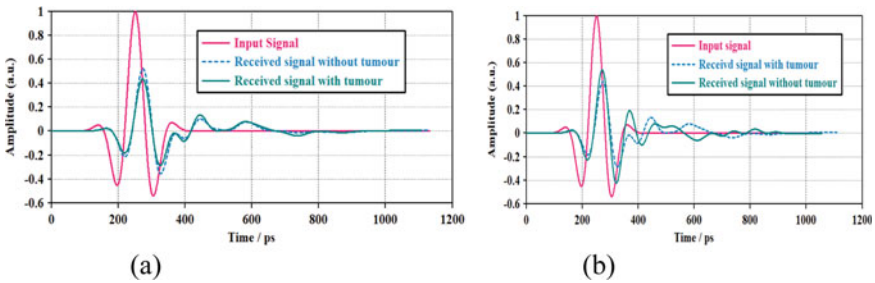


Fig. 8 Transmitted and received pulse **a** face to face setup, **b** side by side setup

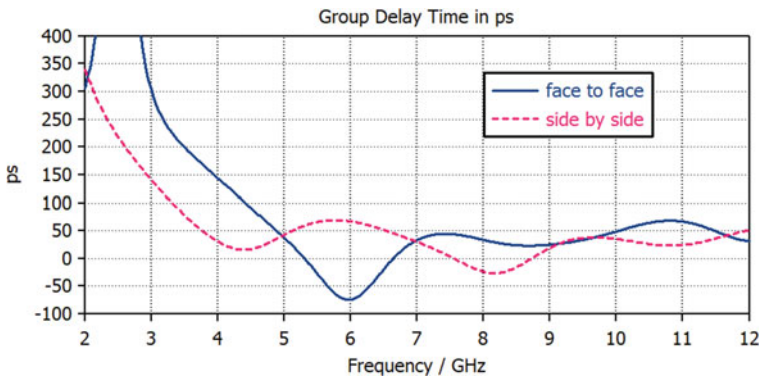


Fig. 9 Group delay of the proposed prototype

(in picoseconds) for both setups is lying in same range; due to this, both setups are convenient for MWI.

5 Conclusions

Configuration of a conventional microstrip patch antenna is modified to attain wide – 10 dB bandwidth at UWB spectrum. Defected ground structure with tapered slotting is merged to design the antenna. Equivalent tissue phantom is evaluating antenna performance for MWI. Antenna has good coupling with heterogeneous tissue. Variations in backscattered signal magnitude with varying electrical properties of tissue are observed that can be used to construct the images. Significant variation in $|S_{11}|$ parameter is obtained with maximum deviation of 6 dB at 5.2 GHz in face to face scenario and 8 dB deviation at 8.1 GHz and 10.8 GHz in side by side scenario. Nonuniform power distribution is evaluated through electric field distribution. Electrical properties of body tissue and depth of penetration of electromagnetic radiations are the functions of frequency. Breast tissue-air-antenna layers are the interference of low and high dielectric materials. It produces internal absorption and scattering of EM waves and causes the uneven power distribution. Thus, reflection coefficient of antenna is varied significantly due to the presence of unhealthy cells. Linearity of the timing signals is analyzed through group delay, received, and transmitted pulse.

References

1. A.M. Hassan, M. El-Shenawee, Review of electromagnetic techniques for breast cancer detection. *IEEE Rev. Biomed. Eng.* **4**, 103–118 (2011)
2. M. Klemm, J.A. Leendertz, D. Gibbins, I. Craddock, A. Preece, R. Benjamin, Microwave radar-based differential breast cancer imaging: Imaging in homogenous breast phantoms and low contrast scenarios. *IEEE Trans. Antennas Propag.* **58**, 2337–2344 (2010)
3. S. Kwon, S. Lee, Recent advances in microwave imaging for breast cancer detection. *Int. J. Biomed. Imaging* (2016)
4. M. Lazebnik, L. McCartney, D. Popovic, C.B. Watkins, M.J. Lindstrom, J. Harter et al., A large-scale study of the ultrawideband microwave dielectric properties of normal breast tissue obtained from reduction surgeries. *Phys. Med. Biol.* **52**, 2637 (2007)
5. S.K. Davis, B.D. Van Veen, S.C. Hagness, F. Kelcz, Breast tumour characterization based on ultrawideband microwave backscatter. *IEEE Trans. Biomed. Eng.* **55**, 237–246 (2008)
6. T. Sugitani, S. Kubota, A.X.X. Toya, T.A. Kikkawa, Compact 4×4 planar UWB antenna array for 3-D breast cancer detection. *IEEE Antennas Wirel. Propag. Lett.* **12**, 733–736 (2013)
7. B.A.J. Mohammed, A.M. Abbosh, P. Sharpe, Planar array of corrugated tapered slot antennas for ultrawideband biomedical microwave imaging system. *Int. J. RF Microwave Comput. Aided Eng.* **23**, 59–66 (2013)
8. S. Subramanian, B. Sundarambal, D. Nirmal, Investigation on simulation-based specific absorption rate in ultra-wideband antenna for breast cancer detection. *IEEE Sens. J.* **18**(24), 10002–10009 (2018)
9. J. Zhang, E.C. Fear, R.H. Johnston, Cross-Vivaldi antenna for breast tumor detection. *Microwave Opt. Technol. Lett.* **51**(2), 275–280 (2009)
10. M. Abbak, M. Çayören, I. Akduman, Microwave breast phantom measurements with a cavity-backed Vivaldi antenna. *IET Microwaves Antennas Propag.* **8**, 1127 (2014)
11. A. Molaei, M. Kaboli, M.S. Abrishamian, S.A. Mirtaehri, Dielectric lens balanced antipodal Vivaldi antenna with low cross-polarisation for ultra-wideband applications. *IET Microw. Antennas Propag.* **8**(14), 1137–1142 (2014)

12. N. Ojaroudi, M. Ojaroudi, N. Ghadimi, UWB omnidirectional square monopole antenna for use in circular cylindrical microwave imaging systems. *IEEE Antennas Wirel. Propag. Lett.* **11**, 1350–1353 (2012)
13. H. Bahrami, E. Porter, A. Santorelli, B. Gosselin, M. Popovic, L. Rusch, Flexible sixteen monopole antenna array for microwave breast cancer detection, in *Proceeding of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3775–3778 (2014)
14. H. Kanj, M. Popovic, A novel ultra-compact broadband antenna for microwave breast tumor detection. *Prog. Electromagn. Res.* **86**(9), 169–198 (2008)
15. V. Selvaraj, D. Baskaran, P.H. Rao, P. Srinivasan, R. Krishnan, Breast tissue tumor analysis using wideband antenna and microwave scattering. *IETE J. Res.* **23**, 1–1 (2018)
16. R.K. Amineh, A. Trehan, N.K. Nikolova, TEM horn antenna for ultra-wide band microwave breast imaging. *Prog. Electromagn. Res. B.* **13**(3), 59–74 (2009)
17. M. koutsoupidou, I.S. Karanasiou, C.G. Kakoyiannis, E. Groumpas, C. Conessa, N. Joachimowicz, et al., Evaluation of a tumor detection microwave system with a realistic breast phantom. *Microwave Opt. Technol. Lett.* **59**, 6–10 (2017)
18. M.M. Islam, M.T. Islam, M.R.I. Faruque, M. Samsuzzaman, N. Misran, H. Arshad, Microwave imaging sensor using compact metamaterial UWB antenna with a high correlation factor. *Materials* **8**, 4631–4651 (2015)
19. M. Mahmud, M.T. Islam, N. Misran, M.J. Singh, K. Mat, A negative index metamaterial to enhance the performance of miniaturized UWB antenna for microwave imaging applications. *Appl. Sci.* **7**(11), 1149 (2017)
20. M. Islam, M.T. Islam, M. Samsuzzaman, M.R. Faruque, N. Misran, M.F. Mansor, A miniaturized antenna with negative index metamaterial based on modified SRR and CLS unit cell for UWB microwave imaging applications. *Materials* **8**(2), 392–407 (2015)
21. C.A. Balanis, *Antenna theory analysis and design*, 3rd edn. (A John Wiley & Sons, Inc., Publication, 2005)
22. M.T. Islam, M. Samsuzzaman, M. Faruque, M.J. Singh, M. Islam, Microwave imaging-based breast tumor detection using compact wide slotted UWB patch antenna. *Optoelectron. Adv. Mater. Rapid Commun.* **1**(13), 448–457 (2019)
23. J.-D. Zhang, L. Zhu, Q.-S. Wu, N.-W. Liu, W. Wu, A compact microstrip-fed patch antenna with enhanced bandwidth and harmonic suppression. *IEEE Trans. Antennas Propag.* **64**(12), 5030–5037 (2016)

Impact of Beam Formation on 5G Mobile Communication Network in mmWave Frequency Band



Nallapalem Neeraj Srinivas, Yasaswini Vellisetty, and P. C. Jain

Abstract 5G is expected to support enhanced Mobile broadband (eMBB), massive machine-type communication (mMTC), and ultra-reliable low latency communication (uRLLC). 5G vision relies on a successful rollout of mmWave frequency technology. The millimeter (mm) Wave frequency band 6–100 GHz range provides more BW to achieve higher data rate but it supports 50–200 m short range only. 5G NR leverages the latest development in multi-input multi-output antennas (MIMO) and beam formation to minimize spectral efficiency and provide better quality of service for large number of users. Beam forming technology has been introduced in mmWave frequency band to compensate high path losses due to mmWave frequencies. Beam formation aims to generate different beams for different user equipment (UE) using an array of antennas elements at the base station (BS) to communicate between different users. The effective beams undergo minimum power loss and interferences and are aimed to direct towards the user equipment. In this paper, we have simulated different types of arrays and analyzed different beam contours that are generated from the arrays of antennas to transmit to the user equipment.

Keywords 5G · Antenna · Array · Beam formation · Beam shape · Directivity · Millimeter wave

N. N. Srinivas · Y. Vellisetty · P. C. Jain (✉)
Department of Electrical Engineering, School of Engineering, Shiv Nadar University, Greater Noida, UP, India
e-mail: premchand.jain@snu.edu.in

N. N. Srinivas
e-mail: ns620@snu.edu.in

Y. Vellisetty
e-mail: yv294@snu.edu.in

1 Introduction

Mobile communication network has successfully connected majority of global population. The vision of 5G focuses on three primary cases: enhanced mobile broadband (eMBB), ultra-reliable and low latency communication (uRLLC), and massive machine-type communication (mMTC). The eMBB needs support for extremely very high data rates to support 4 k/8 k resolution screens, AR/VR, flash download/upload, and congested environments such as stadiums and airports. The mMTC requires much lower throughputs but supports millions of devices with very low energy consumption. The Internet of things (IoT) requires support for massive number of devices, growing exponentially beyond traditional cellular demand. The uRLLC will support industrial control, remote robotics, remote surgery/health, and autonomous car driving. The 5G will provide 10 times high data rates ($>10\text{Gbps}$), 10 times lower latency (1 ms data plane, 20 ms control plane), 10 times higher efficiency compared to 4G networks, and one million per square km devices connection with reliability 99.9999% and enhanced battery life of 10 years for IoT devices.

5G-NR (New Radio) uses two frequency ranges: one FR1 covers sub-6 GHz frequency band including 3 GHz band while other FR2 includes mmWave frequency bands 24, 28, 37, and 39 GHz. FR1 provides 100 MHz BW while FR2 provides 400 MHz BW. The 5G will use millimeter (mm) Wave frequency band to get more BW to achieve very high data rate (20Gbps). However, mmWave frequencies signal attenuate very fast and do not propagate well through obstacles such as walls. The mmWave frequency supports 50–200 m short range only. This will need highly directional beam formation using array of antennas to compensate higher path losses at mmWave frequency band. The beam formation concentrates the signal into beam pointed in the direction of the user equipment (UE) rather than radiating in all the directions as done by omnidirectional antenna. In array of antennas, multiple antennas are spaced by $\lambda/4$ to $\lambda/2$ to avoid inter-antenna interference. Small cells at mmWave frequency band provide large density in smaller ranges mitigating the limitation of path loss, diffraction, and penetration. 5G-NR comprises massive MIMO, mmWave spectrum for mobile and fixed wireless access, and mmWave frequency band, beam formation, and beam tracking. There is a significant increase in cost to upgrade 5G infrastructure and handsets. The 3G network required 4–5 BS per sq. km, 4G network 8–10, and 5G network reaches to more than 100 sites per sq. km. Backhaul to connect 5G BS could be either optical fiber cables or wireless. Manufacturers and operators will tradeoff between cost and complexity of adding 5G RF components.

In a few years, **beam formation** is moved out from research environment into commercial deployment, first in 4G-LTE networks and now in 5G deployments. 5G vision relies on a successful rollout of mmWave frequency technology. The mmWave frequency band suffers heavy path loss due to very high-frequency band and do not propagate well through obstacles such as walls. To compensate for the path loss, high gain antenna elements are required. These antenna elements with input RF signal create narrow beam signals. This approach strengthens the signal

and reduces interference from other user's signal. The increased SNR due to beam formation increases range for both outdoor and indoor coverage. A large number of radiating antenna elements are required to focus beam narrower to transmit all available power in certain direction instead of wasting power in many directions. Each antenna element is fed with same RF signal but phase and magnitude of signal fed to each antenna element are adjusted. It needs full control of magnitude and phase of signal received by every antenna element. Narrow beam formation for a user needs precise real-time channel state information (CSI) from UE to customize beam. The UE uses synchronization signals and system information to establish connection with required beam. The random access response and system information help to refine the beam. Beam refinement is a continuous process during the user session because of the UE movement. The phase and magnitude of antenna element's signal are controlled digitally in nsec. period to achieve faster beam steering of the required beam. Full dimension (FD)-MIMO places large number of active antenna elements in two-dimensional (2D) grid at base station. FD-MIMO can support 3D-beam forming algorithm which exploits the elevation and azimuth dimensions. It creates highly directional beams that can be redirected to specific location or device [1, 2].

2 Related Work

Reference [2] describes five technologies, namely mmWave, massive MIMO, smart devices, device-centric architectures, and machine-to-machine communication, which could lead to both architectural and component disruptive design changes. Reference [3] focuses 5G era in which shift toward network efficiency with 5G systems is based on dense HetNet architecture. However, HetNet incorporates set of frequency bands including macrocells in licensed band (LTE) and small cells in unlicensed bands (Wi-Fi). New higher frequency band (mmWave) may be deployed in small cell to enable ultrahigh data rate services. Reference [4] challenges to rethink relationship between energy, directivity, and spectral efficiency. Redesign of these should be important part of 5G research. It covers the deployment of massive MIMO in the form of irregular antenna arrays where the antenna array elements are embedded into an array geometry. YouTube videos in [6] and [7] were helpful for the basic understanding of antennas and beam forming in 5G technology. Reference [8] helped us in implementing and understanding the concept of antenna arrays in detail.

3 Methodology

High gain antennas are required to compensate for the path loss in mmWave frequency band. The most efficient way to achieve larger gains is by using an array of antenna elements. Antenna arrays consist of two or more antennas where the gain of all these

antennas are combined for an increased and better performance. The antenna arrays also help in generating different kinds of beams, which are being analyzed for the uniform linear arrays (ULA) and uniform rectangular arrays (URA) in this paper. The aim is to increase the directivity and hence to reduce the path losses. Beams formed by the antenna arrays can be directed toward the user equipment. Location coordinates can be helpful to track the user with reference to the base station.

3.1 Uniform Linear Array

In a uniform linear array (ULA), the array elements are uniformly spaced along a straight line. There should be minimum of two elements to form an array. The uniform array is formed by using identical antenna elements with equal RF signal magnitudes and with a progressive phase. The elements of the antenna are arranged in linear pattern like a matrix of $N \times 1$ or $1 \times N$, and they are uniformly spaced from each other. The minimum number of such antenna elements could be 2×1 or 1×2 . The different shapes of the beams can be obtained from these arrays by varying the factors like the spacing of the elements, distance between the elements of antenna, the number of antenna elements, the geometry of the elements of the array, and the phase of the signal in each antenna element. In a uniform linear array, the antenna elements are fed with a RF signal of equal amplitude and equal phase shift between the elements. The RF signal of any antenna array can be concentrated in any arbitrary direction by changing the phase of each antenna element. This type of antenna array is called phased array antenna. This kind of increased directivity helps in increasing the range and reducing the interference. By using the phased array techniques, in this paper we observed that the power density of the beam is not uniformly distributed. The shift in beam pattern is observed by applying individual antenna phase shifts from -90 to 90° . The shape of the beam formed resulted in main lobe which helps to transmit most of the RF signal, and the side lobes which provide interference to the main lobe.

3.2 Uniform Rectangular Array

The array elements are distributed in a y - z plane with the beam direction along the positive x -axis. The spacing between the elements is uniform for uniform rectangular array (URA). The elements of the antenna are arranged in a matrix of $M \times N$, and they are uniformly spaced from each other. We have observed different shapes of the beam of URA arrays by varying the factors like the spacing of the elements, distance between the elements of antenna, the number of antenna elements, the geometric arrangement of the elements of the array, values of M and N , and the phase of each antenna element. The number of rows (M) and columns (N) may be equal in order to form a perfect main lobe. If the M and N are not equal then depending

on the value of M and N , we observed flat beams in the x - and y -directions. All the antenna array responses are simulated and plotted using MATLAB. We have compared directivity and power level of various beams formed. Directivity is the measure of the concentration of an antenna’s radiation pattern in a particular direction. Directivity is a function of the radiation pattern of the antenna where it is assumed that all the power applied to the antenna array is radiated in a particular direction. As the concentration of the signal increases, the directivity of the beam radiated is also increased. Antenna gain is the product of antenna efficiency and directivity. It is expressed in decibels (dB). Uniform rectangular array is better for transmitting the beam to the user, and these arrays are used widely in the 5G communications, which operates in mmWave frequency band around 24–39 GHz. In this frequency range, the signals are blocked by different obstacles which result in poor coverage and reduced data speeds for both uplink and downlink.

4 Results

The simulation of uniform linear array (ULA) with 2D- and 3D-plots and uniform rectangular array (URA) with 3D-plots have been carried out using MATLAB.

4.1 2D-Simulation of ULA

The 2D-polar plots and the magnitude plots for different array elements with different sizes of antennas are observed. The main lobe of the two-dimensional polar plot is shown in Fig. 1. It shows 2D-polar plot for 4 linear antennas and 8 linear antennas with different steering angles $[-90^{\circ}$ – $90^{\circ}]$. Beam width is wider with 4 antenna elements and gets narrow down with 8 elements which effectively increases the range as we increase the number of antennas elements.

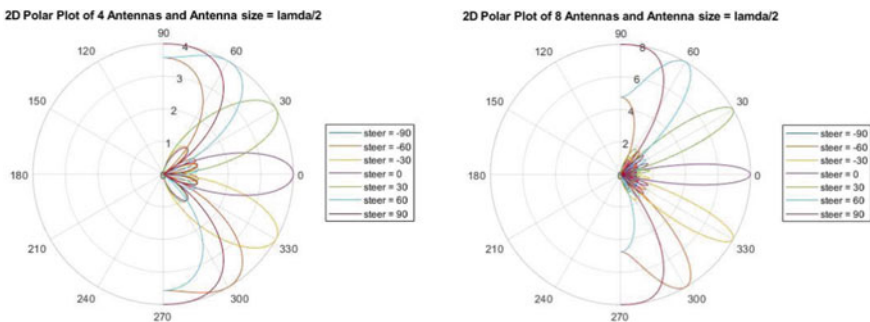


Fig. 1 Two-dimensional polar plot for the 4 and 8 antennas when the patches are phased to direct the main beam toward (0° , 30° , 60° , 90° , -30° , -60° , -90°)

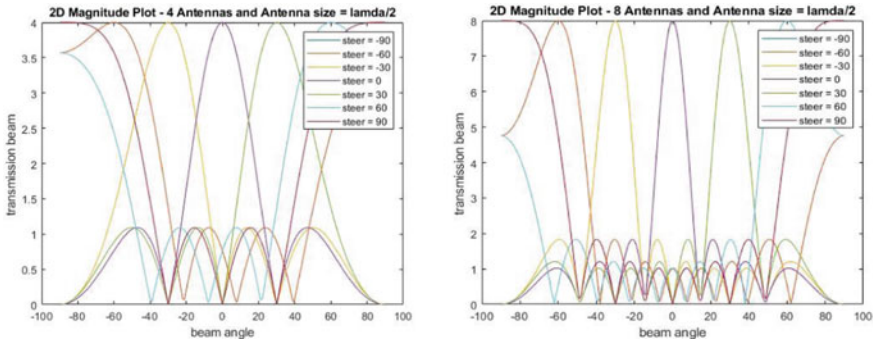


Fig. 2 Two-dimensional **magnitude** plot for the 4 and 8 antennas when the patches are phased to direct the main beam toward (0° , 30° , 60° , 90° , -30° , -60° , -90°)

From the magnitude plot in Fig. 2, we can observe the transmission beams with different steering angles. Here also it is observed that as the antenna elements are increasing, the beam width is decreasing and range of the beam is increasing.

4.2 3D-Simulation of ULA

In Figs. 3, 4, and 5, it can be seen that the maximum amount of normalized power is represented by yellow color of the 3D-plot, and blue color represents the minimum normalized power. The shapes of the transmitted beams in 3D-plot are inappropriate for the user device. If we direct this beam to user device, it leads to a total wastage of power at the 28 GHz mmWave frequency band because the beam formation is not directional as shown in Figs. 3, 4, and 5. Figures 3, 4, and 5 also show the plots for different steering angles. We further worked with the uniform rectangular array (URA) simulations to obtain a proper desired beam for the user device in the 5G environment.

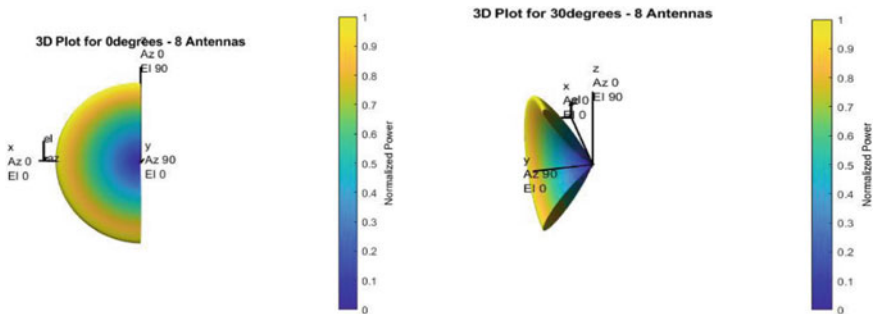


Fig. 3 Three-dimensional **power** plot for the 8 antennas when the patches are phased to direct the main beam toward (0° , 30°)

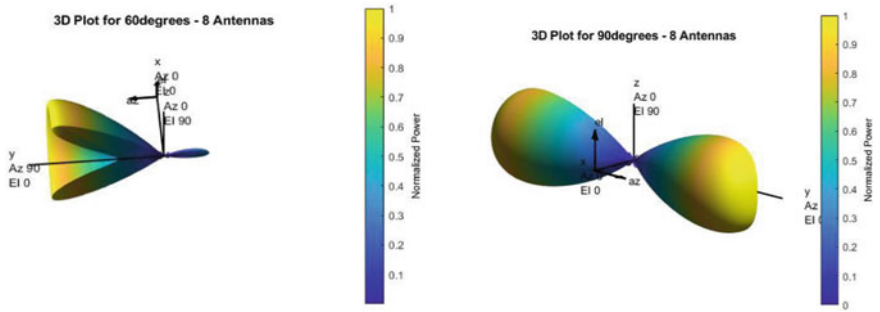


Fig. 4 Three-dimensional power plot for the 8 antennas when the patches are phased to direct the main beam toward (60°, 90°)

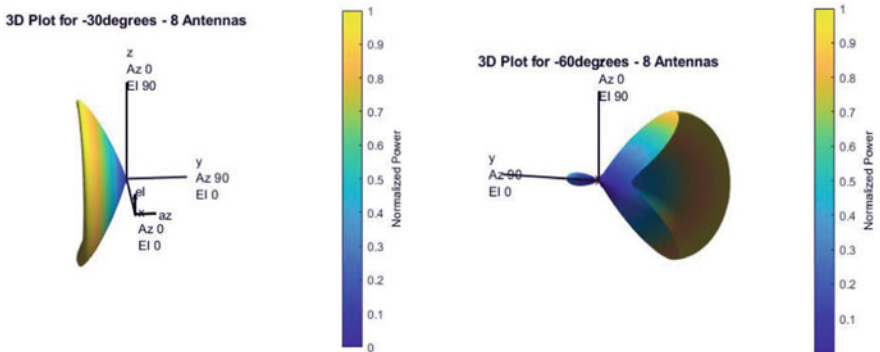


Fig. 5 Three-dimensional power plot for the 8 antennas when the patches are phased to direct the main beam toward (-30°, -60°)

4.3 3D-Simulation of URA

Beam formation using uniform rectangular array has been implemented to obtain 3D-plots considering directivity of the beam and the power of the beam. By increasing the number of antennas, we can observe that the beam width is decreasing. Depending on the value of M and N in the matrix $M \times N$, we can observe different beam shapes. If the M and N are equal, we observe a minimum number of side lobes with a desired main lobe directed to the user equipment in the 5G communication. Directivity and power patterns shown in Figs. 6 and 7 are the normalized graphs. The range is normalized with absolute distances. The range and the power of the different antenna arrays are simulated and represented by different colors for good visualization. Figures 6 and 7 show the magnitude and power plots for 4×4 and 8×8 rectangular antenna arrays. As the number of antennas increases, we observe more directivity and hence more range. The pattern function from MATLAB is used for

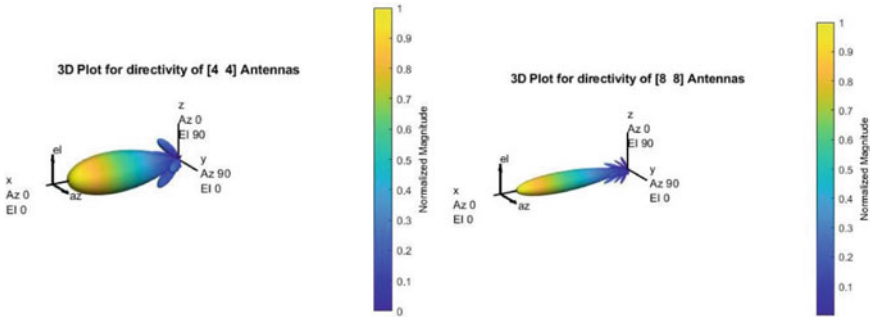


Fig. 6 Three-dimensional normalized **magnitude** plot for the 4×4 array and 8×8 array when the patches are phased to direct the main beam toward (0°)

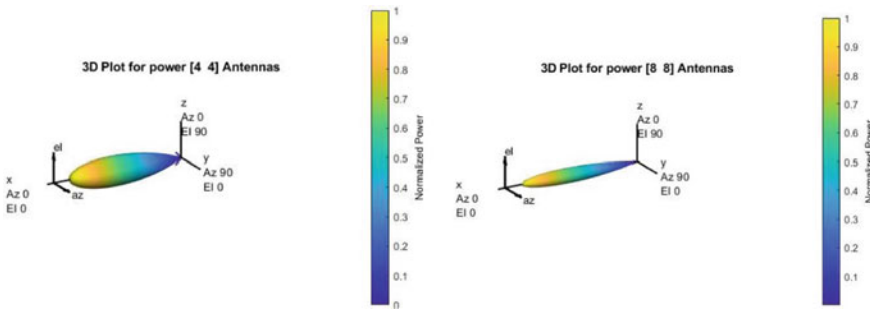


Fig. 7 Three-dimensional **power** plot for the 4×4 array and 8×8 array when the patches are phased to direct the main beam toward (0°)

simulation. It generates a normalized graph by giving the transmitted beam range, power, and directivity.

If the value of M and N in the $M \times N$ rectangular array is different, then we observe that the main lobe has flattened with respect to a smaller number of antennas. If we increase the number of antenna elements in M by fixing the N , we observe further flatness in the beam. Using this type of array, we can reduce the transmitted power of beam and range by changing the M and N antenna elements according to the requirement of the user equipment. Simulation for both normalized magnitude and directivity was done and represented with different colors (yellow with maximum magnitude and blue with minimum magnitude). Figure 8 shows the flatness of the beam for 4×2 and 8×2 rectangular array. We observe that the flatness is more with 8×2 array compared to 4×2 array.

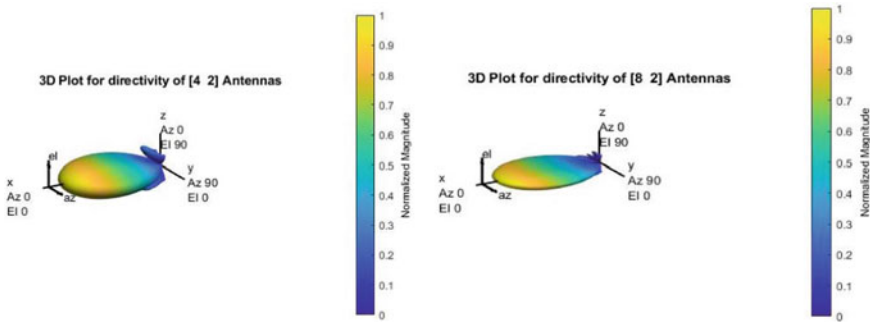


Fig. 8 Three-dimensional normalized **magnitude** plot for the 4×2 array and 8×2 array when the patches are phased to direct the main beam toward (0°)

5 Conclusions

In this paper, we have described important features of 5G and studied how the beam formation at mmWave frequency band can reduce the path loss and interferences. We have discussed different phased arrays in space to implement. We have studied different beam shapes using different kinds of antenna arrays, and then analyzed, and molded according to the requirements of the desired user. The uniform linear and rectangular arrays have been studied, simulated, and analyzed different contours of beams generated by the different array of antennas to be transmitted to the user equipment and found that the effective beams undergo minimum path loss and interferences.

Acknowledgements The authors are very much thankful to Dr. Jitendra Prajapati, EE Dept. to help in the simulation using MATLAB. The authors are also grateful to Prof. Dinkar Prasad, Head, EE Dept., and Associate Dean, School of Engineering, for providing necessary resources and infrastructure to complete this project, and to Prof. S. Sen, Dean, School of Engg., Shiv Nadar University, G. Noida (UP), for their encouragement, and permission to publish this paper.

References

1. M. Safi, et al., 5G: A tutorial overview of standards, trials, challenges, deployment, and practice. *IEEE J. Sel. Areas Commun.* **35**, 1201–1221 (2017)
2. F. Boccardi, F.R.W.J. Heath, A. Lozano, T.L. Marzetta, P. Popovski, Five disruptive technology directions for 5G. *IEEE Commun. Mag.* **52**, 74–80 (2014)
3. B. Bangerter, S. Talwar, R. Arefi, K. Stewart, Networks and devices for the 5G era. *IEEE Commun. Mag.* **52**, 90–96 (2014)
4. S. Chen, J. Zhao, The requirements, challenges, and technologies for 5G of terrestrial mobile telecommunication. *IEEE Commun. Mag.* **52**, 36–43 (2014)
5. International Telecommunication Union (ITU), *ICT Facts and Figures 2017*
6. Use of mm Wavelengths & Beam Forming with 5G [Online]. <https://www.youtube.com/watch?v=6li8CZmrg84>

7. Basics of Antennas and Beam-forming—Massive MIMO Networks [Online]. <https://www.youtube.com/watch?v=xGkyZw98Tug>
8. 5G/NR beam Management [Online]. https://www.sharetechnote.com/html/5G/5G_Phy_BeamManagement.html
9. Antenna Characteristics [Online]. <https://www.waves.utoronto.ca/prof/svhum/ece422/notes/06-antennachar.pdf>
10. Antenna Directivity [Online]. <https://www.everythingrf.com/community/what-is-antenna-directivity>
11. Research articles [Online]. https://futurenetworks.ieee.org/images/files/Tech_Focus_Articles/PDFs/Towards-5GNetwork-Slicing-FINAL.pdf

Multi-transform 2D DCT Architecture Supporting AVC and HEVC Video Codec



K. Phani Raghavendra Sai and I. Mamatha

Abstract In this paper, an area-efficient multi-transform architecture supporting transforms used in most popular video codecs like High Efficiency Video Coding (HEVC) and Advance Video Coding (AVC) is proposed. An eight-point integer DCT is implemented using two four-point integer DCTs. A three-stage pipelined and parallel multiplier-less architecture is designed using shift and add method. Proposed 1D DCT architecture uses 50% less resources as compared to other approach in literature and is capable of producing eight outputs for every clock cycles after the initial latency. Architecture is scalable to 2D DCT, and higher order DCTs can be computed either by reusing 1D structure or by duplicating the 1D structure. Proposed 1D and 2D DCT structures are simulated using Xilinx ISE and MATLAB tools and validated for the results. Further, the design is synthesized on Virtex-6 FPGA board and consumes 38% less area when compared with the standalone architectures. Proposed structure is found to be efficient in terms of resource utilization while comparing with architectures reported in literature.

Keywords Discrete Cosine Transform (DCT) · High Efficiency Video Coding (HEVC) · H.265 · Advance Video Coding (AVC) · H.264 · Multi-transform · Video codec

1 Introduction

Multimedia is an interactive media which offers several ways to powerfully represent the user's details. It offers an interface between digital knowledge and users. Education, training, reference materials, company presentations, advertisement, and documentaries are some of the industries where multimedia is used extensively.

K. Phani Raghavendra Sai · I. Mamatha (✉)

Department of Electrical and Electronics Engineering, Amrita School of Engineering, Bengaluru, India

Amrita Vishwa Vidyapeetham, Bengaluru 560035, India

Video encoding is an essential operation performed in multimedia applications in order to modify/compress the video data and hence, can reduce storage space and bandwidth with enhanced compatibility. Most commonly used video coding standards are advance video coding (AVC) and high efficiency video coding (HEVC). AVC standard (also referred as H.264) can encode or decode the video without compromising the quality of an image and also reduce the size of the file by 80% compared with MPEG and 50% compared with the previous standards. The standard is mostly used in high definition DVDs, HDTV, and mobile television broadcasting. This standard has an advantage of good quality video compression output and good flexibility in transmitting and preserving the video.

As compared to AVC, HEVC has improved coding efficiency and is successor of AVC. HEVC has more compression ratio without compromising the quality of image and gives better video quality with the same bitrates compared to AVC. HEVC has the capability of supporting resolutions up to $8192 * 4320$, including 8K UHD and promises a 50% storage reduction for a video file and needs almost $10\times$ more computing power. There has been few hardware architectures for implementing the integer DCT for HEVC, AVC, and multi-transform proposed for the real-time implementation. Meher et al. [1, 2] proposed an efficient integer DCT architecture for HEVC which is reusable and can implement different DCT lengths like 4, 8, 16, and 32. An area and power-efficient DCT architecture for image compression using signum function is proposed in [3, 4] which had reduced computational complexity. Dias et al. proposed a unified architecture for fast and efficient computation of 2D transforms for the most used video standards [5, 6]. Wahid et al. proposed a resource shared multi-transform architecture for multiple video codecs using delta mapping technique [7, 8]. Mohankumar et al. proposed a DCT architecture used for inserting the water marking for the video streaming during the compression technique [9]. A five-stage pipelined structure of 1D DCT by Megalingam et al. achieved low power and higher speed of implementation [10]. Shyam et al. proposed an architecture to increase the quality of compressed image based on non-zeroing bit truncation method [11]. Distributed arithmetic-based (DA) eight-point multi-transform architectures are proposed in [12, 13] although required less resources but suffers from large computation time. Another multi-transform architecture proposed in [14] supports DCT and DFT which uses cyclic convolution representation and its systolic implementation. Chang et al. proposed a hardware shared architecture supporting different sizes of forward and inverse transforms for multiple video codecs [15]. Although many of these approaches focused on reducing hardware, there is still need for reduction of area as demanded by low-power portable devices.

In this work, the transforms of AVC and HEVC are chosen as these are the popular standards that are in use. The proposed work is mainly focused on implementing the four and eight-point DCT architecture where eight-point architecture is implemented using the four-point structure to reduce the complexity. Rest of the paper is divided into three sections: Section 2 explains the proposed technique for computing DCT and mapping the operations on to an architecture developed. Section 3 discusses about the results and performance of the proposed architecture, and Section 4 concludes the work.

2 Methodology

Discrete cosine transform is widely used in different video codec standards and is available in many forms. The type-2 DCT is extensively used in block-based image and video coding and is represented as follows:

$$X(k) = \sum_{n=0}^{N-1} x(n) \cos \left[\frac{\pi}{n} \left(n + \frac{1}{2} \right) k \right] \quad \text{where, } k = 0, 1 \dots N - s1 \quad (1)$$

The kernel matrices used for four and eight-point integer DCT for HEVC and AVC are given in [1, 5]. In this work, same kernel matrix is used, and the architecture is developed.

The expression for computing 1D integer DCT of an input signal is given by

$$Y = T * X \quad \text{where, } Y = [Y0, Y1, Y2, Y3], X = [X0, X1, X2, X3] \quad (2)$$

T = Transform co-efficient matrix, also referred as kernel matrix.

The matrix form for the four-point 1D DCT is as shown in (3)

$$\begin{bmatrix} Y0 \\ Y1 \\ Y2 \\ Y3 \end{bmatrix} = \begin{bmatrix} a & a & a & a \\ f & g & -g & -f \\ a & -a & -a & a \\ g & -f & f & -g \end{bmatrix} * \begin{bmatrix} X0 \\ X1 \\ X2 \\ X3 \end{bmatrix} \quad (3)$$

Higher order transforms can be computed using lower order transforms. The work proposed discusses eight-point DCT architecture using four-point DCT architecture. The overall computation is divided into three stages and are mapped to computing elements in a three-stage pipelined architecture. The computations that are performed in the three-stage pipelined 1D DCT structure using the lower order DCT's are shown in Table 1, and the proposed pipelined architecture is as shown in Figure 1a where the three stages are referred as:(i) Input adder unit (IAU), (ii) Shift-add unit (SAU), (iii) Output adder unit (OAU).

2.1 Advance Video Coding (AVC)

Input adder unit (IAU) is the first stage of the transform architecture where the additions and subtractions of the initial stage is performed. Figure 1b shows the internal architecture of the input adder unit (IAU). This is similar for the second four-point DCT with the inputs of X3, X5, X2, and X4. The computations that are performed in the IAU are as shown in Table 1.

Table 1 Computations performed in 3-stages for 1D eight-point DCT using four-point DCT for AVC

Stage-1	Stage-2	Stage-3
First four-point DCT $Z1 = X0 + X7;$ $Z2 = X1 + X6;$ $Z3 = X0 - X7;$ $Z4 = X1 - X6;$ $P = Z1 + Z2;$ $Q = Z1 - Z2;$	$m0 = 8 * P;$ $m1 = 8 * Z1 + 4 * Z2;$ $m2 = 8 * Q;$ $m3 = 4 * Z1 - 8 * Z2;$ $m4 = 12 * Z3; m8 = 12 * Z4;$ $m5 = 10 * Z3; m9 = 10 * Z4;$ $m6 = 6 * Z3; m10 = 6 * Z4;$ $m7 = 3 * Z3; m11 = 3 * Z4;$	$Y0 = m0 + s0;$ $Y2 = m1 - s1;$ $Y4 = m2 + s2;$ $Y6 = m3 - s3;$
Second four-point DCT $W1 = X3 + X4;$ $W2 = X2 + X5;$ $W3 = X3 - X4;$ $W4 = X2 - X5;$ $A = Z1 + Z2;$ $B = Z1 - Z2;$	$s0 = 8 * A;$ $s1 = 8 * W1 + 4 * W2;$ $s2 = 8 * B;$ $s3 = 4 * W1 - 8 * W2;$ $s4 = 12 * W3; s8 = 12 * W4;$ $s5 = 10 * W3; s9 = 10 * W4;$ $s6 = 6 * W3; s10 = 6 * W4;$ $s7 = 3 * W3; s11 = 3 * W4;$	$Y1 = m4 + m9 + s10 + s7;$ $Y3 = m5 - m11 - s8 - s6;$ $Y5 = m6 - m8 + s11 + s5;$ $Y7 = m7 - m10 + s9 - s4;$

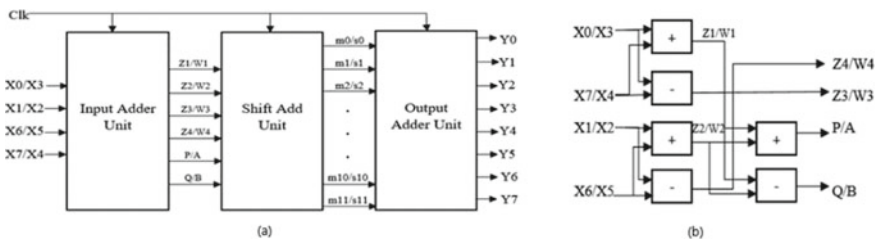


Fig. 1 a Architecture of proposed 1D DCT. b Input adder unit (IAU)

The output of IAU is fed as input to the shift-add unit to get the required product terms at the output of shift-add unit.

As multipliers are more costly; a multiplier-less architecture to achieve the required product terms using shift-add technique [1] is used. The shift-add technique is illustrated in Table 2 where it can be seen that $s0 = 8 * P = (P \ll 3)$. This means if the input P is left shifted by three units, it is equivalent of multiplying by $2^3 = 8$. Similar way, all other multiplications are performed. The structure for performing shift-add operation is shown in Figure 2. The shift-add unit is divided into two parts as even and odd where even SAU produces the product terms required

Table 2 Three-stage computation of 1D eight-point DCT using four-point DCT for HEVC

Stage-1	Stage-2	Stage-3
First four-point DCT $Z1 = X0 + X7;$ $Z2 = X1 + X6;$ $Z3 = X0 - X7;$ $Z4 = X1 - X6;$ $P = Z1 + Z2;$ $Q = Z1 - Z2;$	$m0 = 64*P;$ $m1 = 83*Z1 + 36*Z2;$ $m2 = 64*Q;$ $m3 = 36*Z1 - 83*Z2;$ $m4 = 89*Z3; m8 = 89*Z4;$ $m5 = 75*Z3; m9 = 75*Z4;$ $m6 = 50*Z3; m10 = 50*Z4;$ $m7 = 18*Z3; m11 = 18*Z4;$	$Y0 = m0 + s0;$ $Y2 = m1 - s1;$ $Y4 = m2 + s2;$ $Y6 = m3 - s3;$
Second four-point DCT $W1 = X3 + X4;$ $W2 = X2 + X5;$ $W3 = X3 - X4;$ $W4 = X2 - X5;$ $A = Z1 + Z2;$ $B = Z1 - Z2;$	$s0 = 64*A;$ $s1 = 83*W1 + 36*W2;$ $s2 = 64*B;$ $s3 = 36*W1 - 83*W2;$ $s4 = 89*W3; s8 = 89*W4;$ $s5 = 75*W3; s9 = 75*W4;$ $s6 = 50*W3; s10 = 50*W4;$ $s7 = 18*W3; s11 = 18*W4;$	$Y1 = m4 + m9 + s10 + s7;$ $Y3 = m5 - m11 - s8 - s6;$ $Y5 = m6 - m8 + s11 + s5;$ $Y7 = m7 - m10 + s9 - s4;$

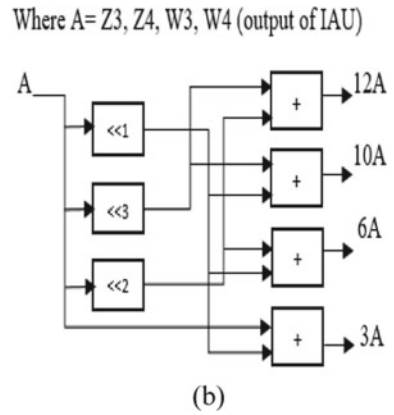
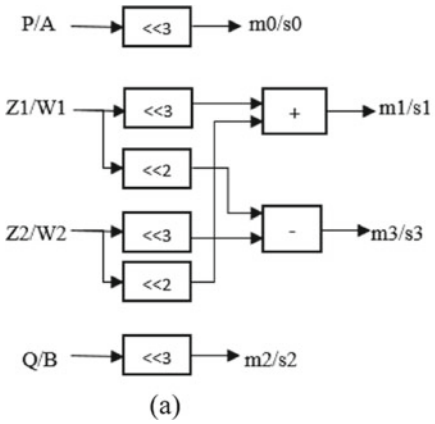


Fig. 2 Architecture of shift-add unit (SAU). **a** Even SAU. **b** Odd SAU

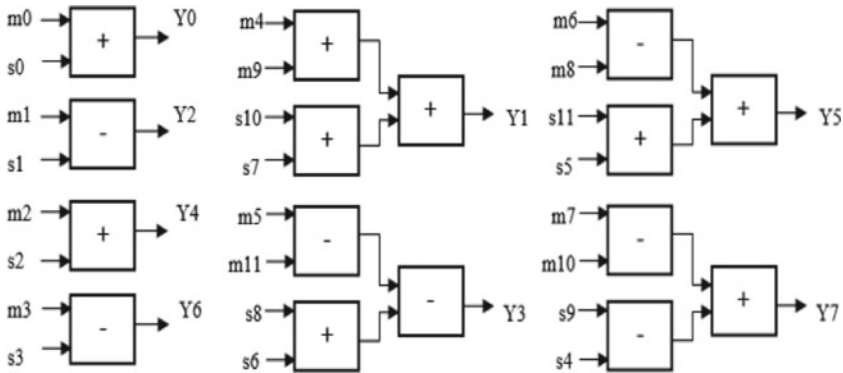


Fig. 3 Architecture of output adder unit (OAU)

for the computation of even outputs (Y_0, Y_2, Y_4, Y_6), and odd SAU computes the product terms required for the computation of odd outputs (Y_1, Y_3, Y_5, Y_7). The final product terms that are obtained at the end of the shift-add unit are shown as stage-2 as represented in Table 1. The final stage of the architecture is represented as output adder unit as shown in Figure 3 which is computing the operations in stage-3 of Table 1.

2.2 High Efficiency Video Coding (HEVC)

The respective computations in HEVC codec are computed in three-stages and are as shown in Table 2. From Tables 1 and 2, we can observe that the stages 1 and 3 are same for both the video codecs. The shift-add technique performed in stage-2 is illustrated in Table 2 where it can be seen that $s_0 = 64 * P = (P \ll 6)$. Similar way, all other multiplications are performed. Hence, only the structure for stage-2 is derived as shown in Figure 4.

2.3 Multi-transform Architecture

The need of multi-transform architecture is to use single architecture instead of using two standalone architectures for AVC and HEVC. As the first and third stage computations are similar, these units are shared leading to reduction in hardware cost with a minimal control overhead. Hence, a common IAU and OAU units are used. In addition, a part of shift-add unit is also common. The SAU of the proposed multi-transform architecture is as depicted in Figure 5 where the output of AVC is used to obtain product terms used for HEVC as shown in the shaded part of Figure 5.

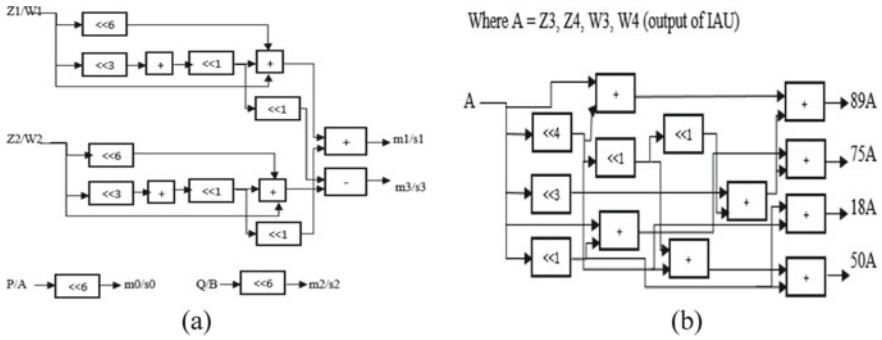


Fig. 4 Architecture of shift-add unit (SAU) for HEVC. a Even SAU. b Odd SAU

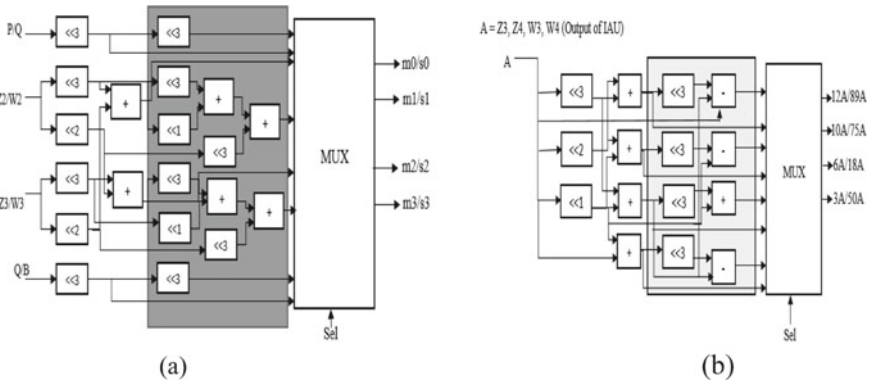


Fig. 5 Multi-transform architecture of SAU. a Even SAU. b Odd SAU

The overall structure for the eight-point 1D DCT using the four-point 1D DCT is represented in Figure 6a, where the two four-point DCTs are combined to get eight point.

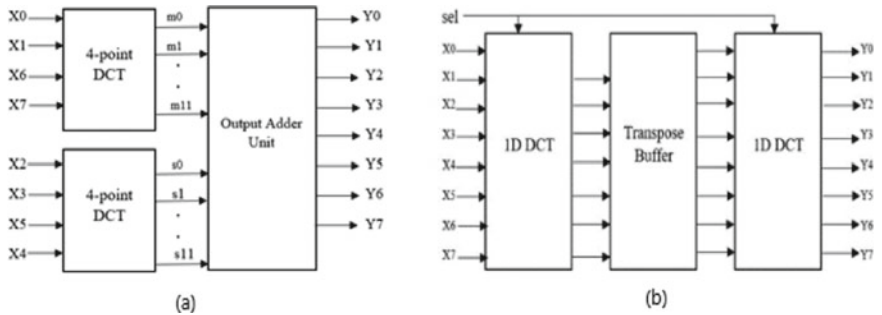


Fig. 6 a Eight-point 1D DCT using four-point DCT. b 2D DCT architecture

2.4 2D Integer DCT

The expression for computing 2D integer DCT of an input signal is given by $Y = T * X * T'$ where T —transform co-efficient matrix.

$$X = [X0, X1, X2 \dots X15]$$

$$Y = [Y0, Y1, Y2 \dots Y15]$$

The matrix form for the four-point 2D DCT is shown in (4)

$$\begin{bmatrix} Y0 & Y4 & Y8 & Y12 \\ Y1 & Y5 & Y9 & Y13 \\ Y2 & Y6 & Y10 & Y14 \\ Y3 & Y7 & Y11 & Y15 \end{bmatrix} = \begin{bmatrix} a & a & a & a \\ f & g & -g & -f \\ a & -a & -a & a \\ g & -f & f & -g \end{bmatrix} * \begin{bmatrix} X0 & X4 & X8 & X12 \\ X1 & X5 & X9 & X13 \\ X2 & X6 & X10 & X14 \\ X3 & X7 & X11 & X15 \end{bmatrix} * \begin{bmatrix} a & f & a & g \\ a & g & -a & -f \\ a & -g & -a & f \\ a & -f & a & -g \end{bmatrix} \tag{4}$$

The 2D integer DCT structure is represented in Figure 6b contains two 1D DCT's and a transpose buffer. The output of first 1D DCT is transposed and is fed as input to the second 1D DCT to get the final 2D DCT output.

2.5 Transpose Buffer

The output of the first 1D DCT is stored in the transpose buffer column wise. When the transpose buffer is filled, the input for the second 1D DCT is fetched as row wise from the transpose buffer. The output of the transpose buffer is fed to the second 1D DCT structure. A typical four-point transpose buffer structure as in [6] is used in this work. For an N-point DCT, total of N cycles are required to fill the transpose buffer. So, there is a delay of N cycles before feeding the input for the second 1D DCT. For the first N cycles, if the transpose buffer is filled row wise, then after completely filling transpose buffer, the input to the second 1D DCT is fed as column wise from the transpose buffer.

Hence, for every four cycles, the inputs stored in the transpose buffer are stored as column wise and row wise to avoid the delay in feeding the input to the second 1D DCT.

Table 3 Computational complexity of standalone and multi-transform architectures

Resource	HEVC	AVC	Total	Multi-transform
Adders	42	32	74	43
Shifts	32	18	50	42

2.6 Computational Complexity

The total number of computations that are performed in the standalone architectures of HEVC, AVC, and multi-transform architecture is shown in Table 3. Standalone architecture needs a total of 74 adders and 50 shift operations. Instead, if a multi-transform architecture is used, only 43 adders and 42 shift operations are required to achieve the final output, resulting in 41% saving in adders and 16% saving in shift operations.

3 Results and Discussion

The proposed standalone and multi-transform architecture is described using Verilog HDL and simulated and synthesized using Xilinx Vivado tool. The word length considered for input and coefficients is eight-bits and that for output is 16 bits. An input image in JPEG format is used as input for the simulation tool. The image is converted into binary format using MATLAB and given as input to the design as one dimension of eight-pixel values per clock cycle. The output obtained is validated by comparing it with output from MATLAB.

The architecture designed is synthesized using Xilinx Vivado 14 tool with the target device as Artix-7 FPGA board. The total number of resources used in the standalone architectures and multi-transform architecture which supports both HEVC and AVC is shown in Table 4. The total number of slice registers, and LUTs used for standalone architectures are 7152 and 10,795. The total number of slice registers, and LUTs used

Table 4 Device utilization summary

Model	AVC	HEVC	Multi-Transform
Device used	Artix-7 xc7a200t-ffg1156		
No. of slice registers	3249	3903	4109
No. of slice LUT's	4927	5867	6993
No. of fully used LUT-FF pairs	3186	3873	3968
No. of bonded IOB's	417	417	419
Min time (ns)	4.065	4.309	4.69
Max frequency (MHz)	246.03	232.09	213.82
Dynamic power (watts)	0.249	0.295	0.312

for multi-transform architecture are 4109 and 6993 which results in reduction of area by using multi-transform architecture instead of using standalone architectures. The designed model can be operated at a maximum frequency of 246.03 MHz for AVC and 232.09 MHz for HEVC and 213.82 MHz for multi-transform which is slightly less due to additional control overhead.

The total power consumed by the standalone architectures and multi-transform architecture is represented in Table 4. It may be observed that there is a minimal increase in the dynamic power in multi-transform architecture as compared to standalone structures. However, the total dynamic power of standalone structures will add up to 0.544 W which is quite large compared to 0.312 W as consumed by multi-transform architecture. The comparison of resource utilization of the proposed method with the earlier work is tabulated in Table 5 where it is observed that total number of LUTs and registers used is reduced by 67% and 43%, respectively, as compared with the methodology proposed in [6]. The work in [16] uses less LUTs and registers but uses DSP48Es as it is multiplier-based approach and supports only transforms of HEVC.

The comparison of performance of proposed architecture with related works is shown in Table 6. The maximum frequency of operation of the proposed architecture is higher and offers an increased throughput by at least $4\times$ as compared to the designs in [7, 8, 14]. However, the approach in [8] and [14] does not support HEVC standards. Although the architecture presented in [6] has high frequency and throughput, the proposed architecture is using substantially less area which makes it advantageous.

Table 5 Comparison of resource utilization

Design	Dias et al. [6]	Awab et al. [16]	Proposed
Device used	Virtex-7	Kintex Ultra	Artix-7
No. of LUT's	21,568	5049	6993
No. of registers	7309	2509	4109
Max frequency (MHz)	279.4	86.19	213.8
DSP48Es	–	368	–
Supported formats	AVC, HEVC	HEVC	AVC, HEVC

Table 6 Comparison of performance with other architectures

Design	Technology	Maximum frequency (MHz)	Latency [ns]	Throughput [GSamp/s]	Supported standards	
					AVC	HEVC
[6]	Virtex-7	279.4	57.2	2.2	Y	Y
[8]	180 nm	146	438.4	0.2	Y	–
[7]	180 nm	211.4	56.8	0.2	Y	Y
[14]	130 nm	100	1000	0.4	Y	–
Proposed	Artix-7	213.82	120	1.6	Y	Y

4 Conclusion

An architecture for supporting integer DCTs of multiple transforms, namely AVC and HEVC, is proposed in this work. Initially, a four-point 1D DCT architecture having three stages of computation is designed for both AVC and HEVC standards where multipliers are realized using shift-add technique. Further, by combining four-point DCTs, eight-point DCT is obtained. The technique is extended to 2D DCT as well. Proposed architecture is described using Verilog HDL and simulated using Xilinx ISE tools. Results obtained are validated by comparing it against the output obtained from MATLAB. The architecture is synthesized and analyzed for area, power, and timing by implementing on FPGA Virtex-6 device. Compared to standalone architectures, proposed multi-transform architecture results in 41% saving in adders and 16% saving in shift operations. On implementation, proposed architecture uses 67% less LUTs and 43% less registers as compared to an approach in literature. Proposed architecture has an increased throughput (8 outputs/clock cycle after the initial latency), is scalable to higher orders and higher dimensions, highly parallel and pipelined which resulted in better performance.

References

1. P.K. Meher, S.K. Lam, T. Srikanthan, D.H. Kim, S.Y. Park, Area-time efficient two-dimensional reconfigurable integer DCT architecture for HEVC. *Electronics* **10**(5), 603 (2021)
2. P.K. Meher, S.Y. Park, B.K. Mohanty, K.S. Lim, C. Yeo, Efficient integer DCT architectures for HEVC. *IEEE Trans. Circuits Syst. Video Technol.* **24**(1), 168–178 (2013)
3. V. Dhandapani, S. Ramachandran, Area and power efficient DCT architecture for image compression. *EURASIP J. Adv. Signal Process.* **2014**(1), 180 (2014)
4. K.K. Senthil kumar, K. Kunaraj, R. Seshasayanan, Implementation of computation-reduced DCT using a novel method. *EURASIP J. Image Video Process.* **2015**(1), 1–18 (2015)
5. T. Dias, N. Roma, L. Sousa, Unified transform architecture for AVC, AVS, VC-1 and HEVC high-performance codecs. *EURASIP J. Adv. Signal Process.* **2014**(1), 108 (2014)
6. T. Dias, N. Roma, L. Sousa, High performance multi-standard architecture for DCT computation in H. 264/AVC high profile and HEVC codecs, in *2013 Conference on Design and Architectures for Signal and Image Processing*, pp. 14–21 (IEEE, 2013)
7. M. Martuza, K. Wahid, A cost effective implementation of 8×8 transform of HEVC from H. 264/AVC, in *2012 25th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pp. 1–4 (IEEE, 2012)
8. K. Wahid, M. Martuza, M. Das, C. McCrosky, Resource shared architecture of multiple transforms for multiple video codecs, in *2011 24th Canadian Conference on Electrical and Computer Engineering (CCECE)*, pp. 000947–000950 (IEEE, 2011)
9. N. Mohankumar, M.N. Devi, D.B. Nath, A. Scaria, VLSI architecture for compressed domain video watermarking, in *International Conference on Digital Image Processing and Information Technology*, pp. 405–416. (Springer, Berlin, Heidelberg, 2011)
10. R.K. Megalingam, V. Vineeth Sarma, B. Venkat Krishnan, M. Mithun, R. Srikumar, Novel low power, high speed hardware implementation of 1D DCT/IDCT using Xilinx FPGA, in *2009 International Conference on Computer Technology and Development*, vol. 1, pp. 530–534. (IEEE, 2009)

11. B. Shyam, V. Vignesh, Image quality compression based on non-zeroing bit truncation using discrete cosine transform, in *2020 4th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech)*, pp. 1–5 (IEEE, 2020)
12. V. Chandran, I. Mamatha, S. Tripathi, NEDA based hybrid architecture for DCT—HWT, in *2016 International Conference on VLSI Systems, Architectures, Technology and Applications (VLSI-SATA)*, pp. 1–6 (IEEE, 2016)
13. M. Nair, I. Mamatha, S. Tripathi, Distributed arithmetic based hybrid architecture for multiple transforms, in *Advances in Signal Processing and Communication*. (Springer, Singapore, 2019), pp. 221–232
14. I. Mamatha, J. Nikhita Raj, S. Tripathi, T.S.B. Sudarshan, Systolic architecture implementation of 1D DFT and 1D DCT, in *2015 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, pp. 1–5 (IEEE, 2015)
15. C.W. Chang, H.F. Hsu, C.P. Fan, Unified forward and inverse integer transforms design with fast algorithm based hardware sharing architecture, in *2020 2nd International Conference on Computer Communication and the Internet (ICCCI)*, pp. 126–135 (IEEE, 2020)
16. A.H. Awab, A.A.H. Ab Rahman, M.S. Rusli, U.U. Sheikh, I. Kamisian, G.K. Meng, HEVC 2D-DCT architectures comparison for FPGA and ASIC implementations. *Telkomnika* **17**(5), 2457–2464 (2019)

Performance Analysis of SDN-Inspired Swarm Intelligence-Based Routing Optimization Algorithm in Vehicular Network



K. Abinaya, P. Praveen Kumar, T. Ananth kumar, R. Rajmohan, and M. Pavithra

Abstract Vehicular network focuses on the development of ad hoc networks on telecommunication vehicles which use vehicle security, like routing algorithms which communicate with one another in the decentralized network controlled by software-defined networking (SDN) provides different transport infrastructure. It can be used in vehicular communications, in two different ways are as follows: transmission between both the internal vehicle and station. That involves the connection stabilization interface is being used for network data transmission; however, it is not effective in handling the complex functionalities, as well as the source routing would still be fairly constant and hence de-automate using constructive hop-by-hop communication, as well as the selecting of routing is not essentially done, and as such the proposed concept of bio-inspiring is sometimes delayed in transmission. This technique is applied in the network context of vehicles to reduce delays and increase the number of packets per frame. This was built in to process the service request from the worldwide client, who uses the proposed swarm intelligence routing algorithm to identify a corresponding ideal portal and base station. The main objective of the ground-to-based gateway connection is to help establish and encourage primary source communication efficiency. It is employed to route efficiently through the networking nodes SD is applied to networking in order to keep the vehicular communication protocols steady. It has the benefits of monitoring traffic, forecasting vehicle demand, directing drivers, and preventing accidents.

Keywords Swarm intelligence · SDN · VANET · Routing · Bio-inspired algorithm

1 Introduction

Ad hoc networks on wireless communication vehicles are used for vehicle protection. It involves bio-inspired network algorithms that interact with one another in the decentralized network optimized by different infrastructure and services supported

K. Abinaya (✉) · P. Praveen Kumar · T. Ananth kumar · R. Rajmohan · M. Pavithra
Department of Computer Science and Engineering, IFET College of Engineering, Villupuram,
Tamilnadu, India

by SDN-software-defined networking [1]. Also, for reasons of coordination, the base station/controller can be used to provide and delete services to vehicles. Also, it includes the transmission gateway that serves as the medium in order to secure the interactions between the two final services in it. The idea underlying software-defined networks (SDN) is to split vertical integration by distinguishing the physical network logic in its physical infrastructure (routers and switches) by supporting (logical) core network concentration of power and by adding the ability to set up the network. It also has the capacity to drive on several network platforms when moving nodes from the source to the source. Ad hoc vehicle networks are composed of a mobile infrastructure where each one of these are used and viewed as a cellular network layout node. This contact is set for approximately 100–300 m that also relies on the reach of the range of the vehicular network. In MANET, its users transmit without even a network connection, but all of them will be fitted with Internet connectivity in comparison to vehicular networks, which all their nodes provide independence and are vehicle oriented [2]. Vehicular network is a variant of the mobile ad hoc network (MANET).

Communication, they determine that vehicular network often has optimization techniques than MANET when connected or disconnected from the network. The vehicular networks applications are developed for diverse reasons by much of the following representatives [3]: (1) Safety-oriented applications: Each program helps to improve road and user safety. This software plays an important and critical role in preserving the lives of people involved in injuries, most of which are the product of driving crashes or lack of attention. The node functions as routers in the ad hoc network to transmit and receive information. Reliability, versatility, and agility are a benefit of the scheme. In order to maximize efficiency, the ad hoc network is able to analyze the radio channel environment. The following are the key contributions.

2 Related Works

In VANET based SDN routing [4], using route optimization and source initialization, it is possible to achieve enhanced throughput between different nodes. This also contains exponential message allocation methodologies to solve the issue of routing in much less period. In the long term, the emphasis is on increasing the performance of latency issues. Moreover, in geographical aware, routing protocols use the technologies SDN [5] using SDN innovations, procedure for regional routing, procedure for improved features, framework for enhanced centralization. In the terms of low delay, communication channels and storage gateways, it operates throughout the SUMO, MINNET-WIFI. Research on connection networks and diverse functionalities will help improving. Cognitive routing protocol [6] involving techniques such as SDN, cognitive routing algorithm, frequency sensing, automated system of pack includes used for precise network security routing. It also has the benefits of sustainable routing, added information intelligence, keeps user behavior safe, as well as the drawback is that primary signal activities are reduced by the reduced optimal

likelihood. Vehicular network optimization involving ant colony techniques [7] uses sophisticated GSR technology to efficiently trace geolocation data. It also has the benefit of obtaining optimal connectivity, lowest possible weight percent path, and the suggested idea is the ant colony routing protocol, routing depending on the position. Through routing, it has been used from the source sequence to the later part of the desired location [8]. It has disadvantages in operational costs clustering, frequency overheads. The main goal of WSN is to make it possible to locate various entities and environmental parameters continuously so that you can monitor them for long term. Since energy is a constraint in the task development process, recent innovations have produced efficient protocol design principles for WSNs [9]. Data movement incurs considerable energy. Over the last few years, there have been numerous innovative heuristic clustering protocols proposed for use in this case. Resource management task in WSN configuration is the means the method for restraining power consumption while holding network behavior constant. A WSN will be judged a success if it fulfills its primary goal, but not secondary targets (such as staying alive after battery power is exhausted) [10]. When one of the sensor nodes is determined to be dead, it is removed from the data flow. There is a novel approach discussed in this paper that utilizes the sensor nodes to prevent duplication of information and improve the accuracy and integrity of data transmission from one node to another. However, one disadvantage of WSN is that it prevents the use of applications such as short battery life, energy consumption, battery life, and area deployed. Reducing the packet delay was suggested as a new way to increase the network lifetime and transmit speed of WSN in this paper [11]. The multi-hop data transfer mechanism is part of the WSN architecture. When using large numbers of hops, inaccurate location values can be calculated. We present a recently discovered social networking concept; we could be called small world characteristics and discuss a novel methodology for locating nodes on a small world network via social media (SW-WSN). One important fact about data compression is that data gathered by nodes must be highly correlated in order to be effective. While this is often true in practical applications, other scenarios, it is different, for example, in water data, the Internet of Things (IoT) is already a mature, and the new long-range single-hop (L-R-SH) network (LRE) is growing rapidly. In comparison, in a multiple-copy-based routing strategy, the control system temporarily includes and advises the clusters to choose the very next finest carriers. In the network, lots of copies of the mobile node are transmitted, and the control system constantly encourages to help determine the copies' propagation to the desired location. However, the handling of dynamic source routing across the specification medium is disadvantageous.

3 Swarm Intelligence in VANET Routing

The bio-inspired swarm intelligence technique can be embedded in identifying the successful route discovery over through the nearby vehicles. In addition to delivering the path to the destination end, the bio-inspired routing discovers an efficient way of

connecting with both the ground station. The service providers are available as part of request/response techniques.

It has a commonly agreed base station (BS), which reacts to the customer's query, which is only accomplished when the automobile is under the transmission range, and when not protected, then the transmission permit to the nearby vehicles personifies its route from the source point to both target ends. The base station is used to control the network working. The basic station that also includes the controller is used for the means of delivering services to the vehicle node and is used for the control of things and for the protection against traffic to protect the propagation over the medium. The networking addresses are being used to analyze the influence of physical fitness.

The value of fitness (VF) is calculated by the following formula:

$$VF = UL_E(N_C + I_C + H_C) \quad (1)$$

where UL_E —upper limit energy, N_C —network operation costs, I_C —node interaction cost, H_C —hardware cost.

3.1 Architecture of Path Routing Source

The common communication specification using SDN and vehicular network is shown in Fig. 1. The three distinct planes have various stacks, which include data protection, monitoring, and data planes. The access point is accountable for the aerial maintenance and coordination of services. The performance measures include the ratio of received packets, end-to-end latency, and failure rate. In order to define the routing over all the social networking mostly on vehicle sites, it is a very effective system. Vehicular ad hoc networks can be used for numerous application domains pertaining to security, such as preventing injuries, road traffic, and emergency warning. Every one of these applications requires vehicles during regular interval to telecast their position, speed, path, and individuality. The private information of the destination of the swarm intelligence-based vehicle routing must be preserved. The swarm intelligence-based vehicle routing algorithm is explained below:

Algorithm 1: Swarm intelligence-based vehicle routing algorithm

Input: ID_{sv}, ID_{dv} Output: output path for transmission

Step 1: Discovery of the procedural path (Source ID, Destination ID)

Step 2: Check if Destination ID exists

Step 3: Then route < - bidirectional (ID_{sv}, ID_{dv}).

Step 4: else

Step 5: Source ID, destination ID identified.

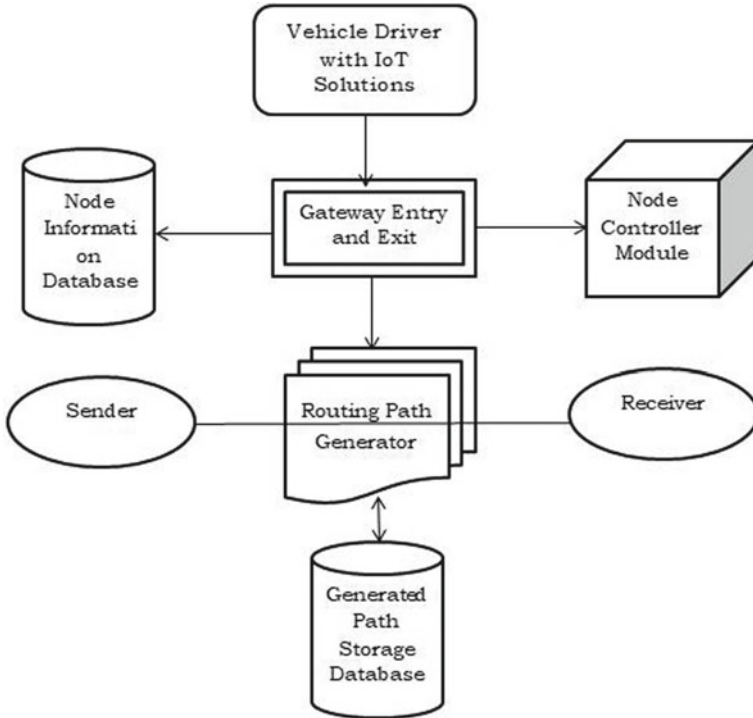


Fig. 1 Proposed routing network model

Step 6: Develop the path networking model.

Step 7: Transmit applications.

Step 8: Activate (Time signal).

Step 9: Calculate minimum time period neighbor for destination ID.

Step 10: Construct the delay for processing.

Step 11: else

Step 12: Restart the discovery path

Step 13: Termination of the procedure

Notations: DSV = source vehicle ID, DDV = Destination ID

DNV = Neighbors of the existing vehicle

RT = Tree-based route model

TR = Transmitter vehicle time span

RE = Receiver vehicle time span

3.2 Proposed Routing Methodology

Although the source routing is fairly stagnant and as such cannot enhance the use of dynamic hop-by-hop communication and the availability of paths is not effectively done, sometimes there is also even a delay in processing. This same objective of the problem definition is to enable accurate detection over the vehicular nodes. Vehicular network is used throughout the networking medium besides interaction applications, and SDN is being used to customize the network architecture from over medium. The principle of bio-inspired method is being used for optimal network; it utilizes the notion of animal behavior to find the best route between both the sender and recipient; it also facilitates efficient path prediction and consciousness of the location, etc. This same interface of the route optimization processing element in vehicular network is regarded as having the able to manage the functionality within it. Also, it increases this same response time of processing by discovering the route with many other modules from both ends. The various process modules include:

1. **Communication Module:** This encrypts the message on the physical medium of the channel, trying to represent a connection to some other device. Unless the sources are implementing a finest message service, as eventually as the text has also been encrypted mostly on medium, this same confirmation primitive is generated. It has been used from either the source or destination of the pattern for the purpose of communication.
2. **Network Topology Establishing:** Network architecture is a systemic structure of vehicular nodes that makes reference to a network's physical or virtual layout. The reason multiple devices are positioned and interdependent with one another is defined. Alternatively, it can define how well the information between such nodes is transmitted.
3. **Generate Path:** The above device is being used to create the route from the source to the end of the departure point using the shortest route routing computational segments methodologies by which it can identify and communicate with the closer node from the sources by providing the knowledge between them and making it appropriate for itself to commute from the sender to the receiver medium at low cost.

Algorithm-2: Bio-inspired routing path generation

Input: Application for service

Output: Access Point/Efficient Path

Vehicle Data: Vehicle ID (VID), Vehicle IP Address (IPA), (E) Nth Energy

Neighbor vehicle, (IOC) Nth neighbor vehicle IO cost, (CC) Communication Cost, (IOC)

Cost of the nth neighbor vehicle (PC) of the nth neighbor vehicle processor cost.

Path Discovery for Procedure

{

The scan of the requested vehicle for the base station

If it is detected

Send request to base station Base Station

Else

{

The requester vehicle dynamically scans within its requester vehicle for the neighbor vehicle and the area of coverage. The vehicle requester sends each neighboring vehicle a beacon message and collects their vehicle information.}}

Algorithm 3: Requestor for vehicle information

Do assuming 'g' number of neighboring vehicles for $n = 1 \cdot g$.

Acquire vehicle information from neighboring countries: VID, IPA, ME, IOC, CC, PC.

The calculation of the fitness value is done by:

$$E/(IOC + CC + PC) = \text{Optimal vehicle (OV)}$$

Determine that OV.

Develop the path networking model

Send your application to OV.

Discovery of restart path (Requester vehicle = OV)

End procedure

4 Experimental Setup

Implementation of the initial setup medium for nodes: The graph represents the construction process where it takes the input data of time (m/s) by operating it the configuration set-up, at first, and the duration was null; the devices are in fixed mode and then when duration improves as input it representations transition of nodes from one to another through deformation from source to destination establishing by of grouping of transfer information between the vehicle. That has inputs in system time (m/s) and provides output in node transfer medium form among other nodes in the network configuration.

In the implementation, initially, the time was zero; the nodes are in static mode, and when time increases as input, it forms transformation of nodes from one to other by transverse from source to the destination ends by form of clustering to transfer information among the vehicular nodes managed by base station (Bs) through gateway sources. It has input in the form of time in (m/s) and produces the output in the form of node transfer mediums among other nodes in the configuration of the networks.

The nodes are all from different media to configure their qualities between one another each node through their pattern network coverage is regulated by the ground service provider, respectively.

5 Result Analysis

The implementation process is a lightweight application that utilizes the Omnett++ simulator tool as well as utilizes the C++ and TCL languages for both the reasons of network security configuration (Figs. 2 and 3).

The extremely effective processor with such a lowest possible speed of 2.5 ghz was permitted and was adaptable to operate in any environment. The *x*-graph formalized defines the characterization of 2D points in a graphics of components density and RGS. The moment modules are used in the procedure of Adap_RT, packet delivery ratio, and PCR. This diverges from one another under different restrictions. The comparative analysis of time implications in vehicle range on 3G-CMGM, OHBR, CMGM, and SIA is shown in Fig. 4, wherein the proposed SIA algorithm does have

Fig. 2 Adap_RT in X-graph

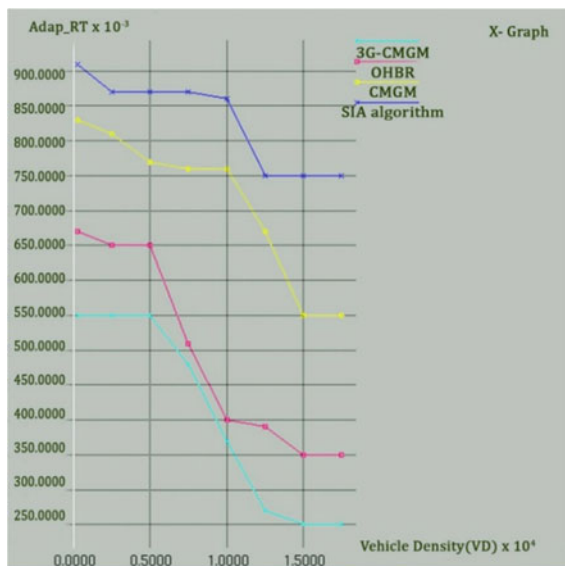


Fig. 3 ROR time processing in X-graph

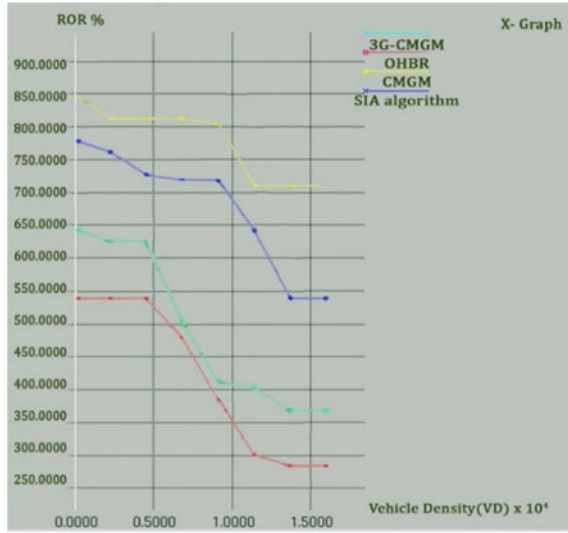
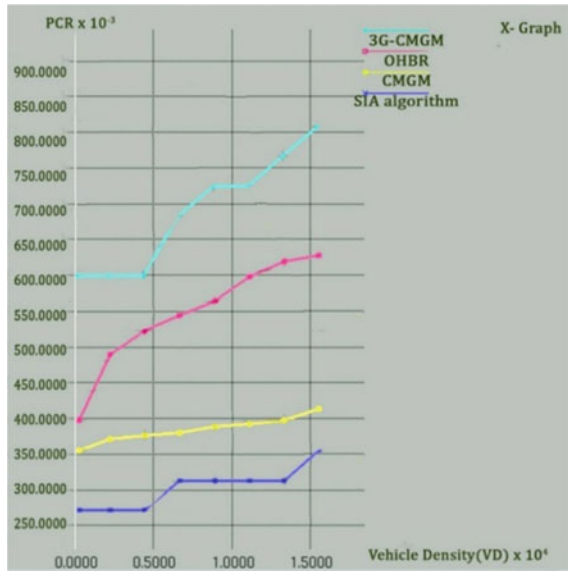


Fig. 4 PCR time based on X-graph



best time complexity.

The contrast of interval implications in vehicle range on 3G-CMGM, OHBR, CMGM, and SIA is shown in figure, wherein the proposed SIA algorithm does have best time-efficient ROR for processing. The comparative analysis of time implications in RGS on 3G-CMGM, OHBR, CMGM, and SIA is described in Fig. 4, wherein the proposed SIA algorithm does have best PCR processing time-efficiency.

Table 1 Comparative analysis of protocols

Metrics	3-CMGM	OBHR	CMGM	SIA
Adap_RT	0.24	0.08	0.15	0.38
Packet concurrency ratio	370	180	250	450
Rate of return	82	75	93	85

The table represents the minimum cost- and time-efficiency of the X-graph, whereas the proposed SIA algorithm functions efficiently with much less complexity in all of the vehicle range restrictions relative to several other time parameters.

Table 1 illustrates that the least time and cost-effective in the service of packets between one and other resources in period, wherein the X-graph predicated on Adap_RT is much more expensive equated to many metrics on the message ratio, and the proposed SIA achieves objectives with less complexity in different forms of constraints, by comparing two tables based on data collected.

6 Conclusion

The proposed methodology in this paper utilizes the swarm intelligence procedure, and the project inference is being used to enhance effective routing over the network security nodes, promoting efficient transmission of messages in much less time and density-related restrictions, including the idea of SDN and vehicle network utilizing simulated network security modules. This structure of modularity aims to detect damage and prevents the malicious impact from expanding across the whole system. The performance of the proposed framework is analyzed using Adap_RT, packet delivery ratio, and PCR metrics. The results suggest that bio-inspired techniques can handle unacceptable networking issues better without any of the human operator's intervention. The future work suggests the use of variants algorithm in swarm intelligence to improvise the results of network issues.

References

1. F.Y. Okay, S. Ozdemir, Routing in fog-enabled IoT platforms: a survey and an SDN-based solution. *IEEE Internet Things J.* **5**(6), 4871–4889 (2018)
2. J. Kaur, G. Singh, Review study on MANET routing protocols: challenges and applications. *Int. J. Adv. Res. Comput. Sci.* **8**(4) (2017)
3. N. Bhalaji, Performance evaluation of flying wireless network with Vanet routing protocol. *J. ISMAC* **1**(01), 56–71 (2019)
4. O. Alzamazami, I. Mahgoub, Link utility aware geographic routing for urban VANETs using two-hop neighbor information. *Ad Hoc Netw.* **106**, 102213 (2020)
5. M.T. Abbas, A. Muhammad, W.C. Song, SD-IoV: SDN enabled routing for internet of vehicles in road-aware approach. *J. Ambient. Intell. Humaniz. Comput.* **11**(3), 1265–1280 (2020)

6. A.R. Sangi, M.S. Alkathairi, S. Anamalamudi, J. Liu, Cognitive AODV routing protocol with novel channel-route failure detection. *Multimedia Tools Appl.* **79**(13), 8951–8968 (2020)
7. Z. Khan, S. Fang, A. Koubaa, P. Fan, F. Abbas, H. Farman, Street-centric routing scheme using ant colony optimization-based clustering for bus-based vehicular ad-hoc network. *Comput. Electr. Eng.* **86**, 106736 (2020)
8. M. Adimoolam, A. John, N.M. Balamurugan, T. Ananth Kumar, Green ICT communication, networking and data processing, in *Green Computing in Smart Cities: Simulation and Techniques. Green Energy and Technology* ed by B. Balusamy, N. Chilamkurti, S. Kadry (Springer, Cham, 2021). https://doi.org/10.1007/978-3-030-48141-4_6
9. E.N. Al-Khanak, S.P. Lee, S. Ur Rehman Khan, A. Verbraeck, H. van Lint, A heuristics-based cost model for scientific workflow scheduling in cloud. *Comput. Mater. Continua* **67**(3), pp. 3265–3282 (2021)
10. M.J. Awan, M.S.M. Rahim, H. Nobanee, O.I. Khalaf, U. Ishfaq, A big data approach to black Friday sales. *Intell. Autom. Soft Comput.* **27**(3), pp 785–797 (2021)
11. M. Krichen, S. Mechti, R. Alroobaea, E. Said, P. Singh et al., A formal testing model for operating room control system using internet of things. *Comput. Mater. Continua* **66**(3), 2997–3011 (2021)
12. Z.H. Ali, M.M. Badawy, H.A. Ali, A novel geographically distributed architecture based on fog technology for improving Vehicular Ad hoc Network (VANET) performance. *Peer-to-Peer Netw. Appl.* **13**(5), 1539–1566 (2020)
13. K.P. Sampooram, S. Saranya, S. Vigneshwaran, P. Sofiarani, S. Sarmitha, N. Sarumathi, A comparative study on reactive routing protocols in VANET, in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 726–731 (IEEE, 2020)
14. S. Hasan, A.H. Al-Bayatti, S. Khan, Critically analysing routing protocols in a vehicular ad-hoc network, in *Proceedings of the 2020 4th International Symposium on Computer Science and Intelligent Control*, pp. 1–7 (2020)
15. A. Gopalakrishnan, P.M. Bala, T.A. Kumar, An advanced bio-inspired shortest path routing algorithm for SDN controller over VANET, in *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*, pp. 1–5 (IEEE, 2020)

Blending of Window Functions in Sonar Array Beamforming



S. Vijayan Pillai, T. Santhanakrishnan , and R. Rajesh

Abstract Extensive research is being carried out in beamforming techniques to suppress the sidelobe levels in the radiation pattern of sonar arrays. Several optimization techniques have been established over the years and practiced in a variety of sonar systems. Most of such techniques use separate window functions with optimized coefficients for beamforming. This paper investigates the blending of two such window functions for forming an actual beam through an innovative product beamforming method. Four popular window functions are considered for the design, and their effective utilization is assessed for sonar applications. The designed product beamformers are evaluated on a 24 elements sonar array through simulation, and their tolerances to array deformations like errors in hydrophone positions are evaluated. The obtained results suggest that the beamformer constructed using the blending of rectangular and null steered Dolph-Chebyshev window functions performs better and is stable against array deformity than traditional beamformers.

Keywords Beamforming · Sonar array processing · Window functions · Blending of window functions · Target detection · Localization

1 Introduction

The primary setback in antenna arrays is high sidelobe levels (SLL) in their radiation pattern. Low sidelobe powers are essential to curtailing the problem of false target identification in sonar or radar systems. The low SLLs can be obtained by varying the number of array elements, the inter-element spacing and coefficients of window functions [1, 2]. Beamforming helps in achieving high gain in the desired directions by reducing the powers of sidelobes. Several applications exist for beamforming, especially in radars, sonars, mobile, and satellite communications [3, 4]. The unsolicited clusters like echoes from the ocean bottom, acoustic channel reverberations, ambient ocean noises, and other interferences emanated from different sources, etc., limit the

S. Vijayan Pillai · T. Santhanakrishnan (✉) · R. Rajesh
Naval Physical and Oceanographic Laboratory, Kochi, Kerala 682021, India
e-mail: tsanthan@npol.drdo.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022
P. Verma et al. (eds.), *Advances in Data Computing, Communication and Security*,
Lecture Notes on Data Engineering and Communications Technologies 106,
https://doi.org/10.1007/978-981-16-8403-6_48

521

detection and tracking of targets by sonar systems [5–7]. Underwater surveillance systems like sonar arrays use panoramic display systems for a 360° inspection across the ocean [8, 9]. Suppressing ambient acoustic noise and other interferences from other directions, while enhancing the acoustic signal received from underwater platforms is essential to detect targets and their direction of arrival (DoA). Beamformers have been employed for this purpose [10–12].

Sonar arrays formed by several individual hydrophones in linear, circular, cylindrical, conical, spherical, etc., configurations are used for forming preformed or steered beams in all directions. Such preformed beams are expected to be identical and unperturbed. Non-identical beams result in errors in bearing, resolution, and target discrimination. Beamwidth, SLL, beam sensitivity, and tolerance to the array integrity are the deciding factors of beam identity. Thus, narrow beamwidth, low SLLs, high sensitivity, and high tolerance to the array integrity are the most desirable factors for a maximum probability of detection with minimum probability of false alarm [13, 14]. The beamwidth, SLL, beam sensitivity can be controlled by window functions by providing optimized individual weighting called windowing.

Several window optimization algorithms exist to address the above conflicting factors but are associated with computational complexities [15–19]. These window optimizations are mainly suitable for uniform linear arrays (ULA) with omnidirectional hydrophones. They are often sensitive to sensor mismatches viz., variation in receiving sensitivity of individual hydrophones, errors in hydrophone location, and variation in the acoustic center. All window functions have shortcomings in one or more properties as every window function has its own characteristics. Thus, a window function good in one property may lag in other properties. These shortcomings can be reduced greatly by suitably blending different window functions [20]. However, the research in this area is scarce, and many advances are expected.

In this paper, a comparative study of different window functions and their blending is carried out to choose the best among them. The investigation is limited to a linear array and can be extended to any array. The proposed beamformers are tested through simulation on a 24-element ULA. The tolerance of the beamformers to array deformations like errors in hydrophone positions is investigated. The results obtained with the present method are compared with that of the traditional techniques to highlight the improvements and the robustness of the present method.

2 Geometry and Methodology for Simulation

Contemporary sonars use multiple hydrophones in an array. More hydrophones produce more signal information and can help suppress noise and interference. The beamformer enhances the amplitude of the signal by operating on the data obtained from an array of sensors [21]. Sonars use beamforming to estimate the DOA of the target by digital processing of the signals obtained from the array. The main function of a beamformer is to exploit the signals arriving from the different hydrophones at different times due to the signal angle. An example of a ULA with

24 hydrophones showing a source and its wavefronts is shown in Figure 1, and a broadside beamforming scheme is shown in Figure 2.

Assume a linear array of hydrophones with an inter-element spacing of $\lambda/2$ as shown in Figure 1. Here, λ is the wavelength concerning the maximum acoustic frequency of the considered acoustic band in water that is emanated from a far-field acoustic source as shown in Figure 1. Supposing the acoustic source is located at the broadside of the sonar array and its wave enters perpendicular to the plane of the array, then the outputs of each hydrophone are in phase and will add up coherently.

Fig. 1 Schematic of an ideal linear array of 24 hydrophones showing source signal and wavefronts

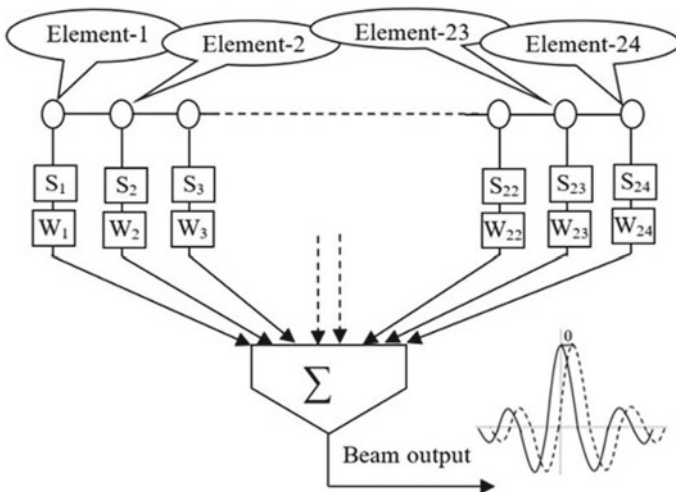
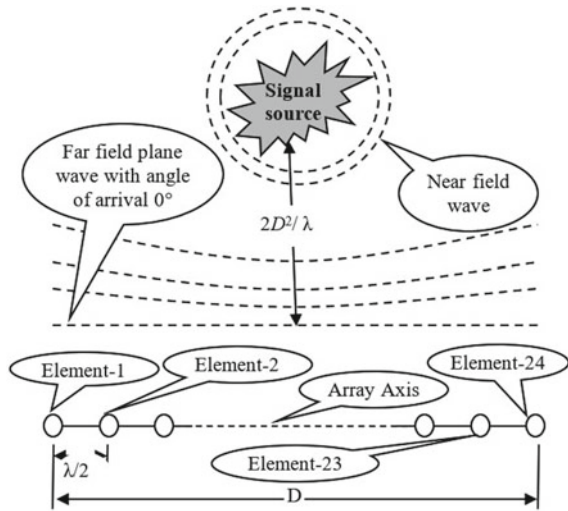


Fig. 2 Schematic of broadside beamforming with a uniform linear array of 24 hydrophones

Therefore, a beam pattern will be formed with centring at 0° , as shown by the solid line in the bottom curve of Figure 2. Likewise, if the acoustic source moves around the sonar array or if the sonar array is rotated around the acoustic source, then the hydrophones receive signals with different time delays. In such conditions, the outputs of each hydrophone no longer add coherently, and hence, the peak shifts from the center. Suitable mathematical operations or beamforming techniques need to be applied to ensure a better beam pattern with higher gain from all the hydrophone elements to detect and track the targets. The desired beam pattern is synthesized by employing suitable window functions with suitable window coefficients and by adjusting the spacing between hydrophones [22–25].

2.1 Window Functions and Beam Patterns

Window functions are operated in spectral analysis for spectral modification, design of filters, radar, and sonar. The way of windowing functions is the most widespread method for inhibition of spectral leak. Window functions are also called tapering functions. Several window functions have been developed for truncating and shaping signal segments. The process of extracting a portion of the signal and finding the spectrum of that portion is called windowing. No window function is the best in all aspects. Thus, it should be selected as per the design requirements. Four special window functions namely rectangular, Dolph-Chebyshev, triangular, and Tukey are examined in this article for sidelobe suppression and for blending through product beamforming. The mathematical equations and other common details are not given as they are available in many of the textbooks and literature [26–28].

Beam patterns obtained using these four window functions for an array with 24 hydrophones are shown in Figure 3. All the hydrophones are given the equivalent weights of the respective windows. The beam pattern with the rectangular window exhibits the first SLL at -13.25 dB and the second SLL at -17.73 dB. The width of the main lobe is $\sim 4.2^\circ$. The first nulls are seen at $\sim \pm 5^\circ$, second nulls at $\sim \pm 10^\circ$, and the third nulls at $\sim \pm 15^\circ$. Since all the hydrophones with equal weights have contributed to beamforming in the rectangular window, the array sensitivity is found to be maximum with a high gain in array SNR of 13.8 dB. As seen from Figure 3, the higher SLL is the most undesirable feature of this window function.

The SLLs of the other three window functions are observed to be lower than that of the rectangular case but are suffered by higher main lobe width; thereby, resolving the targets in the complex ocean is limited. The Dolph-Chebyshev window function exhibits very low but equal SLLs approximately at -55.91 dB. The 3 dB beamwidth is increased to 6.4 dB with a factor of 1.7 compared to that of the rectangular case. A reduction in array gain by a factor of 0.45 is observed, which is undesirable. No single-window function produces a beam pattern with a narrow main lobe and low SLLs. Thus, the blending of window functions is motivated to get a reasonable narrow main lobe with the lowest SLLs through the concept of product beamforming.

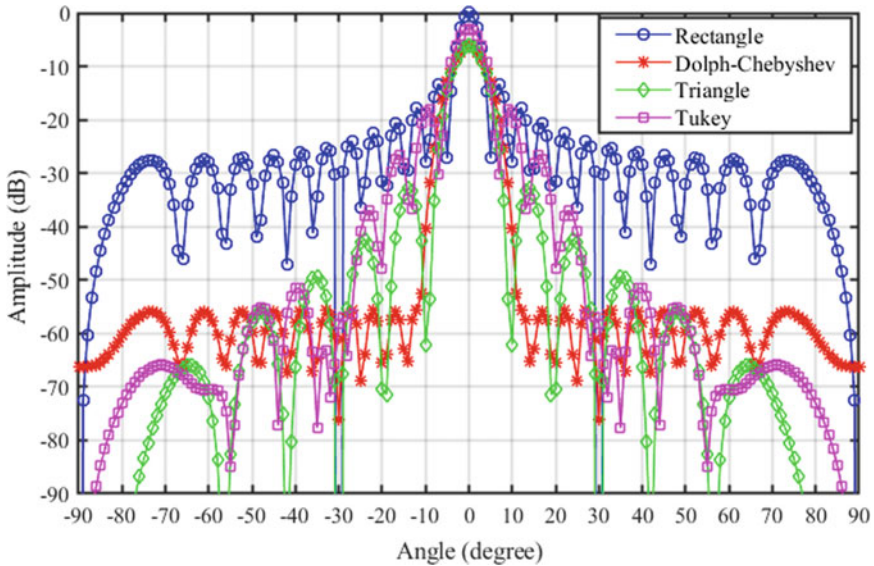


Fig. 3 The beam pattern of selected window functions with 24 elements array

3 Product Beamforming Methodology

The product beamforming method is schematically shown in Figure 4. The output of each hydrophone on the array is passed through two chosen window functions. Subsequently, they are added and computed the energy; thereby, two independent beam patterns are formed. Then, a single beam pattern is obtained by taking the product of them. This final beam pattern is used for the detection and tracking of targets.

Figure 5 displays the product beam patterns of the Dolph-Chebyshev, triangular, and Tukey window functions with the rectangular window function.

Figure 5 indicates that applying a product on two beam patterns tries to retain the quality of the best among them. On the other hand, it can be said that the product beam

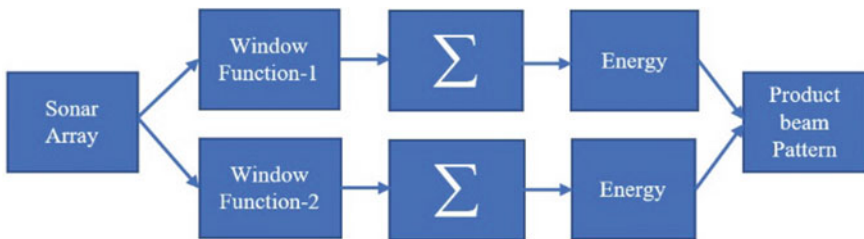


Fig. 4 Schematic of the product beamforming method

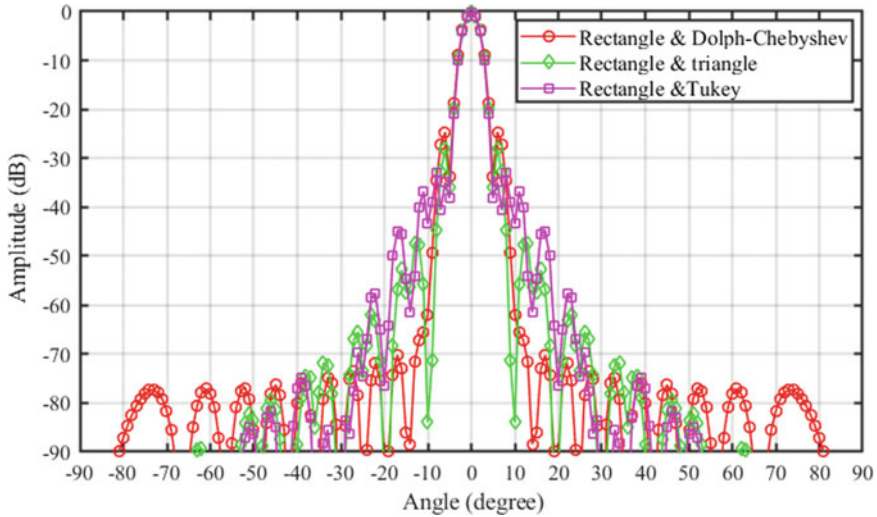


Fig. 5 Product beam patterns of the Dolph-Chebyshev, triangular, and Tukey with rectangle

pattern corresponding to the rectangular window with that of the Dolph-Chebyshev exhibits the first SLL now at -24.68 dB. A reduction of ~ 11.43 dB in sidelobes and a marginal increase in the main lobe width is achieved. Likewise, the product beam patterns of the triangle and Tukey functions report an improvement in SLL reduction of about 14.76 dB and 21.43 dB with a very marginal improvement in the main lobe width. However, it is interesting to note that the blending of rectangular and Dolph-Chebyshev windows reports better performance as the SLLs other than the first sidelobe are drastically reduced. Figure 6 displays the product beam patterns relating to the triangular, and Tukey with the Dolph-Chebyshev window function and the product beam pattern relating to the triangle and Tukey functions.

Figure 6 also indicates that operating a product on two beam patterns tries to maintain the attribute of the best among them. Unlike the results shown in Figure 5, it can be seen from Figure 6 that the product beam patterns of the Dolph-Chebyshev with that of the Tukey and triangle windows exhibit better performance. But, this is still inferior to the product beam pattern relating to the rectangle and Dolph-Chebyshev window functions. The product beam pattern relating to the triangle and Tukey windows does not yield any tangible outcome suitable for sonar surveillances.

The above analysis confirms that the rectangular window produces a very narrow main lobe with higher array sensitivity, whereas the Dolph-Chebyshev window produces very low SLLs. The sonar designer expects these two main parameters from a single-window function. But unfortunately, no single-window function or a simple product function yields these results. Thus, another form of beam pattern blending is motivated to get such a desired result using rectangular and Dolph-Chebyshev windows [20]. The details of getting the same are narrated in the following section.

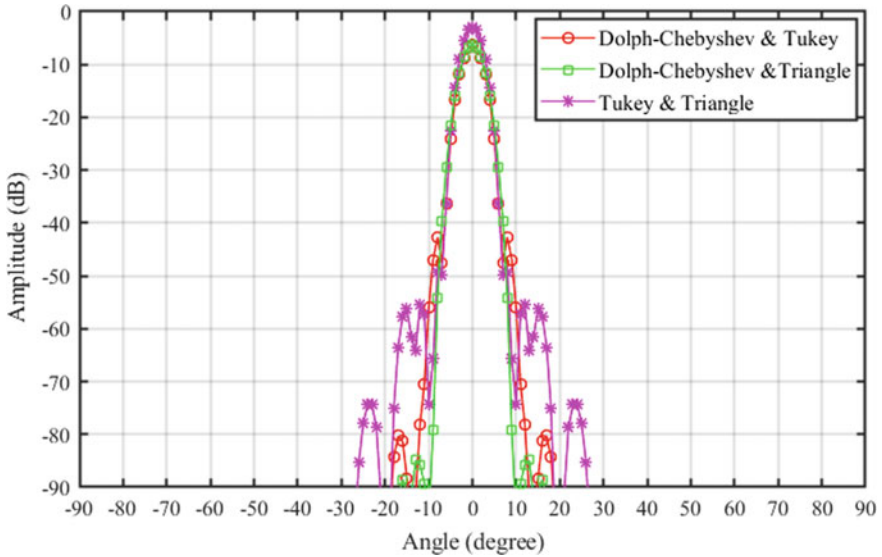


Fig. 6 Product beam patterns of the triangular, and Tukey with the Dolph-Chebyshev and the same relating to the triangle and Tukey windows

3.1 Null Steered Product Beamforming

The null steered product beamforming method is shown in Figure 7. The output of each hydrophone on the array is passed through two chosen windows. Here, the first window has to be rectangular to maintain the supremacy of narrow main lobe width and higher array gain. The second window has to be the Dolph-Chebyshev to maintain the supremacy of the low SLLs. The Dolph-Chebyshev window is; however, null steered such a way that the trough points belonging to the sidelobes of Dolph-Chebyshev beam pattern coincide with the crest points belonging to the sidelobes of the rectangular beam pattern without affecting the main lobe of the rectangular beam pattern. Later, they are added, and energy is computed; thereby, two independent

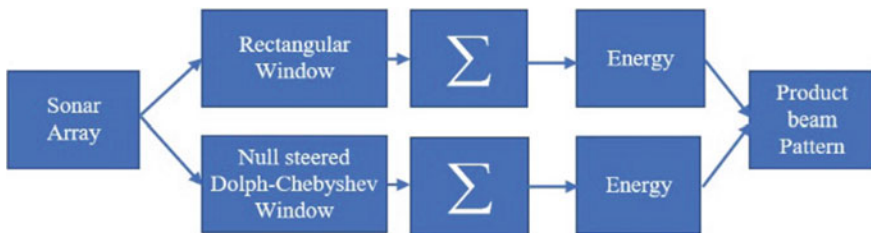


Fig. 7 Schematic of the null steered product beamforming method

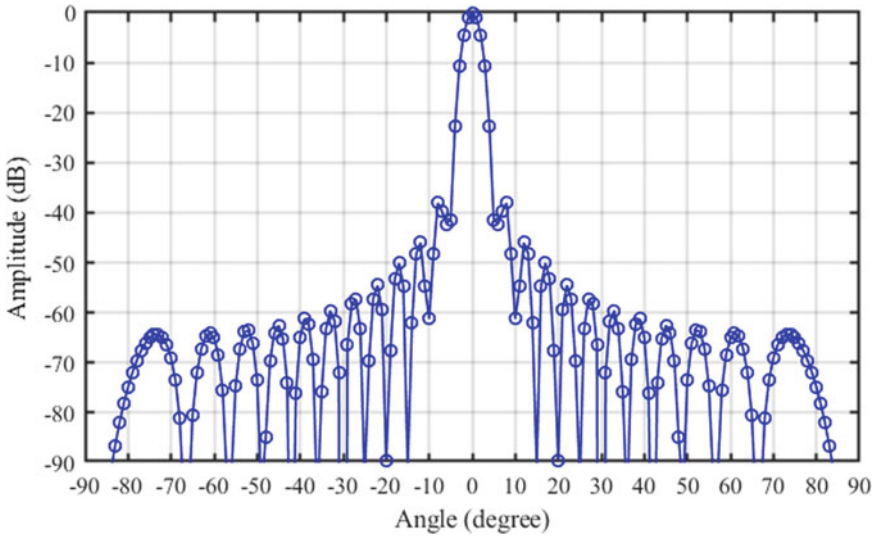


Fig. 8 Product beam pattern resulted from rectangular and null steered Dolph-Chebyshev

beam patterns are formed. Then, a single pattern is obtained by taking a product of them. This final beam pattern is used for the detection and tracking of targets.

Figure 8 shows the resulted beam pattern. The steered nulls of the Dolph-Chebyshev window helped in bringing down the prominent sidelobes of the rectangular window greatly without affecting the width of the main lobe and array sensitivity.

The obtained product beam pattern using the present approach exhibits the first SLL at -38.05 dB instead of the traditional -13.25 dB and the second at -45.98 dB instead of the traditional -17.73 dB, and the trend continues toward the lower side lobes. The width of the main lobe is maintained at $\sim 4.2^\circ$. The first nulls are seen at $\sim \pm 6^\circ$, the second at $\sim \pm 10^\circ$ and the third at $\sim \pm 15^\circ$, and so on. Effectively, the sidelobes are suppressed by more than 25 dB using this novel technique. On the other hand, Figure 8 demonstrates the efficacy of the present windows blending method through the use of product beamforming in suppressing the side lobes in beamforming.

4 Tolerance of the Method to Array Deformation

A perfect beamformer assumes a planar wavefront. Conversely, all the hydrophones in the array are expected to encounter the wavefronts in the same phase for forming a perfect broadside beam. For a given λ and aperture length D , the planar wavefront can be realized once the array is at a distance greater than $2D^2/\lambda$ as shown in Figure 1. A wavefront with wavelength λ traveling through a distance of d will experience a

phase change of $2\pi d/\lambda$. On the other hand, a wave with a wavelength of 0.5 m can experience a phase change of ± 0.0628 rad ($\pm 3.6^\circ$) which corresponds to an error of ± 0.5 mm in its sensor element position. For the considered array, the far-field distance is 288λ . The main factor which affects the beam pattern is the phase error manifesting out of variation in inter-element distance referred here as ‘the position’ and planarity of array aperture. A practical array with possible positional errors is shown in Figure 9. The other array parameter which affects the beam pattern is the variation in sensitivity among hydrophones. An ideal beam is formed in broadside only when the hydrophone element sensitivity is omnidirectional and identical.

To see the effects of aforesaid ambiguities in beamforming, errors in three variables, namely (i) an error in $\lambda/2$ along the x -axis (ii) an error in mounting position along the y -axis and (iii) variations in amplitude due to the hydrophone sensitivity are introduced randomly. The computation is carried out 25 times to verify the repeatability of tolerance. The observed magnitude of perturbation in the case of the simple product using rectangular and Dolph-Chebyshev are shown in Figure 10.

The observed magnitude of perturbation in the case of the product beamformer blended using rectangular and null steered Dolph-Chebyshev is shown in Figure 11. Comparison of Figs. 10 and 11 suggest that the product beamformers produced by the blending of the rectangle with either the traditional or the null steered Dolph-Chebyshev windows are better tolerant to array deformations. However, the product beamformer produced by blending the rectangle with the null steered Dolph-Chebyshev windows is the best. The reasons are obvious from the results of Figure 10. It is observed that the main lobe of the beam pattern is getting disturbed at many instances for the product beamformer without null steering. Likewise, the consecutive sidelobes are raising with more energy than the former in many instances. Therefore, it can be postulated that the product beamformer constructed without null steering in the Dolph-Chebyshev window is less tolerant than that of the null steered Dolph-Chebyshev window. This observation is obvious from Figure 11, wherein the main lobe of the beam pattern is not disturbed due to the array deformation.

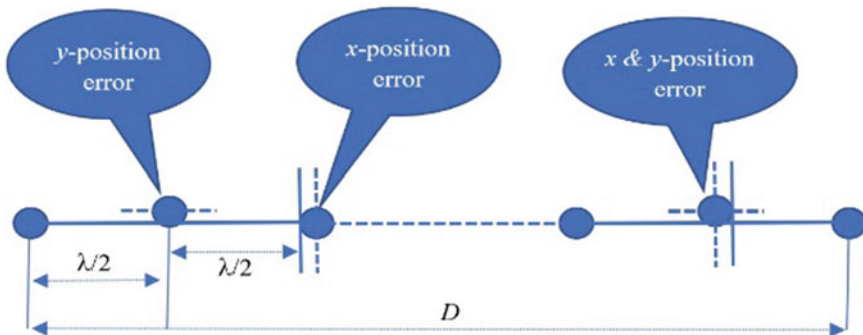


Fig. 9 Practical sonar array with possible positional errors during mounting

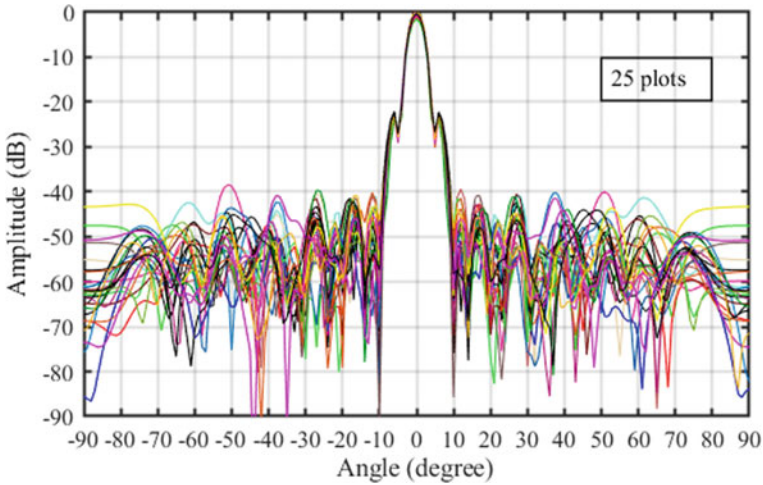


Fig. 10 Perturbation in simple product beamformer using rectangular and Dolph-Chebyshev

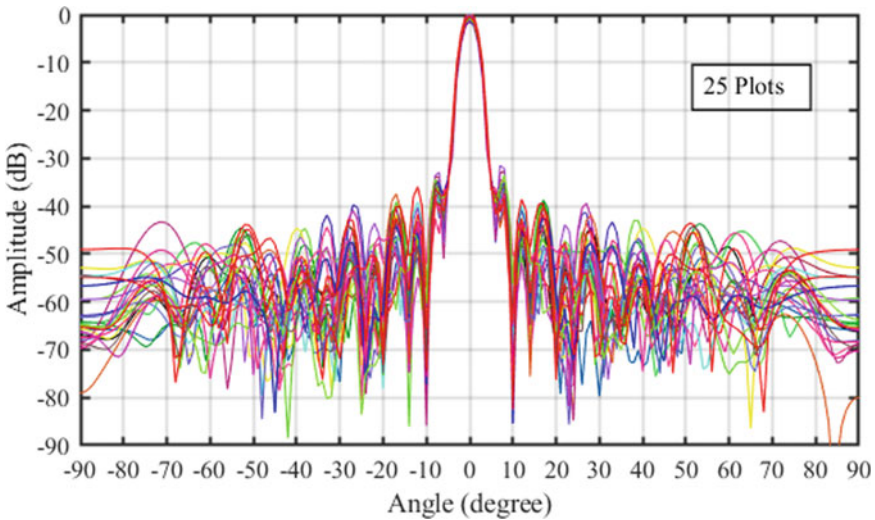


Fig. 11 Perturbation in product beamformer using rectangle and null steered Dolph-Chebyshev

5 Conclusion

The blending of window functions for sonar beamforming through the product of the beam pattern approach is investigated in this article. The four most commonly used functions are analyzed, and their effective usage is assessed for sonar applications. The designed beamformers are tested through simulation, and their tolerance to array deformations is examined. The obtained results suggest that the product

beamformer constructed using the blending of the rectangular and null steered Dolph-Chebyshev functions is the best and most tolerant to array deformations. The product beamforming is accomplished innovatively by applying the destructive interference principle used in optical physics.

Acknowledgements The authors thank Dr. Samir V. Kamat, Director General of Naval Systems and Materials, DRDO, Ministry of Defense for granting permission to publish the paper.

References

1. F. Le Chevalier, *Principles of Radar and Sonar Signal Processing* (Artech House, London, 2002)
2. C.A. Balanis, *Antenna Theory: Analysis and Design* (Wiley, New York, 2016)
3. J.E. Gaudette, J.A. Simmons, Observing the invisible: Using microphone arrays to study bat echolocation. *Acoustics Today* **10**(3), 16–25 (2014)
4. S. Kutty, D. Sen, Beamforming for millimetre wave communications: an inclusive survey. *IEEE Commun. Surv. Tutor.* **18**(2), 949–973 (2016)
5. R.C. Agarwal, V. Chander, S.P. Pillai, Issues in the design of towed array sonar systems. *IETE Tech. Rev.* **10**(2), 93–99 (1993)
6. R.J. Vaccaro, A. Chhetri, B.F. Harrison, Matrix filter design for passive SONAR interference Suppression. *J. Acoust. Soc. Amer.* **115**(6), 3010–3020 (2004)
7. H.A. Jackson, W.D. Needham, D.E. Sigman, Bottom bounce array sonar submarine (BBASS). *Naval Eng. J.* 59–72 (1989)
8. T.M. Buzug, J.-P. Babst, J. Ziegenbein, Image processing for activated towed array sonar displays, in *Proceedings of OCEANS '94*, vol. 2, pp. II/455-II/460 (1994)
9. H.M. South, D.C. Cronin, S.L. Gordon, T.P. Magnani, Technologies for sonar processing. *Johns Hopkins Appl. Tech. Digest* **19**(4), 459–469 (1998)
10. D.D. Ariananda, G. Leus, Direction of arrival estimation for more correlated sources than active sensors. *Signal Process.* **93**(12), 3435–3448 (2013)
11. R.G. Lorenz, S.P. Boyd, Robust minimum variance beamforming. *IEEE Trans. Signal Process.* **53**(5), 1684–1696 (2005)
12. R.A. Mucci, A comparison of efficient beamforming algorithms. *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-32**(3), 548–558 (1984)
13. B. Friedlander, A. Zeira, Detection of broadband signals in frequency and time dispersive channels. *IEEE Trans. Signal Process* **44**, 127–145 (1996)
14. S.M. Kay, Maximum Likelihood Estimation, in *Fundamentals of Statistical Signal Processing: Estimation Theory*, pp. 157–191 (Prentice-Hall, Englewood Cliffs, NJ, USA, 1993)
15. C.L. Dolph, A current distribution for broadside arrays which optimizes the relationship between beamwidth and side-lobe level. *Proc. IRE* **34**, 335–348 (1946)
16. S.W.A. Bergen, A. Antoniou, Design of ultraspherical window functions with prescribed spectral characteristics. *EURASIP J. Appl. Signal Process.* **14**, 2053–2065 (2004)
17. S. He, J.-Y. Lu, Sidelobe reduction of limited diffraction beams with Chebyshev aperture apodization. *J. Acoust. Soc. Am.* **107**(6), 3556–3559 (2000)
18. J.W. Adams, A new optimal window. *IEEE Trans. Signal Process.* **39**(8), 1753–1769 (1991)
19. P. Lynch, The Dolph-Chebyshev window: a simple optimal filter. *Mon. Weather Rev.* **125**, 655–660 (1997)
20. S. Vijayan Pillai, T. Santhanakrishnan, R. Rajesh, An efficient destructive interference based side lobe suppression method in SONAR beamforming. *Adv. Milit. Technol.* **16**(1), 107–120 (2021)

21. D.E. Dudgeon, D.H. Johnson, *Array Signal Processing: Concepts and Techniques* (Prentice-Hall, Englewood Cliffs, NJ, USA, 1993)
22. R.B. Blackman, J.W. Tukey, *The Measurement of Power Spectra* (Dover Publications, New York, NY, 1958)
23. F.J. Harris, On the use of windows for harmonic analysis with the discrete Fourier transform. *IEEE Proc.* **66**(1), 51–83 (1978)
24. H.D. Helms, Digital filters with equiripple or minimax responses. *IEEE Trans. Audio Electroacoust.* **AU-19**, 87–94 (1971)
25. A.H. Nuttall, Some windows with very good side-lobe behaviour. *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-29**, 84–91 (1981)
26. K.M.M. Prabhu, *Window Functions and their Applications in Signal Processing* (CRC Press, London, 2014)
27. N.C. Geckinli and D. Yavuz, Some novel windows and a concise tutorial comparison of window families. *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-26**, 501–507 (1978)
28. A.D. Poularikas, *The Handbook of Formulas and Tables for Signal Processing* (Springer-Verlag GmbH, Heidelberg, 1999)

Photonic Crystal Fiber Based Refractive Index Sensor for Cholesterol Sensing in Far Infrared Region



Amit Kumar, Pankaj Verma, and Poonam Jindal

Abstract Cholesterol is an important liquid in human body and everyone needs to balance the cholesterol level for living a healthy life style. A photonic crystal fiber (PCF)-based biosensor for sensing the cholesterol in far infrared region is proposed in this manuscript. A novel decagonal shape solid core photonic crystal fiber is designed. The sensing holes ring are placed in the solid core. For durability and stability of the sensor, Topas is the single fiber material. A perfectly matched layer (PML) is covered the PCF and also used as a scattering boundary condition. The numerical analysis is investigated by the finite element method (FEM). The highest sensitivity of the designed sensor is 97.32% for detecting the cholesterol. The parameters such as numerical aperture (NA), spot size (SPS) and beam divergence (θ_{bd}) are also analyzed. The designed decagonal PCF structure can be used for sensing the different biological and chemical analytes due to its simple structure and promising results.

Keywords Photonic crystal fiber · Refractive index biosensor · TOPAS material · FEM · Sensitivity · Numerical aperture

1 Introduction

Today, the need of fast and low-cost detection of biological substances is increasing exponentially. The sensing and identification of the biological substances are crucial to diagnose the patients in health sector. There are number of commercial biosensors used for detecting the biochemicals. The PCF-based sensors attract the researcher's interest because of its compact size, low-cost and rigidity. Environmental monitoring [1], biosensing [2], chemical sensing [3] are some common applications of the PCF-based sensors. Few biosensing applications of the PCF are blood cell detection like pH detection [4], glucose level detection [5], cholesterol detection [6], RNA analysis [7], cancer cell detection [8], and illegal drug detection [9], etc.

A. Kumar (✉) · P. Verma · P. Jindal

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

In last decade, the PCF structure like hollow core, solid core [10], porous core [11], hybrid core [12] have been developed. A solid core PCF (SC-PCF) is used as a biosensor in this article. The light traveled through the fiber by index guiding mechanism [12]. The terahertz frequency or T-rays (0.1–10THz) are opted because of its numerous applications in sensing and medicine [13]. Terahertz rays do not damage the substances of sensing analyte. The PCF is single material fiber having tiny holes in the cladding area. The arrangement of these tiny holes has the capability to control the propagation of light through the fiber. There are few materials like TOPAS, ZEONEX, Teflon and high-density polyethylene can be used as a fiber material. In addition, the high performance and designing flexibility of the PCF makes it a smart biosensor [14–16].

Cholesterol is a very important substance in our body that helps to build healthy cells. The human body gets the cholesterol from foods. The plants can not make it, the cholesterol can be finding in eggs, meats and the dairy product. The cholesterol serves in the production of bile in the liver, tissues and sex hormones in the human body. It is waxy substance found in blood. The high level of cholesterol leads to the heart disease, it creates the fatty deposits in blood vessels and making difficulty to the flow of blood through the arteries. Sometime, it forms the fatty clots that increase the risk of heart stroke. The high level of cholesterol has no symptoms. It is basically categories in good cholesterol and bad cholesterol. Cholesterol is attached with protein and it is called lipoprotein. The high-density lipoprotein (HDL) is a good cholesterol because it picks up the excess cholesterol and revert back to the liver. The low-density lipoprotein (LDL) is called bad cholesterol because it transports the cholesterol through the arteries and increase the risk of heart disease. The chemical formula of cholesterol is $C_{27}H_{46}O$ and belongs to steroid family. The estimated refractive index of the cholesterol is 1.525 [6, 17].

In literature, large number of PCF-based sensors has been developed [18]. Chopra et al. reported a PCF with diamond shaped for detecting the blood glucose, cancer tissue and other blood component [19]. In 2017, the authors proposed a blood sensor using circular solid core. The investigated sensitivity response for hemoglobin is 66.47% and CL is 1.928×10^{-8} (dB/m) for hemoglobin [20]. But they fail to gain significant sensitivity for all components. The sensors developed for near infrared region could not achieved the sensitivity more than 75%. The sensor operating in THz region overcome the low sensitivity. In 2004, the first THz sensor has been presented for biomedical application [21]. Md. Saiful Islam et al. [22] reported the THz chemical sensor using a ZEONEX-based hollow core PCF and gained the sensitivity of 96% for benzene, but this PCF structure has poor fabrication feasibility because of different dimension and rectangular shape of air holes. In 2019, authors have reported the blood sensor in THz regime with partial type-b crystalline structure in a symmetric manner for core and cladding [23]. They numerically achieved the maximum sensitivity of 80.93% for RBCs at $f = 1.5$ THz. The investigated effective area is $1.55 \times 10^5 \mu\text{m}^2$ to $1.85 \times 10^5 \mu\text{m}^2$ for different blood cells. They also investigate the different optical properties like V-parameter, SPS and θ_{bd} . In [24], authors reported a THz RI spectroscopy for chemical sensing with high sensitivity based on hollow core PCF, the structure with poor tensile strength is employed to

investigate the sensing performance and achieved very good results for the analyte RI range of 1.26–1.50. In the same year, Hossain et al. [25] have proposed a HC-PCF-based blood component detector and reported the sensitivity of 93% for RBCs. In 2021, Kumar et al. reported a SC-PCF-based blood sensor and also investigate the sensitivity, CL, EA and birefringence at THz frequency [26].

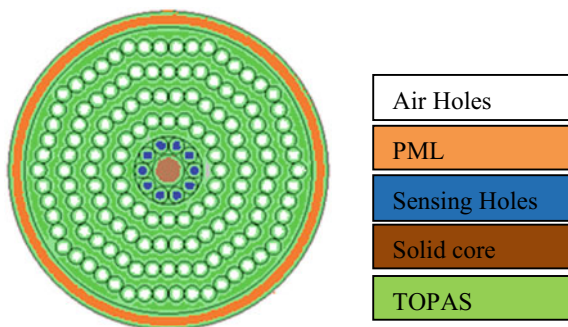
AS going through the literature, it clears that all the article covers the analyte RI range 1.26–1.50 and gained the maximum sensitivity of 87–95%. To the best of my knowledge, no article published to detect the cholesterol with solid core PCF. Thus, there is a scope to propose the highly sensitive RI biosensor with feasible fabrication possibilities.

In this paper, PCF-based refractive index sensor for cholesterol sensing in far infrared region is proposed. The SC-PCF with circular air holes arranged in a decagonal pattern is introduced for biosensing application in THz band. Cholesterol is taken as the biological sensing liquids. Numerically, the designed PCF-based sensor is achieved very high sensitivity, NA, SPS and beam divergence (θ_{bd}) at frequency range of 3.4–4.4THz. The article is further organized as follows; Sect. 2 describes the designed structure of the proposed PCF. In Sect. 3, methodology and numerical analysis are described. In Sect. 4, the important results and comparison with the previous study are elaborated. Finally, Sect. 5 concludes the paper.

2 Insight Geometry of the PCF

The proposed decagonal air holes pattern with solid core geometry of the PCF is shows in Fig. 1. In order to design a sensing ring in the core, the sensing holes in the same decagonal pattern are placed inside the core. The sensing holes act as the analyte core area, these holes are infiltrated with the sensing analytes that provides the better interaction to the evanescent field, the remaining core area have the same material as cladding. The solid core PCF allows the index guiding mechanism because the RI of the core is greater than the RI of cladding. To achieve the maximum sensitivity, cholesterol has been considered as the sensing analytes. In the cladding region, four

Fig. 1 Cross-sectional view of the proposed structure



layers of circular air hole are formatted in decagonal shape. These layers of air hole reduce the overall refractive index of cladding region, and it accentuates the light into the core area. The air holes are selected circular in shape. The size and pitches of the holes are selected optimally to ensure the confinement of optical energy into the core and minimize the radiation out of the core region. The same size and shape of the core and cladding holes make it fabrication friendly structure. The radius of holes $R_1 = 41 \mu\text{m}$, each air holes in cladding are separated by $D_1 = 87 \mu\text{m}$. The total radius of the core and whole PCF structure is $R_2 = 182 \mu\text{m}$ and $R_3 = 760 \mu\text{m}$, respectively. The PML is placed outside the cladding for absorbing the incident radiation without producing the reflection. TOPAS is the fiber material for the PCF, it has been shown the large tensile strength, high temperature and humidity resistance, more than 90% optical transmission, constant RI ($n = 1.53$) and insignificant material dispersion.

3 Numerical Results and Analysis

The PCF-based RI sensor performance depends on the evanescent field distribution which is shown in Fig. 2. Due to the light propagation through the core, the field distribution of the proposed structure is shown in Fig. 2. The high model field is required to ensure the powerful interaction between propagating light and the sensing analyte.

For numerical investigation, finite element method (FEM) is used to investigate the sensing performance of the proposed PCF sensor [27]. The analysis has been done on COMSOL Multiphysics software versus 5.4. In the proposed RI sensor, cladding holes are filled with air ($\text{RI} = 1$) and the core holes with cholesterol ($\text{RI} = 1.525$). According to the Beer-Lambert law, the optical power is modulated with the RI of the sensing analyte [22]. The effective mode index (N_{eff}) within the core is varied

Fig. 2 field distribution in a proposed PCF

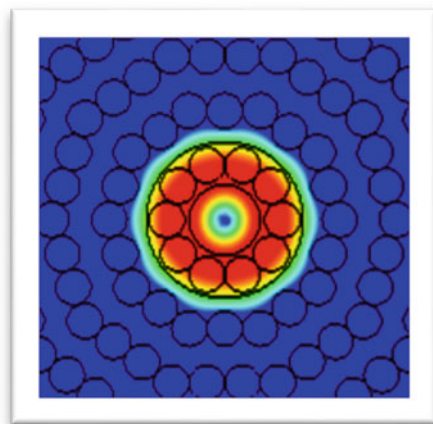
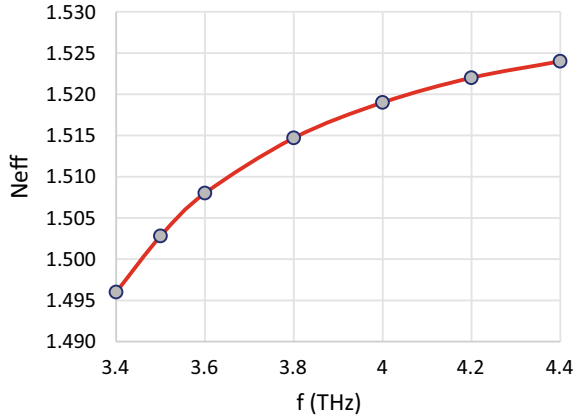


Fig. 3 Neff curve with frequency



around the sensing analyte RI 1.525 which is indicated in Fig. 3. As frequency is increasing, the N_{eff} is slightly increasing. The intensity of transmitted optical signal depends on the absorption (A) and it can be obtained by the following equation [25],

$$A = \log \frac{I_0(f)}{I(f)} = r\epsilon l C \tag{1}$$

Here, $I_0(f)$ denotes the incident optical power and $I(f)$ is the reflected optical power, r is the sensitivity coefficient, ϵ is molar absorption coefficient, l is the length of the fiber, and C is the concentration of the cholesterol in the liquid. The sensor performance is measured by the parameter sensitivity and it is expressed by the given function [26],

$$r = \frac{n_r}{n_{eff}} f \tag{2}$$

Here n_r and n_{eff} are the RI of the cholesterol and $Re|N_{eff}|$, f is the ratio of analyte fractional power to the total power which can be derived as [35],

$$f = \frac{\int (E_x H_y - E_y H_x) \partial x \partial y (\text{sample})}{\int (E_x H_y - E_y H_x) \partial x \partial y (\text{total})} \tag{3}$$

Here E and H denote the TE and TM fields. The sensitivity curve is obtained by the Eq. (2) and it is shown in Fig. 4. The sensitivity is followed the frequency and gained the maximum sensitivity of 97.21%. The light confinement in the core is getting strong at high frequency. So, the sensitivity is high at high frequency.

Numerical aperture (NA) is an important optical property for measuring the light compile capability of a PCF. The variation of the NA with frequency is shown in Fig. 5. In sensing performance, a wide NA is expected. The NA analysis is carried out by the refractive index difference of core and cladding as in equation [23],

Fig. 4 Sensitivity curve with frequency

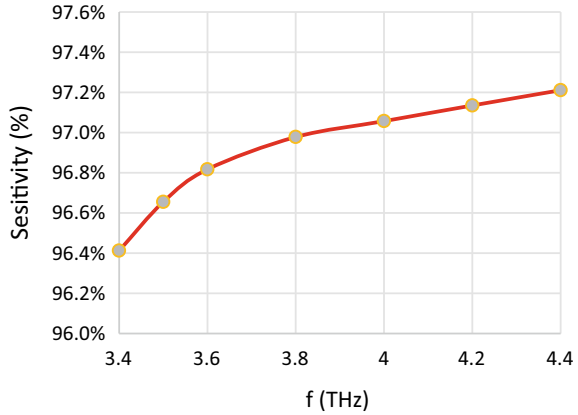
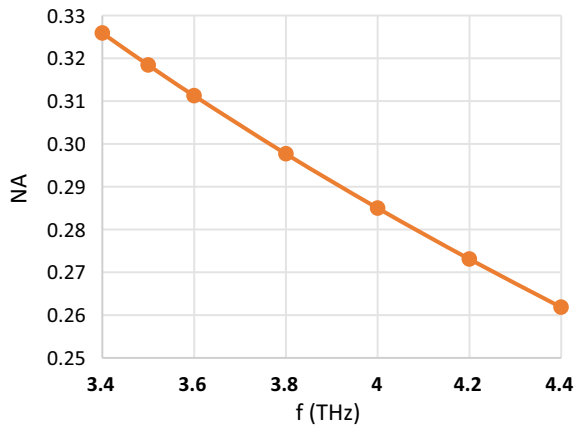


Fig. 5 NA curve with frequency



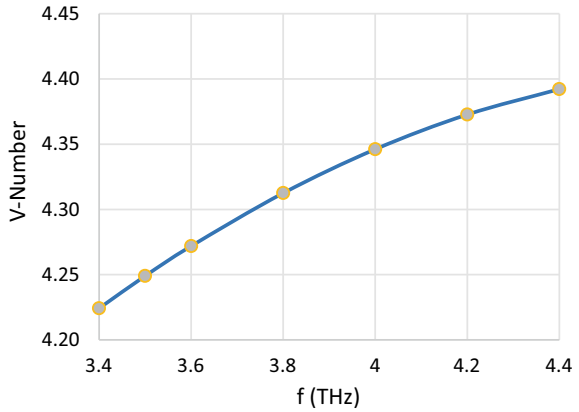
$$NA = \sqrt{n_{\text{core}} - n_{\text{clad}}} \tag{4}$$

Here, n_{core} and n_{clad} represents the RI of the core and cladding region in a PCF. The NA values are reduced slightly as the frequency increases, the negligible change in the NA value indicating the stability of the proposed sensor. The parameter depends on NA is known as V -parameter, it is a normalized frequency parameter and used to find the number of modes. V -parameter can be calculated as [23]

$$V = \frac{2\pi f}{c} r \sqrt{n_{\text{core}} - n_{\text{cladding}}} \tag{5}$$

Here r is radius of core, f is the operating frequency, and c is speed of light. For single mode propagation $V \leq 2.405$, and for multimode $V > 2.405$. The V -parameter curve with frequency is shown in Fig. 6. It has almost constant value near

Fig. 6 V-number curve with frequency



or above 4.00, so we can say the multimode transmission of optical power over this frequency spectrum.

Model SPS and beam divergence (θ_{bd}) are two crucial parameters for investigating the THz biosensor. Model SPS is closely related to the V -parameter and it can be obtained by the following equation [23]

$$W_{\text{eff}} = r \left(0.65 + \frac{1.619}{V^{\frac{3}{2}}} + \frac{2.879}{V^6} \right) \tag{6}$$

Here r is radius of core of the PCF and V is V -parameter. The SPS response with frequency is shown in Fig. 7, which is constant around 34 and negligibly decreasing with increasing frequency. The θ_{bd} in radian is also follow the SPS. The θ_{bd} in radian variation is shown in Fig. 8. SPS and θ are interrelated and both are related with tangent function to each other. According to the beam theory the BD is defined by

Fig. 7 Model SPS curve with frequency

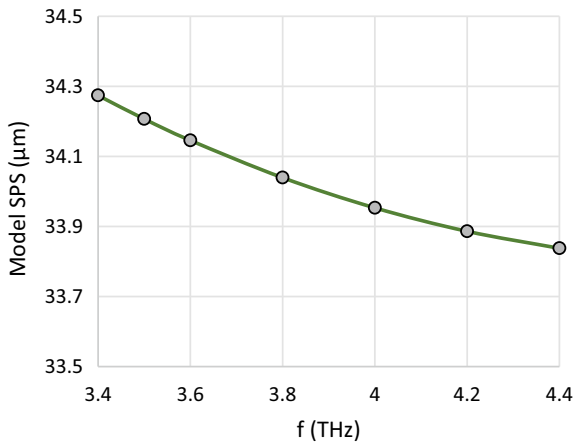
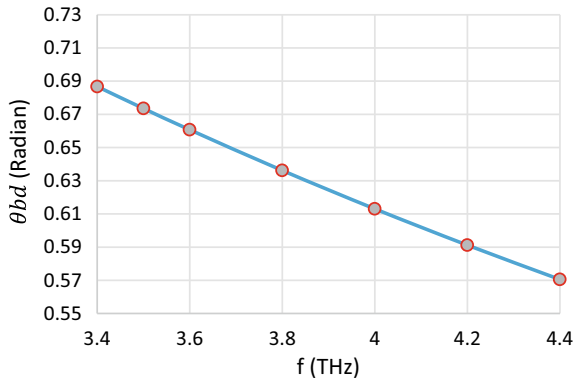


Fig. 8 θ_{bd} curve with frequency



the following equation [23]

$$\theta_{bd}(\text{radian}) = \tan^{-1}\left(\frac{c}{\pi f W_{\text{eff}}}\right) \tag{7}$$

$$\theta_{bd}(\text{degree}) = \theta_{bd}(\text{radian}) \times \left(\frac{180}{\pi}\right) \tag{8}$$

With the numerical analysis, the impact of the structural properties of the designed sensor on the sensing performance are studied and analyzed in the next section.

The effect of variation of sensing hole radius on sensitivity is described in Fig. 9. The sensitivity of the proposed sensor is decreased when radius of sensing holes increases above 40 m. the sensitivity is slightly varying with radius of the hole. The effect of sensing hole radius provides the optimized radius of the holes. The sensitivity of the sensor is increased with size of hole up to some extent after that it can decrease. Therefore, we can say that the 41 μm radius is the optimized radius for the proposed structure. The performance comparison the proposed sensor with reported sensor till date is listed in Table 1.

Fig. 9 Sensitivity curve with frequency with sensing hole radius 40, 41, 42

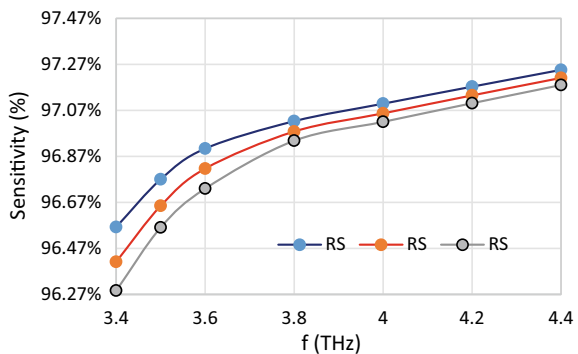


Table 1 Sensing performance comparison of the proposed and previously reported sensor

Reference	Frequency (THz)	Sensitivity (%)
[22], 2018	0.8–2.0	96.6
[23], 2019	1.5–3.5	81.93
[24], 2020	0.5–1.5	95.5
[26], 2021	1.0–4.0	87.68
Proposed work	3.4–4.4	97.21

4 Conclusion

In this article, a SC-PCF-based cholesterol sensor is proposed. The decagonal pattern is used for making the air hole layers in core as well as in cladding. The fiber is prepared by the TOPAS material and the outer layer of the PCF has been considered as the PML with scattering boundary condition. The FEM has been employed to obtain the numerical results. The results shown the outstanding sensitivity of 97.21% for cholesterol liquid. The manuscript also discussed the different optical properties of the PCF like NA, V -number, SPS and beam divergence. These properties help to consider the design feasibility of the sensor. The proposed sensor can be used for different application related to high refractive index liquid like chemical detection, biochemical detection. So, the proposed sensor is very useful in various industries like healthcare, chemical, marine and food industries.

References

1. N. Ayyanar, D. Vigneswaran, M. Sharma, M. Sumathi, M.M. Rajan, S. Konar, Hydrostatic pressure sensor using high birefringence photonic crystal fibers. *IEEE Sens. J.* **17**(3), 650–656 (2017)
2. R. Cheng, L. Xu, X. Yu, L. Zou, Y. Shen, X. Deng, High-sensitivity biosensor for identification of protein based on terahertz Fano resonance metasurfaces. *Opt. Commun.* **473**, 125850 (pp1–4) (2020)
3. S. Asaduzzaman, K. Ahmed, T. Bhuiyan, T. Farah, Hybrid photonic crystal fiber in chemical sensing. *SpringerPlus*, 5(748) (2016)
4. W. Li, H. Cheng, M. Xia, K. Yang, An experimental study of pH optical sensor using a section of no-core fiber. *Sens. Actuat. A.* **199**, 260–264 (2013)
5. A. Natesan et al., Tricore photonic crystal fibre based refractive index sensor for glucose detection. *IET Optoelectron.* **13**(3), 118–123 (2019)
6. S. Soylemez, Y.A. Udum, M. Kesik, C.G. Hizhates, Y. Ergun, L. Toppare. Electrochemical and optical properties of a conducting polymer and its use in a novel biosensor for the detection of cholesterol. *Sens. Actuat. B Chem.* **212**, 425–433 (2015)
7. B.M. Fischer, M. Hoffmann, H. Helm, R. Wilk, F. Rutz, T. KleineOstmann, M. Koch, P.U. Jepsen, Terahertz time-domain spectroscopy and imaging of artificial RNA. *Opt. Express* **13**(14), 5205–5215 (2005)
8. N. Ayyanar, G. Thavasi Raja, M. Sharma, D. Sriram Kumar, Photonic crystal fiber based refractive index sensor for early detection of cancer. *IEEE Sens. J.* **18**(17), 7093–7099 (2018)
9. A. Al-Mamun Bulbul, F. Imam, M.B. Hossain, M.A. Awal, E. Podder, H.S. Mondal, M.E. Rahaman, M.S. Ahmed, F. Iqbal, Highly sensitive photonic crystal fiber for illegal drugs

- detection in THz regime, in *11th ICCCNT 2020 July 1–3, 2020—IIT—Kharagpur*, pp. 49239 (IEEE)
10. P.S.J. Russell, Photonic crystal fibers. *Science* **299**(5605), 358–362 (2003)
 11. S. Sen, S. Chowdhury, K. Ahmed, S. Asaduzzaman, Design of a porous cored hexagonal photonic crystal fiber based optical sensor with high relative sensitivity for lower operating wavelength. *Photon. Sens.* **7**(1), 55–65 (2017)
 12. S. Asaduzzaman et al., Hybrid photonic crystal fiber in chemical sensing. *Springer plus* **5**(1), 748 (2016)
 13. H. Park, M. Cho, J. Kim, H. Han, Terahertz pulse propagation in plastic photonic crystal fiber. *Phys. Med. Biol.* **47**(21), 2634–2636 (2002)
 14. K. Nielsen, H.K. Rasmussen, A.J.L. Adam, P.C.M. Planken, O. Bang, P.U. Jepsen, Bendable, low-loss Topas fibers for the terahertz frequency range. *Opt. Express* **17**(10), 8592–8601 (2009)
 15. S. Sen, M. Abdullah-Al-Shafi, A.S. Sikder, M. Selim Hossain, M. Mohammad Azad, Zeonex based decagonal photonic crystal fiber (D-PCF) in the terahertz (THz) band for chemical sensing applications. *Sens. Bio-Sens. Res.* **31**(1–7), 100393 (2021)
 16. M. Goto, A. Quema, H. Takahashi, S. Ono, N. Sarukura, Teflon photonic crystal fiber as terahertz waveguide. *Jpn. J. Appl. Phys.* **43**(2B), Art. No. L317, (2004)
 17. M.M. Rahmana, F.A. Moua, M.I.H. Bhuiyanb, M.R. Islamc, Photonic crystal fiber based terahertz sensor for cholesterol detection in human blood and liquid foodstuffs. *Sens. Bio-Sens. Res.* **29**, 100356 (2020)
 18. T. Zhang, Y. Zheng, C. Wang, Z. Mu, Y. Liu, J. Lin, A review of photonic crystal fiber sensor applications for different physical quantities. *Appl. Spectrosc. Rev. Taylor Francis* **53**(6), 486–502 (2017)
 19. H. Chopra, R.S. Kaler, B. Painam, Photonic crystal waveguide-based biosensor for detection of diseases. *J. Nanophotonics* **10**(3), 036011 (2016)
 20. P. Sharma, P. Sharan, Design of photonic crystal-based ring resonator for detection of different blood constituents. *Opt. Commun.* **348**, 19–23 (2015)
 21. T.W. Crowe, T. Globus, D.L. Woolard, J.L. Hesler, Terahertz sources and detectors and their application to biological sensing. *Philos. Trans. R. Soc. A* **362**, 365–377 (2004)
 22. M. Islam, J. Sultana, A.A. Rifat, A. Dinovitser, W.-H.N. Brian, D. Abbott, Terahertz sensing in a hollow Core photonic crystal Fiber, *IEEE Sens. J.* **18**(10) (2018)
 23. K. Ahmed, F. Ahmed, S. Roy, B.K. Paul, M.N. Aktar, D. Vigneswaran, M.S. Islam, Refractive index-based blood components sensing in terahertz Spectrum, *IEEE Sens. J.* **19**(9) (2019)
 24. T. Yang, L. Zhang, Y. Shi, S. Liu, Y. Dong, A Highly birefringent photonic crystal fiber for terahertz spectroscopic chemical sensing. *Sensors* **21**, 1799 (2021)
 25. M.B. Hossain, E. Podder, Design and investigation of PCF-based blood components sensor in terahertz regime. *Appl. Phys. A* **125**, 861 (2019)
 26. A. Kumar, P. Verma, P. Jindal, Decagonal solid core PCF based refractive index sensor for blood cells detection in terahertz regime. *Opt. Quant. Electron.* **53**, 165 (2021)
 27. A. Gautam Prabhakar, et al., *Finite Element Analysis of Solid-Core Photonic Crystal Fiber* (IEEE, 2012)

An Efficient, Low-Cost, and Reliable Monostatic Microwave Imaging (MWI) Approach for the Detection of Breast Tumor Using an Ultra-Wideband Dielectric Resonator Antenna



Gagandeep Kaur and Amanpreet Kaur

Abstract In this article, a low-cost, unproblematic, and efficient microwave imaging technique have been proposed to detect the presence of breast tumor using an inverted “L” shaped ultra-wideband (UWB) dielectric resonator antenna. The proposed antenna has been used as a sensor to transmit and receive the radio frequency (RF) signals toward and from the target, i.e., human breast phantom. The proposed DRA has been designed and simulated in CST MWS with fractional bandwidth 70.96% (6.1–12.6 GHz) and maximum peak gain 3.92 dB at 12.5 GHz of frequency. To accomplished the tumor detection process in CST software, the proposed DRA is rotated around the breast phantom in circular manner in elevation from $0-\pi$ and azimuthal planes from $0-2\pi$ and backscattered signals (with and without presence of tumor inside breast phantom) are recorded at different positions. The recorded reflection parameters are proceed in different beam-forming algorithms: delay and sum (DAS) and delay multiply and sum (DMAS) to reconstruct the 2D image of the breast tumor in XY plane using MATLAB R2018a.

Keywords Microwave imaging · Bio-model · Breast tumor · Beam-forming algorithms CST · MATLAB

1 Introduction

Now days, breast cancer is the one of the most fatal diseases among females [1, 2]. That is mostly occurs when some unwanted malignant tissues start growing around the healthy body tissues [3]. To detect the presence of the breast tumor, several clinical methods are available such as X-ray mammography [4], ultra-sound [5], and magnetic resonance imaging (MRI) [5]. But all these existing methods have some shortcomings like ionized exposure of the radiation along with painful, costly and time-consuming treatment, etc. [6, 7]. Therefore, these methods are not used widely for the detection of breast cancer. An alternate method, i.e., monostatic microwave

G. Kaur (✉) · A. Kaur
Thapar University Patiala, Patiala, India

imaging (MWI) is gaining the attention of numerous researchers because of its simple, non-ionized exposure, cost-effectiveness, and high image resolution of the scanned tissues. It is essentially working on the principle of large deviation in dielectric properties of malignant and healthy tissue [8]. This method is preferred basically, due to its simplicity, cost-effectiveness, confronts (provides least amount anxiety to the patient), and availability to detect small sized cancerous cells [9, 10]. Different types of antennas are available and used in microwave imaging applications such as biconical [11], bow-tie [12, 13], fractal [14], Vivaldi [15], horn [16], and Monopole antenna [17] are available in the literature for use in MWI applications. But since, the detection of breast cancer requires conformable antennas mainly with UWB properties; the larger antenna structures cannot be used in the proposed research work. In monostatic MWI procedure, only single antenna is used as a sensor. It consists of an inverted “L” shaped ultra-wideband (UWB) dielectric resonator antenna (DRA) has been used as sensor to transmit and receive the signals toward and from the scanned body area of female breast. The proposed DRA has total thickness of 5.67 mm, i.e., 0.035 mm of ground, 1.6 mm of substrate, 0.035 mm of feed point, and 4 mm of dielectric material that has been proposed, designed (in CST), and fabricated (using photolithography) on a layer of FR4 substrate with thickness 1.6 mm. On the top and bottom of this substrate, a layer of copper metal with thickness 0.035 mm is deposited to use as feedline and ground plane, respectively. At last, an inverted “L” dielectric resonator (DR) of alumina with thickness 4 mm is stacked over the feedline. The proposed DRA excites an ultra-wideband response with impedance bandwidth of 6.5 GHz (6.1–12.6 GHz), voltage to standing wave ratio (VSWR) < 2, and a high peak gain of 3.92 dB.

The proposed DRA is used as a microwave sensor to transmit and receive the non-ionizing signals toward and from the bio-model of breast in two cases; presence and absence of tumor (4 mm) inside the phantom. For MWI setup, the proposed DRA is placed parallel to the breast phantom at some distance (8 mm). The bio-model consists three layers skin, fat and tumor and each have different value of dielectric constant and conductivity. Then, backscattered signals are recorded on the vector network analyzer (VNA) by the same DRA in terms of *S*-parameter responses by rotating it around the phantom from 0 to 2π in elevation and 0 to π in azimuthal plane with fixed interval of $\pi/6$. These recorded responses are used in different beam-forming algorithms mainly DAS and DMAS to remove the unwanted clutter signals, i.e., from fat and skin and only amplify the tumorous signals. These amplified data used in MATLAB R2018a to construct an image of the scanned breast tissues using.

The effectiveness of the proposed DRA sensor has been confirmed by comparing its performance with some existing antenna structures as given in Table 1. From Table 1, it is observed that the existing work on MWI is done mostly by using array of antennas (multistatic MWI), but the work done in present research work is suggested by using single antenna (monostatic MWI). Additionally, the proposed technique is able to detect smaller size (4 mm) tumor than the earlier microstrip patch antenna (MPA) detected tumor of large sized using a. Therefore, the proposed DRA is one step ahead of the existing research work.

Table 1 Comparison of the proposed sensor with existing antennas

Reference number	Type of antenna	Achieved frequency bands	Type of imaging used	Size of tumor (radius)
[18]	Biconical MPA	Na	Multistatic	>5 mm
[19]	L-shaped	3.7–9.35	Na	Na
[20]	H-shaped	3.47–8.42	Na	Na
[21]	MPA with defected ground	2.4–4.7 GHz	Monostatic	6.5 mm
Proposed antenna	L-shaped DRA with defected ground	6.1–12.6 GHz	monostatic	4 mm

2 Configuration of the Proposed DRA Structure

The main purpose of designing the ultra-wide band (UWB) dielectric resonator antenna (DRA) is to identify the presence of breast tumor by comparing the electrical properties between the normal and malignant breast tissues. For this aspect, an inverted “L” shaped UWB DRA along with 6.5 GHz bandwidth, 5.4 dB high peak gain, and high-resolution are achieved by the proposed DRA. The geometry of the proposed DRA configuration is optimized with L-shaped resonator, inverted L-shaped feed point, and a defected ground structure (DGS) are depicted in Fig. 1a–c. The proposed DRA is designed and modeled on a mechanically robust FR-4 lossy substrate that has total dimensions of $22 \times 28 \times 1.6 \text{ mm}^3$. It consists four different layers defected ground structure (DGS) followed by FR4 substrate ($\epsilon_r = 4.4$), feed-line, and “L” shaped dielectric resonator of alumina ($\epsilon_r = 9.8$). For the improvement of antenna’s results in terms of impedance matching, return loss and bandwidth, the dimensions of the proposed antenna is optimized and the optimized dimensions of the proposed DRA are as specified in Table 2.

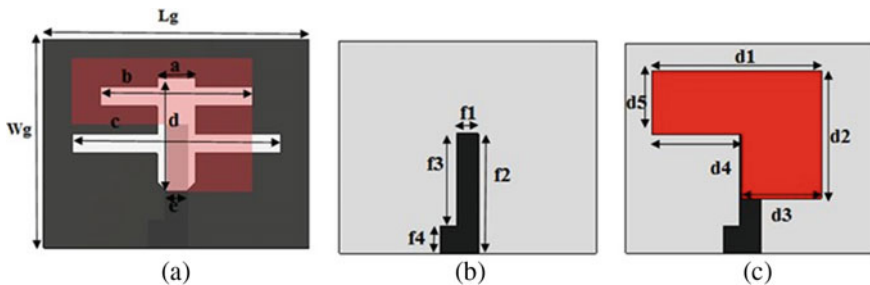


Fig. 1 Proposed DRA’s geometry **a** bottom layer (Ground), **b** intermediate layer (feedline), **c** topmost layer (DR)

Table 2 Optimized parametric values

Parameter	a	b	c	d	e	f1	f2	f3	f4	d1	d2	d3	d4	d5
Value	4	16	20	15	2.3	2.45	15.5	12.5	3	19	14	9	10	7

2.1 Design Procedure Followed to Attain the Desired UWB Characteristics

Lots of intermediate steps are followed to achieve the preferred UWB characteristics (6.1–12.6 GHz) from the conventional DRA as illustrated in Fig. 2a–d and their corresponding responses in terms of S-parameter as illustrated in Fig. 3. The conventional geometry (DRA_1), i.e., rectangular DR, simple microstrip feed line, and full ground plane (Fig. 2a) allows the antenna to excite only one resonances frequencies at 9 GHz with 8.4–9.5 GHz of impedance bandwidth. For bandwidth improvement,

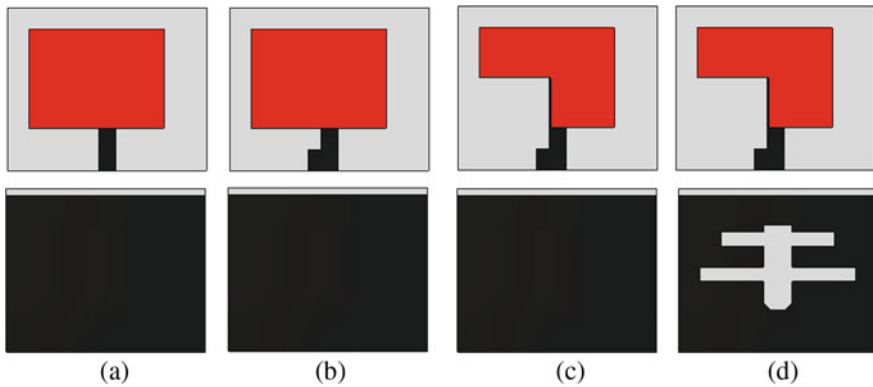


Fig. 2 Intermediate steps for the construction of proposed DRA

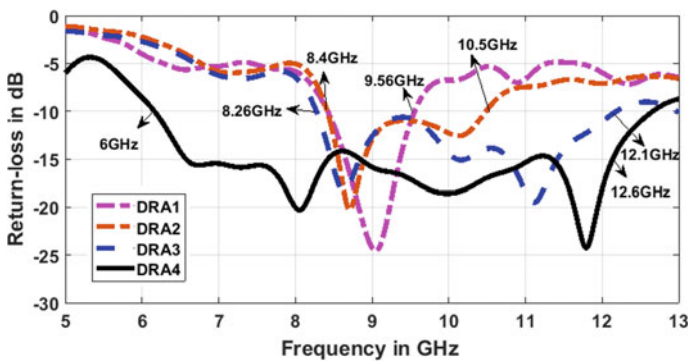


Fig. 3 S11 Parameter results of the four intermediate DRA geometries

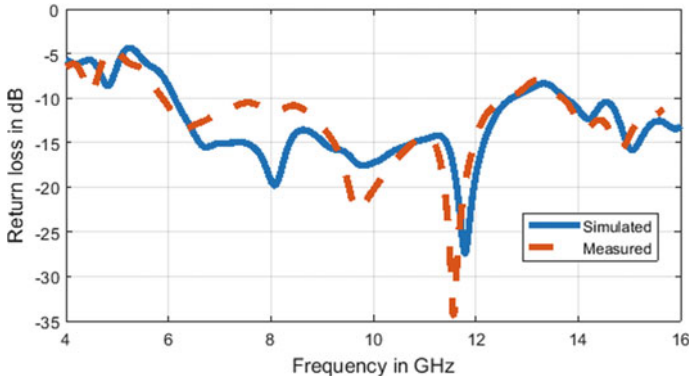


Fig. 4 Combined plot for simulated and measured results

a rectangular stub of dimension $3 \times 2 \text{ mm}^2$ is incorporated at the bottom end point of the feedline (DRA_2). It allows the DRA to excite wider frequency band from 8.4 to 10.5 GHz. To get better results in terms of bandwidth and return loss at upper frequency, a rectangular slot of dimensions $7 \times 10 \text{ mm}^2$ is etched from the bottom left surface of DR (DRA_3) as depicts in Fig. 2c which allows the antenna to excites 8.26-12.1GHz of frequency band. Finally, the ground surface of the proposed DRA is made defected by etching an airplane shaped slot (DRA_4). It allows the antenna to stimulate desired UWB characteristics from 6.1 to 12.6 GHz of frequency band (solid black color plot).

2.2 S-parameter Responses

Figure 4 shows the S-parameter (simulated and measured) comparison plot of the proposed DRA structure by taking operating frequency along X-axis and corresponding return loss along Y-axis. And it is proficient to excite ultra-wideband characteristics, i.e., 6.1–12.6 GHz with impedance bandwidth of 6.5 GHz along with trf frequencies, i.e., at 8 GHz and 11.8 GHz with return loss of -20 dB and -28.3 dB , respectively. Similarly, it shows measured S_{11} results, i.e., 6–7.68 GHz and 7.9–12.45 GHz along with 9.78 GHz and 11.7 GHz of resonant frequencies.

2.3 Broadband Gain and Voltage Standing Wave Ratio (VSWR) Plots

The broadband gain plot of the proposed DRA is plotted in between operating frequency (X-axis) and gain (Y-axis) for simulated and measured values as depicts in

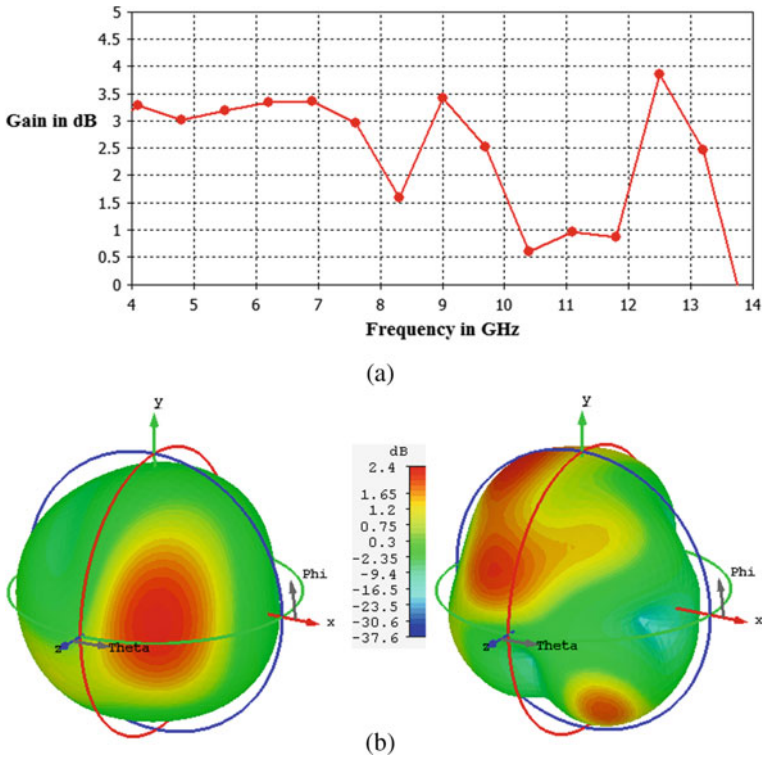


Fig. 5 Simulated results **a** broadband gain, **b** 3D gain plot

Fig. 5a and achieved the simulated gain 3.92 dB at resonant frequency of 12.5 GHz. The 3D gain at resonant frequencies of 8 and 11.8 GHz as illustrated in Fig. 5b having directivity of 3.27 dBi and 4.01 dBi, respectively. To check the impedance matching between the feed point and the antenna, VSWR is calculated. It is concluded that antenna shows excellent matching between the input and output of the signals by showing VSWR value between 1 and 2 for the desired frequency band.

2.4 D Radiation Pattern Plots

The 2D radiation pattern of the proposed DRA in E-filed at resonant frequency of 8 GHz and 11.8 GHz with main lobe magnitude of 1.83 dBi and 3.63 dBi, respectively, as illustrated in Fig. 6a, b. For the measurement of radiation pattern practically, a horn antenna that is used as transmitter placed at a distance of 1 m from the proposed DRA that is used as a receiver. The proposed DRA shows quite directional radiation characteristics so it can be a good candidate for microwave applications.

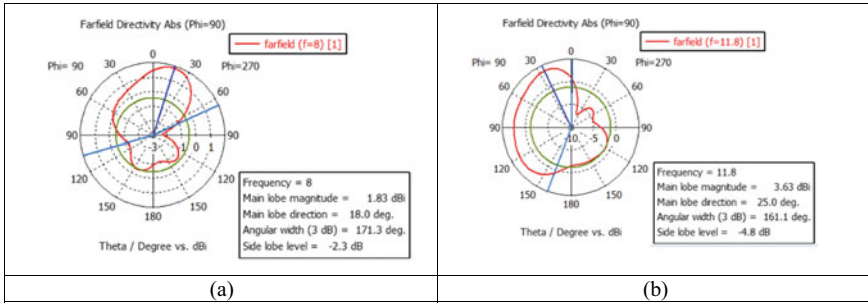


Fig. 6 Polar plot radiation pattern at a 8 GHz and b 11.8 GHz

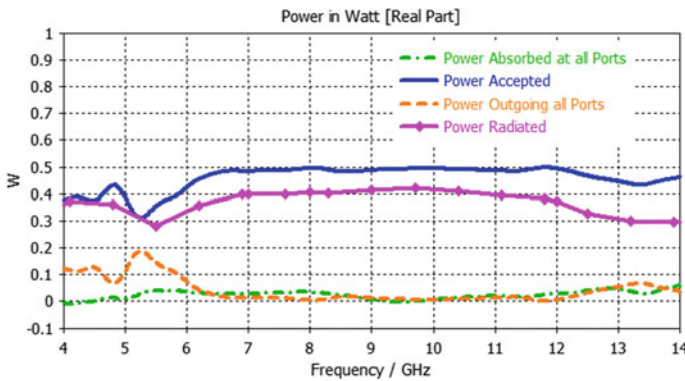


Fig. 7 Simulated power components at operating frequency range

2.5 Power Components for the Proposed DRA

The different power components, i.e., absorbed, accepted, outgoing, and radiated power of the proposed sensor are presented in Fig. 7. From Fig. 7, it is observed only small amount of power is absorbed at port of the antenna. The total radiated power is measured experimentally by an antenna is measured by placing the antenna under an anechoic chamber.

2.6 Radiation Efficiency

It is defined as the ratio of radiated to accepted power and represented a value between 0 and 1 for an ideal case. As depicted in Fig. 8 that the proposed DRA shows a quite good value for radiation efficiency, i.e., in between 0 and 2 which is a quite acceptable value for MWI applications.

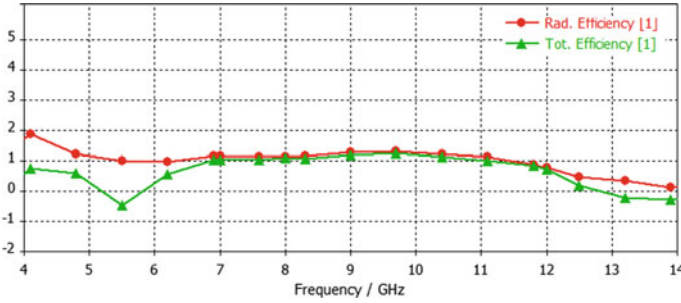


Fig. 8 Simulated radiation efficiency

3 Monostatic MWI Setup and Reconstruction of the 2D Image of Breast

For the simulation process in CST software, the designed DRA is placed parallel to the breast phantom at a distance of 8 mm as depicts in Fig. 9. The breast phantom consists a skin layer (4 mm thickness) followed by fat layer (34 mm thickness) and tumor layer (8 mm thickness). The proposed DRA is rotating around the breast phantom for two cases (i) with placement of tumor inside breast phantom from 0 to 2π in azimuthal plane at approximate focus points represent with black dots as $p1, p2, p3 \dots$ so on depicted in Fig. 7a and $0-\pi$ in elevation plane at focus points as $r1, r2, r3 \dots$ depicted in Fig. 7b.

For practical testing of the results, the proposed prototype with inverted “L” shaped DRA is connected to the VNA and rotates around the breast phantom (as done for simulation in CST) to collect the backscattered signals with fixed interval of 8.9 mm. The two set of data is recorded firstly with presence of tumor then without

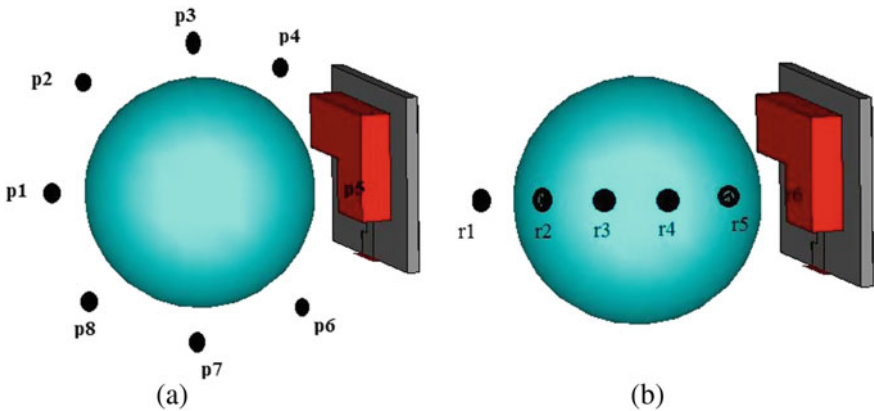


Fig. 9 Microwave imaging setup in CST a azimuthal and b elevation view

presence of tumor inside the breast phantom. A number of backscattered signals are recorded at different positions (with shift of 8.9 mm) that have to processed in two main beam-forming algorithms, i.e., delay and sum (DAS), delay multiply and sum (DMAS) to amplify the tumorous signals only. For the removal of unwanted clutter signals, i.e., skin and fat, the subtraction method is used as specified in Eq. (1) [22]. In this method, a tumor of 8 mm is inserted in the phantom at (0, 0) coordinates and reflections are recorded at each position named as S_{11}^{presence} . Then, reflections that are recorded without the placement of any tumor in breast phantom named as S_{11}^{absence} .

$$S_{11}(x, y) = S_{11}^{\text{presence}}(x, y) - S_{11}^{\text{absence}}(x, y) \tag{1}$$

These results are utilized in different beam-forming algorithms to get a clue about the presence of the breast tumor by amplifying only tumorous signals. But the subtraction method is not practically reasonable to distinguish the position of the breast tumor. So other interference removal beam-forming algorithms have been used to find out the position of the breast tumor that is explained in brief in subsection of this section.

3.1 Delay and Sum (DAS)

In this algorithm, initially, the received S -parameter responses are recorded at individual antenna positions (for example, at $p_1, p_2, p_3 \dots$ and $r_1, r_2, r_3 \dots$) by rotating it around the breast phantom and time-delays for all received signals are calculated based on different position of transmitter and receiver DRA named as $S_{11}(x, y)$. Afterward these time delayed signals are summed together as $S(t)$ according to Eq. (2) [23]. And get a noticeable difference among the backscattered values of presence and absence of the tumor. That amplify the energy level of the tumorous signals only in terms of return losses by adding signals them coherently and S -parameter responses from other cells, i.e., healthy cells are adding incoherently. Same procedure is done for all the focal positions within the breast phantom. Finally, a 2D image of breast tumor is reconstructed in MATLAB as shown in Fig. 10a and observed position of the tumor at (2, 3) in XY coordinates.

$$S(t) = \sum_{i=1}^N \sum_{j=1}^N s_{ij}(t + \text{delta}) \tag{2}$$

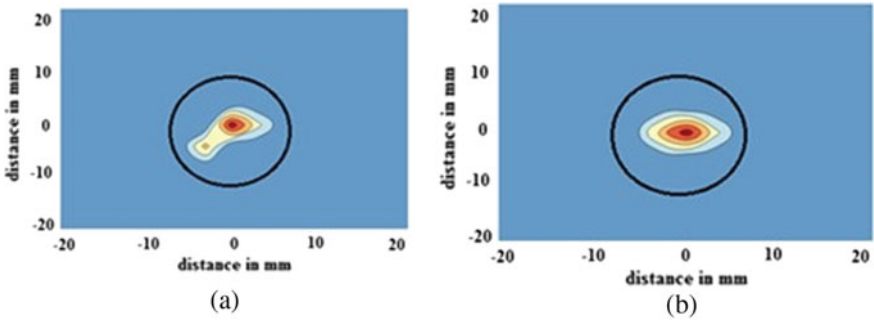


Fig. 10 2D image reconstructed using a DAS, b DMAS algorithms

3.2 Delay Multiply and Sum (DMAS)

DMAS method is the enhancement over the DAS method. In this method, addition of the pair-wise multiplications of the recorded backscattered data is used to amplify energy profile of the tumorous signals only as mentioned in Eq. (3) [24]. For each focal point, time alignment synthetically focuses received signals at that point. The 2D microwave image of breast tumor is reconstructed in the MATLAB software to get a clue about its position as illustrated in Fig. 10b at (1.5, 0) coordinates in XY plane. Where w is window size, N is number of DRA rotation, i and j are represent the different position of antenna around the breast model.

$$y(t) = \sum_{i=1}^{(N \times N)-1} \sum_{j=1}^{N \times N} X(t + \text{delta})Y(t + \text{delta}) \tag{3}$$

In Fig. 10a, b, the red color (existence of tumor) and blue color (absence of tumor) potion characterize the utmost and least amount of intensity of the processed data, respectively. Moreover, it is observed that DMAS provides a 2D breast image with better quality and approximately same positions as that of actual placement of tumor, i.e., (0, 0) by removing the back clutter, i.e., skin or fat reflections as compared to DAS algorithm.

The existing breast screening techniques are quite effective; but each of them has some shortcomings (as discussed in Sect. 1). Breast cancer detection using monostatic MWI has recently attracted the interest of many researchers and scientists. Recent clinical studies have demonstrated that MWI has the potential to become an alternative or additional tool for detecting and diagnosing the breast cancer at early stages. However, there are several problems arises for the experimental demonstrations of MWI technique like (i) the artificial breast models cannot exactly resembled with real human tissues; (ii) type of imaging method chosen for breast; (iii) proper and suitable selection of the operating frequency range; and (iv) resolution of the

breast image. To solve these challenges, a highly dynamic system should need be developed to incarcerate the minor to minor differences in the backscattered signals, i.e., S_{11} .

4 Conclusion

An UWB “L” shaped DRA for the detection of breast cancer using monostatic microwave imaging is proposed and investigated in this research article. The proposed DRA consists of a FR4 substrate, inverted L shaped microstrip feedline and a ground plane with DGS technique, designed to achieve 70.96% fractional bandwidth, 3.92 dB peak gain for resonant frequency of 12.5 GHz along with VSWR in between 1 and 2. To detect the breast cancer, proposed DRA is rotated around the phantom with $\pi/6$ interval in elevation ($0-\pi$) and azimuthal planes ($0-2\pi$) to record the S_{11} signals. And a significance deviation in S_{11} responses is compared among two cases firstly breast phantom with tumor then without tumor inside breast phantom that helps in the detection of breast tumor. DAS and DMAS algorithms are applied to the recorded S_{11} signals. The processed data are then plotted in MATLAB software to form a 2D image to depict the position of breast tumor in XY plane coordinates.

References

1. N. Howlander, A. Noone, M. Krapcho, D. Miller, K. Bishop, S. Altekruse, C. Kosary, M. Yu, J. Ruhl, Z. Tatalovich, Seer cancer statistics review Bethesda. MD: National Cancer Institute, vol. 19, pp. 1975–2013 (2016)
2. R. Siegel, D. Naishadham, A. Jemal, Cancer statistics. *CA Cancer J. Clin.* **63**(1), 11–30 (2013)
3. P.T. Huynh, A.M. Jarolimek, S. Daye, The false-negative mammogram, *Radiographics* **18**(5), 1137–1154 (1998)
4. P.J. Kornguth, F.J. Keefe, K.R. Wright, D.M. DeLong, Mammography pain in women treated conservatively for breast cancer. *J. Pain* **1**(4), 268–274 (2000)
5. L.L. Humphrey, M. Helfand, B.K. Chan, S.H. Woolf, Breast cancer screening: a summary of the evidence for the us preventive services task force. *Annals Intern. Med.* **137**(5_Part_1), 347–360 (2002)
6. C. Li, *Breast Cancer Epidemiology* (Springer, 2010)
7. E. Fear, M. Stuchly, Microwave detection of breast cancer (200). *IEEE Trans. Microw. Theory Techn* **48**(11), 1854–1863
8. W.T. Joines, Y. Zhang, C. Li, R.L. Jirtle, The measured electrical properties of normal and malignant human tissues from 50 to 900 MHz. *Med. Phys.* **21**(4), 547–550 (1994)
9. E.C. Fear, S.C. Hagness, P.M. Meaney, M. Okoniewski, M.A. Stuchly, Enhancing breast tumor detection with near-field imaging. *IEEE Microwave Mag.* **3**(1), 48–56 (2002)
10. R. Chandra, H. Zhou, I. Balasingham, R.M. Narayanan, On the opportunities and challenges in microwave medical sensing and imaging. *IEEE Trans. Biomed. Eng.* **62**(7), 1667–1682 (2015)
11. S.S. Tiang, et al., Radar sensing featuring biconical antenna and enhanced delay and sum algorithm for early stage breast cancer detection. *Progress Electromagn. Res. B*, 299–316 (2015)

12. X. Yun, E.C. Fear, R.H. Johnston, Compact antenna for radar based breast cancer detection. *IEEE Trans. Antennas Propag.* **53**(8), 2374–2380 (2005)
13. X. Li, M. Jalilvand, Y.L. Sit, T. Zwick, A compact double layer on-body matched bowtie antenna for medical diagnosis. *IEEE Trans. Antennas Propag.* **62**(4), 1808–1816 (2014)
14. D. Gibbins, M. Klemm, I.J. Craddock, J.A. Leendertz, A. Preece, R. Benjamin, A comparison of a wide-slot and a stacked patch antenna for the purpose of breast cancer detection. *IEEE Trans. Antennas Propag.* **58**(3), 665–674 (2010)
15. M.T. Islam, M.Z. Mahmud, N. Misran, J.-I. Takada, M. Cho, Microwave breast phantom measurement system with compact side slotted directional antenna. *IEEE Access* **5**, 5321–5330 (2017)
16. M.S. Nepote, D.R. Herrera, D.F. Tapia, S. Latif, S. Pistorius, A comparison study between horn and vivaldi antennas for 1.5–6 GHz breast microwave radar imaging, in *The 8th European Conference on Antennas and Propagation (EuCAP 2014)*, pp. 59–62 (2014)
17. K. Halili, M. Ojaroudi, N. Ojaroudi, Ultrawide band monopole antenna for use in a circular cylindrical microwave imaging system. *Microw. Opt. Technol. Lett.* **54**(9), 2202–2205 (2012)
18. S.S. Tiang et al., Radar sensing featuring biconical antenna and enhanced delay and sum algorithm for early stage breast cancer detection. *Progress Electromagn. Res. B* **46**, 299–316 (2013)
19. P. Suwanta, P. Krachodnok, R. Wongson, Wideband inverted L-shaped dielectric resonator antenna for medical applications, in *IEEE international conference on computational electromagnetic, Kumamoto*, pp. 188–189 (2017)
20. Z. Xu, S. Zhu, R. Wang, R. Xie, An H-shape dielectric resonator antenna with U-slot on the patch. *Progr. Electromag. Res. Sympos. (PIERS)* 4447–4450 (2016)
21. D. Bhaskaran, R. Krishnan, Breast tissue tumor analysis using wideband antenna and microwave scattering. *IETE J. Res.* 1–10 (2018)
22. D.T. Al-Zuhairi, J.M. Gahl, A. Al-Azzawi, N.E. Islam, Simulation design and testing of a dielectric embedded tapered slot UWB antenna for breast cancer detection. *Progress Electromag. Res. C* **79**, 1–15 (2017)
23. H. Bahramiabarghouei, E. Porter, A. Santorelli, B. Gosselin, M. Popovic, L.A. Rusch, Flexible 16 antenna array for microwave breast cancer detection. *IEEE Trans. Biomed. Eng.* **62**(10), 2516–2525 (2015)
24. H.B. Lim, N.T.T. Nhung, E.-P. Li, N.D. Thang, Confocal microwave imaging for breast cancer detection: delay-multiply-and-sum image reconstruction algorithm. *IEEE Trans. Biomed. Eng.* **55**(6), 1697–1704 (2008)

Effect of Link Reliability and Interference on Two-Terminal Reliability of Mobile Ad Hoc Network



Ch. Venkateswara Rao and N. Padmavathy

Abstract Reliability is quite possibly the main measure for arising advancements in communication network performance nowadays. The network reliability of mobile ad hoc network (MANET) is a function of link existence (single-hop/multi-hop) among the source and destination and the link reliability. The creation and deletion of communication links among the nodes are influenced by the dynamic topology nature of mobile ad hoc networks. In addition, these communication links may fail due to node failures, link failures, and interference. The presence of interference in an ad hoc network can vary the path from source to destination (direct path to intermediate path) which inherently exhibits diminish in the reliability of individual topology, which limits the network performance as well. It is decisive to address the effect of interference and link reliability on the link status and path status among network nodes. Therefore; a methodology that computes the two-terminal reliability of mobile ad hoc networks by considering the effect of the link reliability and interference has been proposed. The simulation results evidently designate that the network reliability falls down by 32% with an increase in interfering nodes from a single interference node to five interfering nodes. This work also identifies the finest value for the link reliability is 0.7.

Keywords Interference · Link failure · Link reliability · Mobile ad hoc network · Two-terminal reliability

Ch. Venkateswara Rao · N. Padmavathy (✉)
Vishnu Institute of Technology, Bhimavaram, Andhra Pradesh, India
e-mail: padmavathy.n@vishnu.edu.in

Ch. Venkateswara Rao
e-mail: venkateswararao.c@vishnu.edu.in

1 Introduction

Mobile ad hoc network (MANET) is an assortment of independent mobile nodes which are having wireless communication and networking capabilities that communicate with each other (directly or through intermediate nodes) without the aid of any centralized administrator or base station. These networks exhibit dynamically varying topology at every instant of time because of mobility [1]. The mobile nodes in a MANET must be unite and categorize themselves to offer routing and executive services for a specific application [2]. The typical characteristics such as infrastructure-less environment, multi-hop routing, dynamic topology, and constrained resources (limited bandwidth, battery power, etc.) of MANET attain the concentration researchers to contribute for this field [3]. The mobile nodes (source, terminal, and relay) in MANETs are outfitted with broadcasting transceivers which enable the communication through wireless channel. The communication link between source and destination has established either by single-hop or through multi-hop manner [4]. The creation and deletion of link between a pair of two nodes can be influenced by node failures, link failures, mobility, and presence of interference [5]. Generally, in MANET, the communication among the nodes does not depends upon predefined network environment as the topology varying dynamically. Moreover, the link existence is a function probability of existence of each configuration and path status from source to destination [6].

The reliability of a network can be viewed in terms of continuity of the network which is depends up on establishment of a successful communication link between source and destination. Generally, the reliability can be defined as the ability of enduring to implement a specific process regardless of the effect of faulty and damage. For the MANET, reliability measures are very crucial in order to obtained accuracy in network performance [7, 8]. The reliability analysis forever refers to the terminal reliability [9], which means that all the nodes in the network are able to connect, on the assumptions that there is a failure of link and node due to various reasons (interference, battery, and bandwidth). Hence, in order to implement two-terminal reliability, it is essential to identify the successful communication links among the nodes.

2 Literature Review

The research in the field of MANET has been predominant over past two decades because of their most versatile characteristics. The approach for evaluate reliability by assuming each node as a terminal is always referred to the two-terminal reliability. Numerous amounts of research have been carryout by various authors to evaluate terminal reliability of MANET in recent past. A Boolean algebra method to evaluate the terminal reliability has been proposed [10], which considers the link in the network is self-governing and time reliant and also nodes are absolutely reliable.

The node reliability in the network depends on mobility pattern and also depends on node failures, a recursive algorithm has been proposed [11] for evaluating terminal reliability. A mathematical method [12], which provides subordinate hop for the complication of two-terminal reliability has been proposed.

The dynamic nature of the topology in MANET is mainly due to node mobility, and the mobile nodes move with in their coverage area according to their mobility and exhibit arbitrary network topology. A methodology [13] based on Monte Carlo simulation in which the effect of mobility on reliability has provided. The numerous authors have developed various algorithms for evaluating network reliability by considering path reliability [14, 15], hop count [16] shadow fading environment for estimation reliability metrics [17], effect of scenario metrics (network size, coverage area, transmission range, etc.) [18], propagation-based link reliability model [19], effect of node failures [20], mobility models [21], etc. It has been observed that, the research contributions which deal with consequences of interference in mobile ad hoc network except reliability studies.

A mathematical model [22], which determines the network capacity in presence of interference, has been proposed. An algorithm [23] to decide the intrusion effect in mobile ad hoc network based on log normal model has been developed. A mathematical model has been proposed [24] which estimates the signal to noise plus interference levels in order to improvise network performance. In addition, several studies provide for estimating terminal reliability of MANET in various aspects are failed to consider the interference among the nodes. The presence of interference in the network creates sustainable impact on link formation and alters the path from source to destination. Hence, it is crucial to analyze the effect of interference on link formation (especially for multi-hop) among the nodes in order to achieve better network performance. Therefore, this work is focused on propose a methodology for evaluating two-terminal reliability of MANET by considering interference. The proposed methodology also identifies the effect of link existence probability on two-terminal reliability of MANET.

3 Assumptions

- The first and last nodes in the MANET are assumed as source and destination nodes, respectively.
- The mobile nodes are alike with maximum reliability ($r_i = 0.9$)
- The links between any two nodes are bidirectional
- Any node can act as an interfering node contained by normal nodes

4 Methodology

Traditionally, the mobile ad hoc networks are represented by a graph having n number of nodes connected directly (single-hop) or indirectly (multi-hop) with l communication links among them [25]. At every instant of time duration, MANET can change its topology due to the mobility of nodes. Every node in the network has an associated reliability which can be represented by (1).

$$p(n_i) = r_i \quad \forall i = 1 \text{ to } n \quad (1)$$

Further, each node in the network is connected by a means of communication link. The total number of possible communication links in a MANET is identified by (2). After knowing the number of links among the nodes, the corresponding network topologies are obtained by (3).

$$L = \frac{n(n-1)}{2} \quad (2)$$

$$|C| = 2^L \quad (3)$$

The existence of a link (l_{ij}) between a pair of nodes (n_i, n_j) is bidirectional for all $i, j = 1, 2, \dots, n$. Then, the formation and deletion of a link between a pair of nodes can be defined by using (4). When two neighbor nodes are within the transmission range of each other ($l_{ij} = 1$), then there is a link that exists, else no link has existed between nodes [26].

$$L_{ij} = \begin{cases} 1 & \text{if link exists} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

The existence of a link among the pair of nodes can be a probabilistic, and it has a value and it is represented by λ . The path between (S - D) pairs is influenced by the link failures, node failures, battery restrictions, and interference. The probability associated with each possible configuration is given by (5).

$$p(\alpha_k = 1) = \lambda^m \times (1 - \lambda)^{n_u} \quad (5)$$

The persistence of interference has an ability to divert the path status of each associated network configuration (α_k) [27]. The $2TR_{\alpha_k}$ defines the 2-terminal reliability of configuration α_k , $k = 1, 2, \dots, |C|$ can be defined by using (6). Where m is the interconnected nodes along the path.

$$2TR_{\alpha_k} = r_i^m \quad (6)$$

Finally, the $2TR_m$ can be evaluated by taking the mean of probability of existence for each configuration and associated reliability (with and without interference). The two-terminal reliability can be evaluated by (7).

$$2TR_m = \sum_{k=1}^{|C|} 2TR_{\alpha_k} \times p(\alpha_k = 1) \tag{7}$$

5 Illustrative Example

A fully connected MANET with six nodes has been considered to explain the proposed approach for attaining precision. This network is said to be reliable if a communicating path is available between source node and destination node (see Fig. 1).

The nodes in the network are alike with reliability $r_i = 0.9$. The link existence probability between the nodes is varying, and the value has been considered from 0.1 to 0.9 and $\lambda = 0.7$, then the non-existence would be $(1 - \lambda)$. For a 6-node network as exposed in Fig. 1; all possible communication links are 15 and possible network configurations are 32,768 which can be obtained by using (2) and (3). After all the network configurations have been generated, the probability of existence for each configuration is calculated using (5). The link reliability has been varied from 0.1 to 0.9. Table 1 highlighted the values of probability of existence of network configuration for various values of λ .

For each change in λ , the probability of network configuration has varied accordingly. The path from S - D pair may get changed in the presence of interference which leads to deviation of the shortest path. The concept of path status for both the cases has been provided in Table 2.

It is evident that, the nodes choose a direct path from a set of available paths, when there is no interference among the nodes. Whereas, considering interference among the nodes, the path status from source to destination changes direct path (single-hop) to indirect path (multi-hop), even though, there is a chance for establishment of direct path from source to destination. This inherently affects the individual configuration

Fig. 1 6 node complete network topology [27]

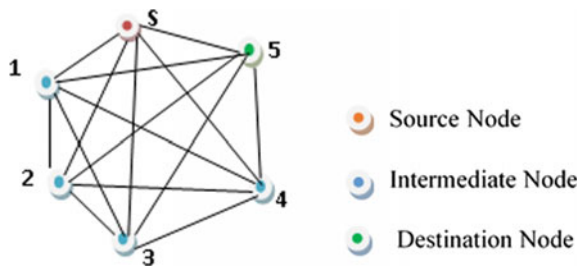


Table 1 Probability of existence of α_k with respect to λ

λ	$p(\alpha_k = 1)$
0.1	0
0.2	0
0.3	0.000000014
0.4	0.000001074
0.5	0.000030518
0.6	0.000470185
0.7	0.004747474
0.8	0.035184372
0.9	0.205891132

Table 2 Path status of each configuration with and without interference

Existing paths	Without interference		With interference		
	Shortest path	$2TR_{\alpha_k}$	Interfering nodes	Shortest path	$2TR_{\alpha_k}$
1-6; 1-2-6; 1-3-6; 1-4-6, 1-5-6; 1-2-3-6; 1-2-4-6 1-2-5-6; 1-3-4-6; 1-3-5-6; 1-4-5-6; 1-2-3-4-6; 1-2-3-5-6; 1-2-3-4-5-6	1-6	0.81	1	1-2-6; 1-3-6; 1-4-6; 1-5-6	0.729
			2	1-2-3-6, 1-2-4-6, 1-2-5-6, 1-3-4-6, 1-3-5-6, 1-4-5-6	0.6561
			3	1-2-3-4-6; 1-2-3-5-6	0.5904
			4	1-2-3-4-5-6	0.5314
			5	1-2-3-4-5-6	0.5314

reliability, which leads to degradation in network reliability (see Table 2). The reliability has been reduced from 0.81 to 0.729. Similarly, further increase in interfering nodes leads to increase the path status from source to destination which ultimately causes the decrease in reliability value. Finally, network reliability is evaluated by using (7).

6 Algorithm

An algorithm has been planned and developed based on the methodology (see Sect. 4) and illustrative example (see Sect. 5) in MATLAB R2018a having a processing speed of 3.9 GHz. The flow of pace involved in algorithm has been explained as flows.

- Step-1: Commence all the parameters such n, λ, r_i .
- Step-2: Calculates total number of links and network configurations by using (2) and (3), respectively.
- Step-3: Define the existence of a link among the nodes using (4).

- Step-4: Find the probability associated with each network configuration by varying the values of λ (0.1 to 0.9) by using (5).
- Step-5: Evaluate two-terminal reliability ($2TR_{\alpha_k}$) for each topology (with and without interference) by observing the path status from source to destination by using (6).
- Step-6: Finally, evaluate the two-terminal reliability of entire network ($2TR_m$) can be evaluated by using (7).

7 Simulation Results

The two-terminal reliability has been evaluated for a mobile ad hoc network by considering the interference among the nodes and also varying values of link reliability λ (0.1–0.9). Based on the illustrative example provided, the two-terminal reliability is a function of path status of (S - D) pair and also depends on probability of existence of configuration ($p(\alpha_k = 1)$) which is depends on λ . The simulation results depicted in Fig. 2 have clearly indicate that the network reliability is gradu-

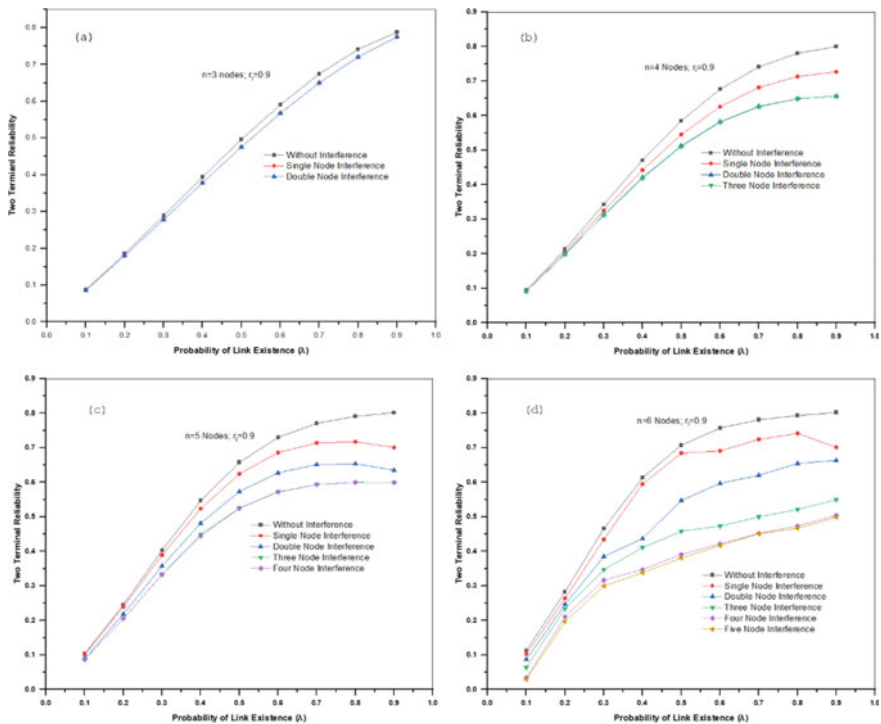


Fig. 2 Influence of link reliability and interference with **a** 3 nodes, **b** 4 nodes, **c** 5 nodes, **d** 6 nodes on two-terminal reliability analysis

ally increasing with increase in value of λ and network size (i.e., number of nodes), whereas, reliability decreases with increase in interfering nodes has been observed.

It is observed that the network reliability by considering single node interference and double node interference is typically equivalent (see Fig. 2(a)). This is because the network composed with 3 nodes has two path sets (1-3, 1-2-3) from source to destination. For example, the network having single node interference (say node 2 or node 3), then it should take path 1-2-3 instead of 1-3. Whereas, consider two interfering nodes (say node 2 and node 3), the path should be 1-2-3. In both the cases, the probability of existence of each configuration will be 0.729 and the corresponding reliability (0.6503) has been achieved for both single node and double node interference with $\lambda = 0.7$, whereas, for no interference case, the network reliability (0.6742) has been achieved. The network reliability has decreased by 2.4% when the network consists of single (double) interfering nodes. If the network size increased to 4 nodes, then the effect of interference on two-terminal reliability is much more significant when compared to 3 nodes. It is clearly depicted in Fig. 2b; the network reliability falls down by 6% with single node interference and 11% with double node interfering nodes. The network reliability has been increased by 7% with increase in network size from 3 to 4 nodes. Similarly, for the network with 5 nodes, the reliability has decreased with increase in interfering nodes.

It is observed from Fig. 2c, the network reliability with no interfering nodes was 0.73, and by considering 4 interfering nodes, the reliability achieved is 0.5245. Therefore, the network reliability has fallen down by 21% and 33% when the interfering nodes are increased to 4 nodes and 5 nodes (see Fig. 2c, d), respectively. Finally, the network reliability has increases from 0.6742 ($\lambda = 0.7$ and $n = 3$ nodes) to 0.7807 ($\lambda = 0.7$ and $n = 6$ nodes) with no interference consideration among the mobile nodes (see Fig. 2a, d). Whereas, the drastic decrease in the network reliability has been observed when interference among the nodes has considered. The network reliability falls down from 0.7807 to 0.4508 ($\lambda = 0.7$ and $n = 6$ nodes) with four nodes as interfering nodes. Hence, it is clearly identified that the network reliability has decreased by 32% with consideration of four interference nodes (see Fig. 2d). The simulation results prove that, the link reliability is directly proportional to the network reliability and also the network reliability has been increased (0.6742–0.7807) by 10% when the network size increased from 3 to 6 nodes. Whereas, the network reliability has been decreased by 32% when considering interference among the nodes. This study clearly demonstrates that the design engineers need to take at most attention while taking decision among network size and interfering nodes.

From the simulation result (see Fig. 2), it is observed that, the link reliability (λ) is proportional to probability of existence of network configuration ($2TR_{\alpha_k}$). Hence, the selection of λ has a predominant role in evaluation of two-terminal reliability. It is observed from Fig. 3, the simulation time required for evaluating two-terminal reliability is increasing with increase in network size. The simulation time is not only dependent on the network size; it will also vary with link reliability. The time taken for the evaluating two-terminal reliability with 3 nodes is 0.0452 s ($\lambda = 0.7$) and 15.73 s with 6 nodes. Similarly, the time taken for the evaluating two-terminal reliability with 3 nodes is 0.0694 s ($\lambda = 0.4$) and 17.392 s with 6 nodes (see Fig. 3).

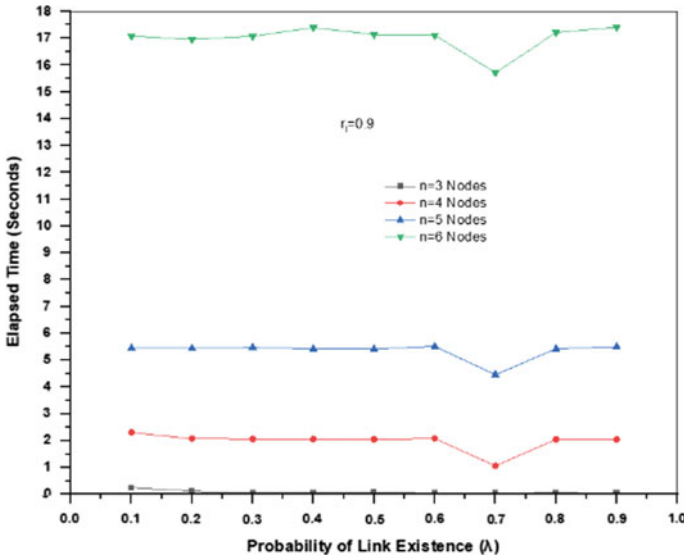


Fig. 3 Simulation time for evaluating two-terminal reliability

Finally, it is understood that the best suited value of λ is 0.7, where the time required for simulation is less and also the values of probability of existence of each topology has been normalized.

8 Conclusion

The two-terminal reliability analysis of mobile ad hoc network by considering various values of link reliability (λ) and interference among the nodes has been analyzed. A methodology for evaluating two-terminal reliability that considers the effect of interfering nodes on path status from source to destination and identifies the effect of λ the probability of network configuration has been developed. A reliability of 0.7807 has been achieved with no interference among the nodes with probability link existence chosen as 0.7. On the other hand, the reliability falls down 32% (0.4508) by considering interference among the nodes. The network reliability increased by 31% (0.6131–0.7807) with the value of λ operated from (0.4–0.7) for six node networks. The simulation results indicate that, the perfect value for λ is 0.7, which is proven in terms of CPU time, normalized values for probability of existence of network configuration.

References

1. C. Perkins, Ad hoc networking: An introduction (Chapter 1), in *Ad Hoc Networking* (Addison-Wesley Longman Publishing, 2001)
2. K. Pahlavan, P. Krishnamurthy, *Principles of wireless networks* (Prentice Hall, Englewood Cliffs, 2002)
3. D.S. Gurjar, P.K. Upadhyay, D.B. da Costa, R.T. de Sousa, Beamforming in traffic-aware two-way relay systems with channel estimation error and feedback delay. *IEEE Trans. Veh. Technol.* **66**(10), 8807–8820 (2017)
4. D.S. Gurjar, H.H. Nguyen, P. Pattanayak, Performance of wireless powered cognitive radio sensor networks with nonlinear energy harvester. *IEEE Sensors Lett.* **3**(8), 1–4 (2019)
5. M. Mandloi, D.S. Gurjar, P. Pattanayak, H. Nguyen (eds.), *5G and beyond wireless systems: PHY layer perspective* (Springer Nature, Singapore, 2020)
6. S.G. Datey, A. Taha, Mobile ad hoc networks its advantages and challenges. *Int. J. Electr. Electron. Res.* **3**(2), 491–496 (2015)
7. S.K. Chaturvedi, N. Padmavathy, Mobile ad hoc network reliability: an imperative research challenge (Chapter 4), *Advances in Reliability and System Engineering* (Springer International Publishing, 2017)
8. X. Xiao Chuan, W. Gang, W. Keping, J. Shilou, Link reliability based hybrid routing for tactical mobile ad hoc network. *J. Syst. Eng. Electron.* **19**(2), 259–267 (2008)
9. N. Padmavathy, An Efficient distance model for the estimation of the mobile ad hoc network reliability, in *International Conference on Intelligent Computing and Communications Technologies* (2020), pp. 65–74
10. L. Fratta, U. Montanari, “A Boolean algebra method for computing the terminal reliability in a communication network,” *IEEE Trans. Circ. Theory*, **20**(3), pp. 203–211 (1973)
11. L. Fratta, U.G. Montanari, A recursive methods based on case analysis for computing network terminal reliability. *IEEE Trans. Commun.* **26**(8), 1166–1177 (1978)
12. A. Majid, A. Mishra, Efficient two-terminal reliability calculations for mobile ad hoc networks, in *International Conference on Communication and Computing (ICCC-2014)* (2014), pp. 33–42
13. N. Padmavathy, S.K. Chaturvedi, Reliability evaluation of mobile ad hoc network: with and without mobility considerations. *Proc. Comput. Sci.* **46**, 1126–1139 (2015)
14. S.B. Venkata, N. Padmavathy, A systematic approach for analyzing hop count and path reliability of mobile ad hoc networks, in *International Conference on advances in computing, Communications and Informatics* (2017), pp. 155–160
15. B. Venkatasai Kumar, N. Padmavathy, A hybrid link reliability model for estimating path reliability of mobile ad hoc network. *Proc. Comput. Sci.* **171**, 2177–2185 (2020)
16. N. Padmavathy, A.S. Sri Vani, Effect of network parameters on hop count estimation of mobile ad hoc network, in *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)* (2019), pp. 1–8
17. N. Padmavathy, S.K. Chaturvedi, Reliability evaluation of capacitated mobile ad hoc network using log-normal shadowing propagation model. *Int. J. Reliab. Saf.* **9**, 70 (2015)
18. S.K. Chaturvedi, N. Padmavathy, The influence of scenario metrics on network reliability of mobile ad hoc network. *Int. J. Performability Eng.* **9**, 61
19. N. Padmavathy, S.K. Chaturvedi, Evaluation of mobile ad hoc network reliability using propagation-based link reliability model. *Reliab. Eng. Syst. Saf.* **115**, 1–9 (2013)
20. N. Padmavathy, J.R.C. Teja, S.K. Chaturvedi, Performance evaluation of mobile ad hoc network using monte carlo simulation with failed nodes, in *Second International Conference on Electrical, Computer and Communication Technologies* (2017), pp 1–6
21. N. Padmavathy, K. Anusha, *Dynamic reliability evaluation framework for mobile ad-hoc network with non-stationary node distribution* (Communication and Computing Systems, CRC Press Taylor and Francis, 2018), pp. 333–342
22. R. Hekmat, P.V. Mieghem, Interference in wireless multi-hop ad-hoc networks and its effect on network capacity. *Wireless Netw.* **10**, 389–399 (2004)

23. R. Hekmat, P.V. Miegheem, Interference power sum with log-normal components in ad-hoc and sensor networks, in *3rd International Symposium on Modeling and Optimization in Mobile Ad Hoc and Wireless Networks* (2005), pp.174–182
24. J.P. Mullen, *A Proposed Method to Estimate Signal to Noise Plus Interference Levels in Order to Improve the Performance of Mobile Ad Hoc Network Routing Protocols*. Center for Stochastic Modeling, Industrial Engineering Department, New Mexico State University (2005)
25. J.L. Cook, J.E. Ramirez-Marquez, Two-terminal reliability analyses for a mobile ad hoc wireless network. *Reliab. Eng. Syst. Saf.* **92**(6), 821–829 (2007)
26. J.L. Cook, J.E. Ramirez-Marquez, Reliability analysis of cluster-based ad-hoc networks. *Reliab. Eng. Syst. Saf.* **93**(10), 1512–1522 (2008)
27. C.V. Rao, N. Padmavathy, S.K. Chaturvedi, Reliability evaluation of mobile ad hoc networks: with and without interference, in *IEEE 7th International Advance Computing Conference* (2017), pp. 233–238

Security

User Information Privacy Awareness Using Machine Learning-Based Tool



Aaditya Deshpande, Ashish Chavan, and Prashant Dhotre

Abstract Today, web services are used by everyone to carry out everyday activities online. In the exchange of these free web services, the user information is collected extensively by service providers that leads to user information privacy concerns. The principles of privacy must be followed by service providers and must adhere to them and communicate with users through their privacy policy. However, this privacy policy document is a legal document that is broad and tiresome to read and understand. This nature of the document drives a user to accept the policy without understanding which may lead the user to release their data or even get sold in the worst cases that puts users' privacy at high risk. To overcome this risk, this paper presents a semi-automatic machine learning-based tool that analyses the privacy policy document and performs text categorization. This technique divides the policy document into meaningful categories by classification algorithms. It will help the user to understand what they are collecting and when. Additionally, this tool will inform users the purpose of data collection chosen by the service providers. This paper presents the results that will enhance the privacy knowledge of users so that their privacy will be protected.

Keywords User information · Privacy · Privacy policy · Machine learning · Security

1 Introduction

Whenever there is personal data involved, it will surely lead to the privacy and proper use of that data. A similar situation occurs when the number of users avails for a service from a service provider. For that communication purpose, a legal document is created that document is the privacy policy. But this document also contains a lot

A. Deshpande (✉) · A. Chavan
Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, India

P. Dhotre
MIT School of Engineering, MITADT University, Pune, India

of technical terms and cryptic words that makes it difficult for a user to understand what is written in it. This leads the user to agree on it and continue to the service [1]. If the user denies the policy, they will not get access to that service. This process makes the user unaware of the terms specified in the policy.

In the year 2020, due to the COVID-19 virus outbreak lead, one online meeting application named “zoom” became popular. The iOS version of that application reported sending data to Facebook, but the application’s privacy policy is not explicit about such data transfer [2]. But no user was aware that their audio was being recorded and listened by Microsoft itself [3]. The reason is the user did not read and understood the privacy policy made by Microsoft.

To avoid such cases European Union’s General Data Protection Regulation (GDPR) suggested service providers to reduce the reading level of their privacy policy. But the surprisingly word count and their respective time were increased in order to reduce complexity [4, 5]. But still, it was not addressing the problem there.

This paper discusses a method to increase user awareness by properly classifying a document into several categories. For that purpose, 50 privacy policy has been analysed and found such categories that most of the privacy policies have [6]. This paper proposes a machine learning-based tool for classifying that legal document into its respective sections. This will make users aware of the clauses in the privacy policy. Also, this tool will provide that document reading time making the user understand the actual complexity of that document.

2 Literature Survey

There is a lot of research going on in this area, most of them focuses on designing a policy in a much better way [7]. However, designing a policy from the users’ point of view is a difficult task as the policy context may vary from service to service. There was a study for calculating completeness of privacy policy based upon the categories that they have mentioned [8]. They have proposed 8 categories that most of the privacy policies should have and depending upon the presence of each category they have determined the complete score of that privacy policy.

In one study it is found that the yearly amount to read privacy policy is \$781 billion in the US [9]. This much amount is because of the recurring charges and complexity in understanding policy. This also shows the severity of the problem from an industry point of view. To reduce the size of privacy policy there is a study that suggests that terms which are already aware of a user should be removed and only new once to be shown [10]. But some privacy awareness studies have shown that users are not aware of clauses present in policy [11, 12]. So, removing certain clauses will mislead the user about policy. In a research, trustworthiness between a user and service providers was demonstrated. The machine learning based tool was presented to trust score [13]. There are different trends, challenges, and opportunities in user privacy and empowerment are presented in a research that includes privacy policy analysis and visualization [14]. The issues includes privacy policy understanding,

analysis, visualization, etc. The privacy and security issues are not only limited to industry but also to users as an individuals [15]. Hence, the issues of privacy and cybersecurity are presented in a reaserch work. The evolving world of the Internet of Everything involves machine-to-machine communication to provide widespread access to users or service providers [16]. However, ubiquitous systems have several problems of efficient ubiquitous system architecture, security and confidentiality problems in contextually aware ubiquitous systems.

3 Methodology

The machine learning model is trained and tested on a dataset to measure its performance. The model which gives the best performance and accuracy is selected. The optimum model is used in websites to show the categorization of sentences and readability score to users.

3.1 Preparation of Dataset

The dataset is created by collecting the top 50 most visited sites in India. Sentences are categorized into respective clauses. The dataset is human-annotated. The model is then trained and tested on this dataset.

There is a total of 8 categories or clauses.

1. Information collected
2. Children data collected
3. Cookies
4. Security
5. Purpose of sending
6. Third-party information shared
7. Way of collection
8. Contact information

3.2 Applying Preprocessing and Splitting Data

Below are the steps included in preprocessing.

- Cleaning the text
- Stop-words removal
- Stemming
- Generating N-gramms
- Count vectorizer

Table 1 Voting classifier accuracy table

Classifier	Accuracy	Precision	Recall	F1 score
Random forest classifier	90.24	0.91	0.90	0.90
Multinomial NB				
NuSVC				
Logistic regression				

- TF-IDF scoring

For testing and training purposes, the dataset is split into 70 and 30%. 70% data for training purpose and 30% for testing with stratification technique.

3.3 Classification

Naïve Bayes provides accurate results and performance within less training. K-nearest neighbours groups text in predefined labels based on internal content. KNN does not use probabilities and is efficient. Decision tree makes decisions on a given set of choices. The advantage of the decision tree is visibility of how choices are made. Support vector machines are also called universal learners [17].

Ensemble learning combines various machine learning algorithms to produce one optimized algorithm. The resultant model contains the advantages and strength of all included algorithms. This tool uses a voting classifier. NuSVC, random forest, multinomial Naïve Bayes, and logistic regression are the best performing according to accuracy, log loss, time, and f1 score (Table 1).

Voting classifier is having overall better performance than individual models in the system architecture (Fig. 1).

4 Proposed Solution

This paper proposes building a website that will aid in creating awareness among users. Privacy policies of the top 50 sites in India are collected and stored in a database.

4.1 Analysing Privacy Policy

Users will visit the site and select the site whose privacy policy is to be analysed. Websites will fetch those sites from the database and users will be presented with a

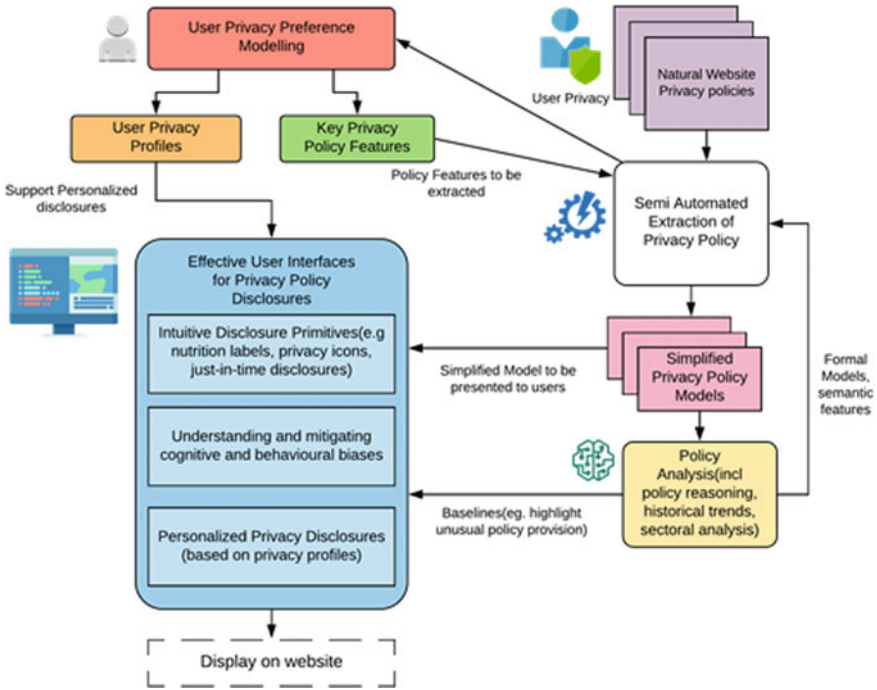


Fig. 1 System architecture

categorized policy with better readability and understandability. A readability score will be provided which will help users to understand the complexity of the document.

This site uses React.js for front end, flask for backend and MongoDB to store collected privacy policies. The user interface is handled by React.js. The aim of the front end is to make users as comfortable as possible and navigation easier. Web server flask handles requests and runs machine learning models and sends responses to the front end, which is then presented in 8 categories to the user. If the site is selected from the existing 50 websites from dropdown these sites are fetched from MongoDB. If a site is not present in dropdown users can copy-paste the privacy policy in the text area and it will be analysed.

4.2 Readability Score

A readability score is also provided which will educate users regarding the time taken to read the document and its complexity. This tool uses various readability scores such as Flesch Reading Ease, Flesch-Kincaid Grade Level, Fog Scale (Gunning FOGFormula), SMOG Index, Coleman-Liau, Automated Readability Index, Linsear Write Formula.

5 Results and Discussion

The main goal of this project is to create awareness among users who use online services. This site will result in education and awareness of users regarding the privacy policy. Users will actually understand what data, sites are collecting and how sites deal with data. This approach will make the user more aware easily by classification of policies and providing readability.

An ensemble model is used for machine learning. Algorithms used in ensemble models are NuSVC, random forest, multinomial Naïve Bayes, and logistic regression. The overall performance of the voting classifier is good (Fig. 2).

From the figures, in this case NuSVC, logistic regression, multinomial NB, and random forest classifier are top-performing classifiers (Table 2).

Log loss penalizes false classifications of the classifier and quantifies accuracy. Lower log loss is an indication of a good model and higher log loss indicates higher entropy. In this evaluation, random forest, multinomial NB, logistic regression, and NuSVC have the lowest log loss signalling these are good models. K-nearest classifier and decision tree classifier performed worst. In terms of predicting time, duration of multinomial NB and logistic regression wins while NuSVC takes long. When the model is tested NuSVC, logistic regression and multinomial NB have good *F1* scores.

5.1 Screenshots

In this feature, the user can select a privacy policy from a dropdown menu. Web sites mentioned in the lists are already collected in the database, and when a user asks for them, then, they are processed and sent back to the front end (Figs. 3, 4, 5 and 6).

The user can copy and paste the policy for which analysis is to be done.

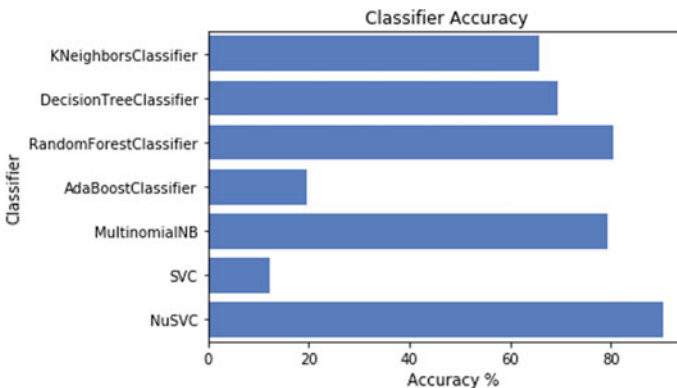


Fig. 2 Accuracy of classifiers

Table 2 Classifier performance metrics

Classifier	Accuracy	Log loss	Training time	Prediction time	Precision	Recall	F1 score	Support
KNeighborsClassifier	65.853659	7.386968	0.002	0.015	0.745893	0.658537	0.667038	82
DecisionTreeClassifier	67.073171	11.372524	0.04	0.002	0.704431	0.670732	0.678246	82
RandomForestClassifier	87.804878	0.831744	0.422	0.018	0.889289	0.878049	0.879898	82
AdaBoostClassifier	19.512195	2.216394	0.476	0.02	0.121467	0.195122	0.115827	82
MultinomialNB	79.268293	1.052502	0.003	0.001	0.833533	0.792683	0.76092	82
LogisticRegression	92.682927	0.91979	0.311	0.001	0.938153	0.926829	0.926508	82
SVC	12.195122	1.882378	0.875	0.022	0.014872	0.121951	0.026511	82
NuSVC	89.02439	0.365994	0.969	0.024	0.900851	0.890244	0.892395	82

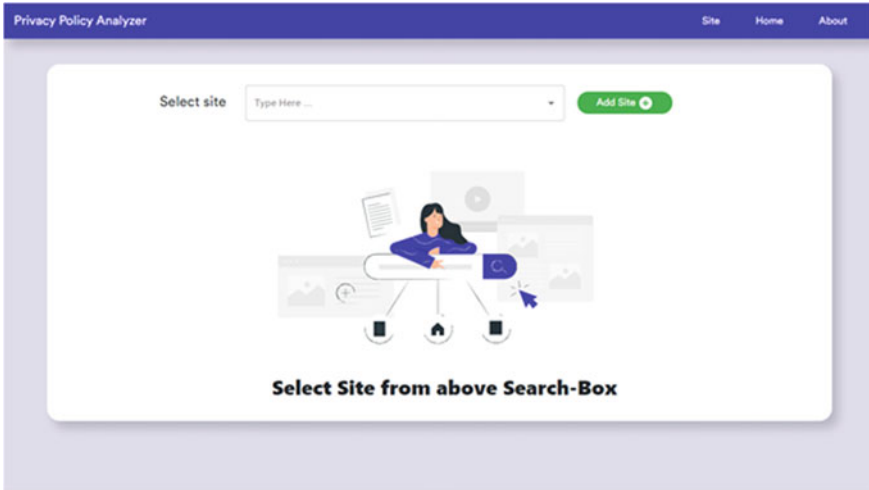


Fig. 3 Search privacy policy

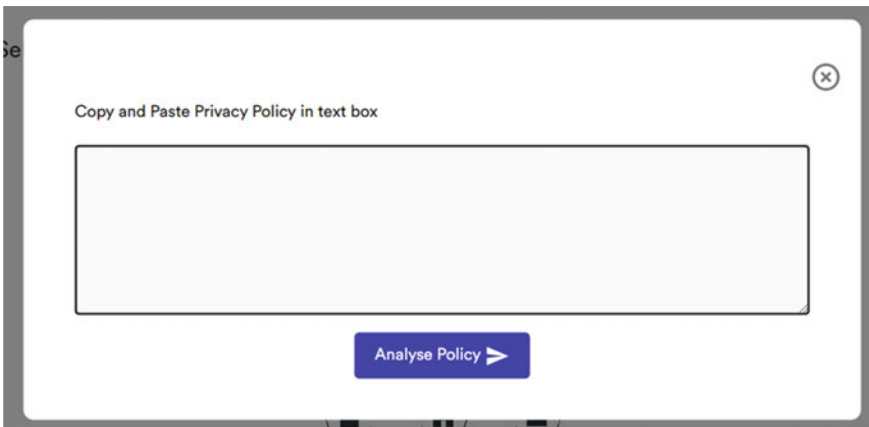


Fig. 4 Copy and paste a privacy policy

The privacy policy is classified into 8 categories. User can navigate through each policy and view related statements.

The user will be presented with a readability score of privacy policy giving them insights about the readability and difficulty of language used in the policy.

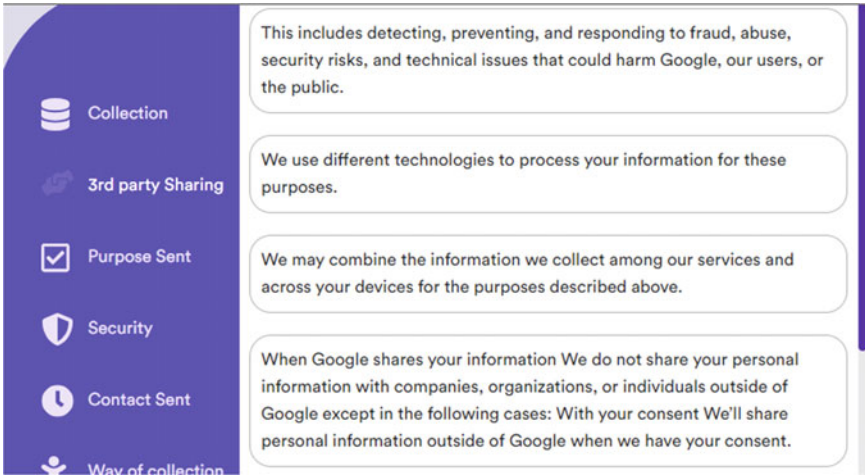


Fig. 5 Classification of privacy policy

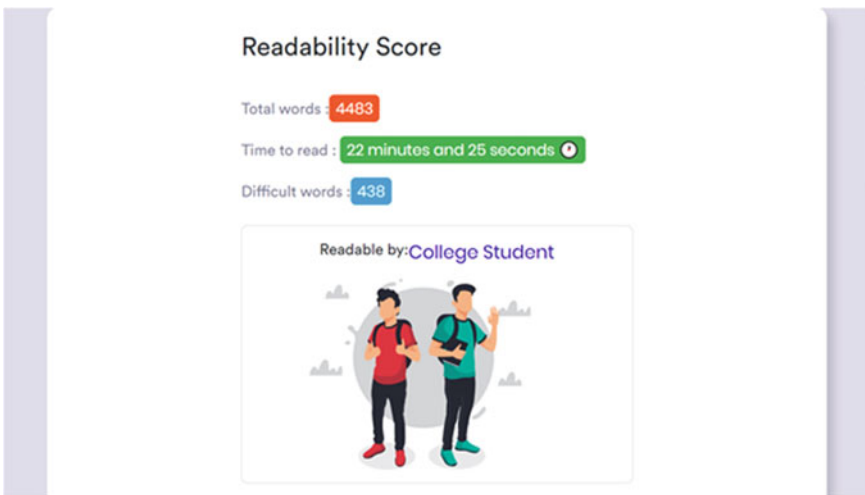


Fig. 6 Readability score

6 Conclusions and Future Scope

The privacy awareness tool is developed based on machine learning approaches to help users develop a belief, increase privacy awareness, builds good connections, and maintain transparency with the service providers. Well, structuredness is introduced in a legal document. This approach will also support the users to read and understand the privacy policy effortlessly and has improved the user’s responsiveness towards the

activities that are undertaken when a user accepts a privacy policy. Hence, this website has achieved its goal of increasing users' knowledge about terms and conditions specified in the privacy policy.

The limitations of the tool are that it will not be able to identify missing clauses other than 8 predefined clauses. It will only aid users to help understand privacy policy and it will not block access to any service infringing user privacy.

Further to extend this research multi-linguistic support can be added to this tool. This tool works on machine learning-based algorithms, whose performance can be increased for better classification.

References

1. P.S. Dhotre, A. Bihani, S. Khajuria, H. Olesen, Take it or leave it: Effective visualization of privacy policies, in *Cybersecurity and Privacy: Bridging the Gap (s. 39–64)*, eds. by I.S. Khajuria, L. Sørensen, K.E. Skouby. Wireless World Research Forum Series in Mobile Telecommunications (River Publishers, 2017)
2. J. Cox, Zoom iOS app sends data to facebook even if you don't have a facebook account, vice.com (2020) [Online]. Available: https://www.vice.com/en_us/article/k7e599/zoom-ios-app-sends-data-to-facebook-even-if-you-dont-have-a-facebook-account
3. M. Hachman, Microsoft's privacy policy admits contractors listen to Cortana, Skype recordings, pcworld.com (2019) [Online]. Available: <https://www.pcworld.com/article/3431701/microsofts-privacy-policy-admits-contractors-listen-to-cortana-skype-recordings.html>
4. R. Sobers, The average reading level of a privacy policy, varonis.com (2018) [Online]. Available: <https://www.varonis.com/blog/gdpr-privacy-policy/>. Accessed 22 Apr 2021
5. T. Linden, R. Khandelwal, H. Harkous, K. Fawaz, The privacy policy landscape after the GDPR [Online]. Available: <https://arxiv.org/pdf/1809.08396.pdf>
6. P.S. Dhotre, H. Olesen, S. Khajuria, Interpretation and analysis of privacy policies of websites in India, in *Proceedings of WWRF Meeting 36, Beijing, China, June 2016 Wireless World Research Forum (WWRF)* (2016). <https://vbn.aau.dk/da/publications/interpretation-and-analysis-of-privacy-policies-of-websites-in-in>
7. F. Schaub, R. Balebako, L.F. Cranor, Designing effective privacy notices and controls. *IEEE Internet Comput.* **21**(3), 70–77 (2017). <https://ieeexplore.ieee.org/document/7927931>
8. N. Guntamukkala, R. Dara, G. Grewal, A machine-learning based approach for measuring the completeness of online privacy policies, in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), Miami, FL, 2015*, pp. 289–294. <https://ieeexplore.ieee.org/document/7424323>
9. A.M. McDonald, L.F. Cranor, The cost of reading privacy policies. *ISJLP* **4**, 543 (2008). <https://lorrie.cranor.org/pubs/readingPolicyCost-authorDraft.pdf>
10. J. Gluck, F. Schaub, A. Friedman, H. Habib, N. Sadeh, L. Faith Cranor, Y. Agarwal, How short is too short? Carnegie Mellon University (2016). <https://www.usenix.org/conference/soups2016/technical-sessions/presentation/gluck>
11. P.S. Dhotre, H. Olesen, A survey of privacy awareness and current online practices of Indian users: motivating and mitigating factors for improving personal information privacy, in *Proceedings of WWRF Meeting 34, Santa Clara, CA, USA, Apr 2015*. <https://core.ac.uk/download/pdf/60615291.pdf>
12. Y. Wang, R.K. Nepali, Privacy impact assessment for online social networks, in *2015 International Conference on Collaboration Technologies and Systems (CTS), Atlanta, GA, 2015*, pp. 370–375. <https://ieeexplore.ieee.org/document/7210451>
13. R. Doshi, A. Ahale, G. Gharti, P. Pathrikar, P.S. Dhotre, Assessment of privacy policies using machine learning (2018). <https://www.irjet.net/archives/V5/I5/IRJET-V5I545.pdf>

14. P.S. Dhotre, H. Olesen, S. Khajuria, User privacy and empowerment: trends, challenges, and opportunities, in *Intelligent Computing and Information and Communication*, eds. by S. Bhalla, V. Bhateja, A. Chandavale, A. Hiwale, S. Satapathy. *Advances in Intelligent Systems and Computing*, vol 673 (Springer, Singapore, 2018). https://doi.org/10.1007/978-981-10-7245-1_30
15. S. Khajuria, L. Sørensen, K.E. Skouby, Cybersecurity and privacy—Bridging the gap. https://www.riverpublishers.com/book_details.php?book_id=434
16. P.N Mahalle, P.S Dhotre, Context-Aware Computing and Personalization, in *Context-Aware Pervasive Systems and Applications*. *Intelligent Systems Reference Library*, vol. 169 (Springer, Singapore, 2018).https://doi.org/10.1007/978-981-32-9952-8_4
17. T. Joachims, Text categorization with support vector machines: learning with many relevant features. https://www.cs.cornell.edu/people/tj/publications/joachims_98a.pdf

Target Node Protection from Rumours in Online Social Networks



Saranga Bora and Shilpa Rao

Abstract With the expansion of social networks, there is a rise in rumours and misinformation being shared among the users. Rumours have the potential to harm people in various ways. The spread of rumours in the network can be prevented by feeding users the true information. In this paper, we model a social network with multiple users connected to each other, where we aim to protect a specified set of target nodes in the network which are deemed as more vulnerable to the spread of a particular rumour as compared to the other nodes in the network. A linear threshold model with one direction state transition (LT1DT) for propagation of information in the social network, which is a modified version of the LT model, is considered in the paper. To counteract the spread of rumours, we select a group of nodes in the network as anti-rumour seed nodes using different selection algorithms to spread the truth in the network. We determine the average number of rumour infected target nodes in the network for random selection, random target node selection, max degree selection and greedy selection algorithm. The average number of rumour infected target nodes for different anti-rumour seed selection algorithms are compared in the paper. We compare the average number of rumour infected target nodes to the varying target set size for different anti-rumour seed selection algorithms. The results demonstrate the effectiveness and efficiency of the different algorithms on real world data set, and how target set size influence our results.

Keywords Online social networks · Rumour · Positive information

1 Introduction

Social networks are online platforms where people can connect with other people all over the world to build social relationship with people who have similar personal or career interests, or anything else in common. People share their thoughts, experiences, knowledge and other information over these social platforms. Some information

S. Bora · S. Rao (✉)
IIIT Guwahati, Guwahati, India
e-mail: shilpa@iiitg.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022
P. Verma et al. (eds.), *Advances in Data Computing, Communication and Security*,
Lecture Notes on Data Engineering and Communications Technologies 106,
https://doi.org/10.1007/978-981-16-8403-6_53

581

spreads rapidly over these platforms which has the potential to influence the users into taking certain actions and decisions. The information shared on social networks may not be always true.

1.1 Rumour in Social Networks

At times, rumour information gets shared on social networks. Rumours can be created with an intent to tarnish the reputation of a person, a brand or a public figure. Rumour could also be created unintentionally due to imperfect or incomplete knowledge. For, e.g. in Andhra Pradesh, a village was boycotted due to swine flu death rumours. Though the swine flu reports of the deceased were negative, it did not allay fears of the neighbouring villages [1].

Rumours can lead to conflict and in extreme cases communal violence among people living in a society, which eventually leads to great destruction, chaos or even loss of lives. For, e.g. on 26 June 2018, a 40-year-old woman was beaten by crowd in Ahmedabad as they suspected her to be a child-lifter [2].

With the overwhelming use of social media these days, rumours are getting more and more popular among the people. For, e.g. during the recent outbreak of COVID-19, there were some rumours spreading among people that consuming garlic and small onion can prevent one from getting infected by coronavirus disease, due to which there was a hike in the prices of garlic and onion resulting from their increasing demands [3].

Therefore, in today's modern world of Internet and social networks, it is very important to stop the spread of such rumours and negative information.

1.2 Rumour Diffusion Model

The social network is depicted by a graph $G = (V, E)$ where the set of vertices V represents the users of the social networks and the set of edges E represents the connections between the users. Let S_0 represent the vertices in the network, which are the sources of rumours. The set S_0 is called the rumour seed set. Let S_t denote the set of vertices influenced by rumour at time $t \geq 1$. We consider the following information diffusion model for spread of rumours and truth.

Linear Threshold model with One Direction state Transition (LT1DT): The LT1DT model [4] is a modified version of linear threshold model where it is assumed that a node (user) which was influenced earlier by rumour can later change his/her mind upon receiving the truth. Just like in LT model, here every edge $(u, v) \in E$ has an influence weight $w(u, v) \in [0, 1]$, indicating the influence of u on v . The weights are normalised such that for all v , the sum of weights of all incoming edges of v is at most 1, i.e. $\sum_{u \in N^{\text{in}}(v)} w(u, v) \leq 1$ for all $v \in V$, where $N^{\text{in}}(v)$ represents the set of in-neighbours of the node v . For all $(u, v) \notin E$, we assume $w(u, v) = 0$.

In the graph $G = (V, E)$, the LT1DT model takes the influence weights $w(\cdot)$ on all edges, and the seed set S_0 as the input, and generates the rumour active vertices S_t for all $t \geq 1$ by the following rule. Each node $v \in V$ in the graph G has a general threshold value θ in the range $[0,1]$ of adopting the true or false (rumour) information and a rumour threshold value θ_r in the range $[0,1]$ of adopting the rumour. Every node in the graph has a state; state 0 denotes the node is inactive (susceptible), state -1 denotes the node is rumour active (infected), and state 1 denotes the node is anti-rumour active.

In each time-step $t > 0$, all the inactive and rumour active nodes check for net incoming positive and negative influences from all its neighbouring nodes, and checks if the sum of the magnitudes of all the positive and negative influences are greater than the general threshold value θ , and correspondingly activates the node if the threshold is exceeded. If the ratio of the magnitude of net incoming negative edge weight to the magnitude of net incoming edge weight is greater than the rumour threshold value θ_r , then the node is rumour activated, else it gets anti-rumour activated. In each time-step, the total number of nodes activated in that time-step is counted. If no node is activated in a particular time-step, it implies that the rumour/anti-rumour diffusion process has terminated.

Assumptions:

Assumption 1: If a user (node) was influenced by rumour earlier, he/she can later change his/her mind upon receiving the truth.

Assumption 2: Once a user is given the truth and has changed his/her state, he/she will remain in that state and will not change his/ her state again in future.

1.3 Rumour Containment in Social Networks

The effect of rumour active nodes on the other nodes in the network can be reduced by taking certain measures. The number of rumour active nodes in the final set can be minimized in two ways:

1. *Spreading the truth (anti-rumour campaign):* In this method, anti-rumour campaigns are held in the network to control the spread of misinformation. Some nodes are selected as anti-rumour agents in the network to spread the true information. Different algorithms are used to efficiently select the initial set of anti-rumour agents/truth starters such that the number of rumour active nodes in the final set is minimum.
2. *Blocking the nodes:* In this method, the nodes infected by rumour are blocked from interacting with the rest of the nodes in the network. Some nodes/edges are selected to be immunized before the start of an epidemic such that misinformation cannot pass through to the immunized nodes/edges. The nodes and edges are identified whose removal would minimize the largest eigenvalue of the remaining network.

2 Related Work

Rumour containment currently is an active research area and based on different strategies. The works given in [4–8] incorporate an immunization strategy where some nodes/edges are selected to immunize such that misinformation cannot pass through to those nodes/edges; while the works mentioned in [9–11] adopt anti-rumour campaigns in the networks to reduce the spread of misinformation.

The work done in [7] involves selection of nodes/edges to be immunized before an epidemic starts. The nodes and edges whose removal would save the largest set of nodes of the remaining network are identified. In [5, 6, 10], nodes are selected to immunize the given infected nodes. A tree structure is constructed rooted at the infected nodes and further nodes are selected which can save the maximum number of descendants in the tree. In [4], a user's experience is considered while being blocked and a time window to model the period is used where users are willing to be blocked from using social service.

It has been found that in broadcasting, spreading the true information is much more effective than immunizing nodes/edges. In [11], the problem of 'eventual influence limitation' or 'influence blocking maximization' is defined in order to identify the set of nodes to spread the truth information. A multi-campaign independent cascade model is designed, where both rumour and anti-rumour campaign are actively propagating in the network. Several methods are proposed based on centrality measures to pick the nodes for anti-rumour campaigns.

In [12], a linear threshold model is proposed where both rumour and anti-rumour campaigns can compete with each other. In [4], a linear threshold model with one direction state transition is proposed which is a modification of the linear threshold model.

In [13–17], the importance of considering time delays in propagating information is recognized. In [13] and [15], it is observed that a user receives information only when he/she is online. The 'independent cascade model with login events' (IC-L) is introduced in the delayed diffusion process. In [16], a decay function is proposed, where nodes further away from the seeds set will have lower chance of being influenced. In [17], a distribution of meeting probability is assigned to each edge, indicating how likely it is for the pair of nodes to meet with each other over a certain time frame. All of these work involves maximizing the influence over a certain period of time in the given timed diffusion models.

3 Problem Statement

In our work, we aim to protect a certain specified set of nodes in the network which are deemed as more vulnerable to the spread of a particular rumour, as compared to the other nodes in the network. Such nodes are termed as the 'target nodes' [18]. Here, we will be using the first method, i.e. anti-rumour campaign, to minimize the

average number of rumour active target nodes in the network. We select anti-rumour agents to spread the truth using different algorithms and analyse how the number of rumour active target nodes in the final set changes. We have considered the LT1DT model in which once a node has been influenced by anti-rumour, it stays in that particular state and does not get infected by the rumour again.

We randomly select the target nodes for the target set ' T ' out of all the nodes $v \in V$ in the graph G . We randomly select a different set of target nodes of size $|T|$ for every sample path of information diffusion to study the average number of target nodes influenced.

Initially, at time $t = 0$, k number of nodes are selected as the rumour agents to spread the rumour and k' number of nodes are selected as the anti-rumour agents to spread the truth. Different selection algorithms are used to select the seed nodes. Rumour and truth is spread in the network until there is no further possible activation of nodes in the graph.

4 Proposed Methodology

In this section, we discuss how we can minimise the spread of rumour information among the target nodes in a network. We emulate [4] to minimise the total number of rumour active target nodes in the final set by spreading the truth using anti-rumour agents. We can achieve this by wisely choosing the initial anti-rumour agents at the beginning. We use the following four algorithms for anti-rumour seed node selection: (i) Random selection algorithm [19] (ii) Random target node selection algorithm (iii) Max degree selection algorithm [19] (iv) Greedy selection algorithm [19].

The greedy selection algorithm has obtained results close to the optimal algorithm [18–20] and hence is used as a baseline algorithm for performance benchmarking.

4.1 Algorithms for Spread of Positive Information

- (i) **Random selection algorithm:** In this algorithm, the ' k ' number of nodes are selected as anti-rumour seed nodes randomly out of the total ' $|V|$ ' number of nodes in the graph. The computational complexity of selecting the anti-rumour seed nodes using this algorithm is in the order of $O(k)$, where ' k ' is the budget.
- (ii) **Random target node selection algorithm:** In this algorithm, the ' k ' number of nodes are selected as anti-rumour seed nodes randomly out of the target set nodes ' T ' and total ' $|V|$ ' number of nodes in the graph. If the budget (k) is less than the ' $|T|$ ' number of nodes in the target set, all the ' k ' number of anti-rumour seed nodes are selected randomly from among the target set nodes; if the budget (k) is greater than the ' $|T|$ ' number of target set nodes, all the ' $|T|$ ' number of target set nodes are selected as anti-rumour seed nodes

and the remaining number of ' $k - |T|$ ' anti-rumour seed nodes are selected randomly from among the other 'non-target' inactive nodes. The computational complexity of selecting the anti-rumour seed nodes using this algorithm is in the order of $O(k)$.

- (iii) **Max degree selection algorithm:** In this algorithm, the ' k ' number of nodes with maximum number of outgoing edges (i.e. maximum degree) are selected out of the total ' $|V|$ ' number of nodes in the graph as anti-rumour seed nodes. The computational complexity of selecting the anti-rumour seed nodes using this algorithm is in the order of $O(|V|^2)$, where $|V|$ is the net number of nodes in the graph.
- (iv) **Greedy selection algorithm:** In this algorithm, in each round i , one node u_i is added into the set S_A of anti-rumour seed nodes, such that this node u_i , along with the other seed nodes previously added to the set S_A , leads to the minimum number of rumour infected target nodes in the final set V . This method of adding nodes to the set S_A keeps recurring until all the ' k ' number of nodes are selected as anti-rumour seed nodes. The computational complexity of selecting the anti-rumour seed nodes using this algorithm is in the order of $O(Rk|V||E|)$, where ' k ' is the budget, ' R ' is the number of sample paths, ' $|V|$ ' is the total number of nodes and ' $|E|$ ' is the total number of edges in the graph.

We assume that the rumour seed set is selected by the rumour agent based on max degree selection scheme. The number of anti-rumour seed nodes ' k ' is also called the budget for rumour containment.

5 Experiments and Results

We describe the simulation set-up and simulation results for rumour containment using random selection, random target node selection, max degree selection and greedy selection algorithm for LT1DT model in the following. We have built the code using C programme. We have performed the simulations on a real world social network data set: soc-firm-hi-tech (social networks) [21], and plotted the results. The simulations were run for 4000 iterations, each time with a different set of general threshold and rumour threshold values for the nodes, and with a different target set consisting of randomly chosen target nodes for each iteration, to get statistically meaningful results.

Figure 1 shows the average number of rumour active target nodes in the final set ' V ' as we increase our target set size ' $|T|$ ' from 1 to 8, keeping the number of rumour seed nodes fixed at 8 and our budget ' k ' of anti-rumour seed nodes fixed at 2 for the soc-firm-hi-tech (social networks) data set. We observe that as we increase our target set size, the number of rumour infected target nodes in the final set increases. This is because more the number of target nodes in the network, higher is the probability of an activated node being a target node; also the probability of the activated target node being a rumour infected node is higher because we have fixed the number of

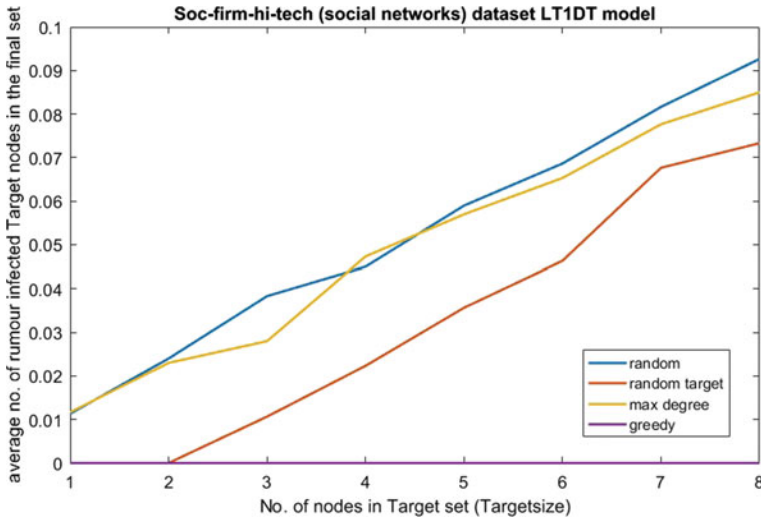


Fig. 1 Average number of rumour active target nodes versus target set size for soc-firm-hi-tech (social networks) data set

rumour seed nodes to be multiple times higher than the number of anti-rumour seed nodes in the network, thus increasing the probability of a target node being infected by rumour.

We can see that max degree algorithm is performing better than the random selection algorithm; because random selection algorithm selects the initial agents (nodes) in a random manner to spread the true information, whereas max degree algorithm has its unique strategy of selecting the initial agents (nodes) with maximum outgoing degree to spread the information. Random target selection algorithm is performing better than the max degree algorithm because random target selection algorithm selects the anti-rumour seed nodes from the target set of nodes itself. Greedy selection algorithm is performing the best because here we simulate the entire process itself beforehand in selection of every next seed node by checking for all inactive nodes, taking one at a time and selecting the one which performs the best along with the previously selected seed nodes.

Figure 2 shows average number of rumour active target nodes in the final set ‘V’ as we increase our budget ‘k’ of anti-rumour seed nodes from 1 to 5, keeping the number of rumour seed nodes fixed at the soc-firm-hi-tech (social networks) data set. For this parameter setting, we observe that the number of rumour active target nodes for greedy algorithm is nearly zero.

We can also see from the above plot that as we increase our budget, the number of rumour infected target nodes in the final set decreases. This is because more the number of initial anti-rumour agents, faster will be the spread of true information in the network, thus decreasing the probability of a target node being infected by rumour.

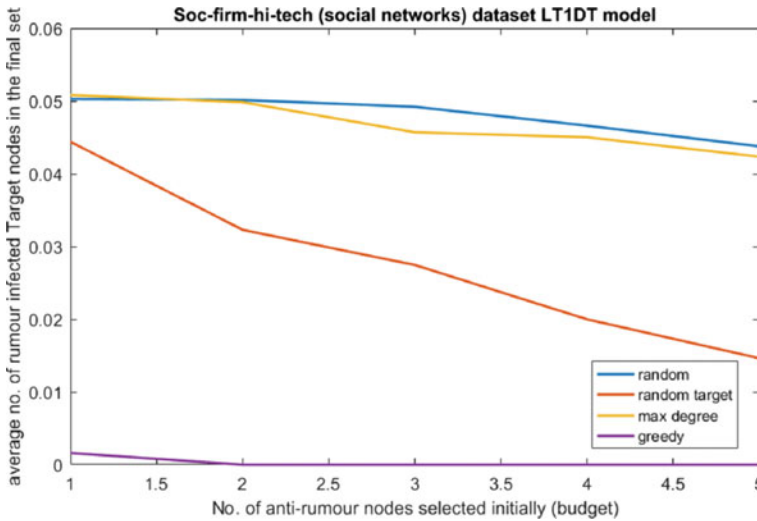


Fig. 2 Average number of rumour active target nodes versus budget for soc-firm-hi-tech (social networks) data set

6 Conclusion and Future Work

Here, we have addressed one of the most pressing issues of the twenty-first century, as the Internet is becoming more and more popular day by day, we can see a rise in the use of social network services among people. Therefore the social platforms have a high potential of influencing people in today's era. Thus the spread of some rumours may cause great harm to an individual and also to the society.

We have seen how the algorithm we use for selection of the initial anti-rumour agents impacts the spread of rumour in the network. Also, we have seen that we get better results with increasing budgets. Therefore more the number of initial agents we can choose to spread the truth, lesser is the number of rumour infected nodes in the final set. Future work would be implementation of a simpler algorithm with lower complexity that performs nearly as good as the greedy algorithm and extensive performance comparison with several low complexity heuristic-based algorithms. The scalability of the seed selection algorithms also forms an interesting avenue of future work.

References

1. Ndtv Homepage, <https://www.ndtv.com/andhra-pradesh-news/village-in-andhra-pradesh-fac-social-boycott-after-rumour-spread-that-2-died-of-swine-flu-1960169>. Last accessed 17 Apr 2020

2. Outlook Homepage, <https://www.outlookindia.com/newscroll/woman-killed-on-suspicion-ofbeing-childlifter-in-ahmedabad/1338304>. Last accessed 17 Apr 2020
3. The Times of India Homepage, <https://timesofindia.indiatimes.com/city/coimbatore/garlic-sha-lot-prices-soaras-rumours-spread-they-could-prevent-covid-19/articleshow/74720219.cms>. Last accessed 17 Apr 2020
4. B. Wang, G. Chen, L. Fu, L. Song, X. Wang, X. Liu, Drimux: dynamic rumour influence minimization with user experience in social networks, in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)* (AAAI Press, 2016), pp. 791–797
5. C. Song, W. Hsu, M.L. Lee, Node immunization over infectious period, in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM'15* (ACM, New York (2015), pp. 831–840.
6. Y. Zhang, B.A. Prakash, Data-aware vaccine allocation over large networks ACM Trans. Knowl. Discov. Data **10**(2), 20–12032 (2015)
7. H. Tong, B.A. Prakash, T. Eliassi-Rad, M. Faloutsos, C. Faloutsos, Gelling, and melting, large graphs by edge manipulation, in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM'12* (ACM, New York, 2012), pp. 245–254
8. Y. Zhang, B.A. Prakash, Scalable vaccine distribution in large graphs given uncertain data, in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. CIKM'14* (ACM, New York, NY, USA, 2014), pp. 1719–1728
9. L. Yang, Z. Li, A. Giua, Rumor containment by spreading correct information in social networks, in *2019 American Control Conference (ACC)* (IEEE, 2019), pp. 5608–5613
10. J. Tsai, T.H. Nguyen, M. Tambe, Security games for controlling contagion, in *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI-12)* (AAAI Press, 2012), pp. 1464–1470
11. C. Budak, D. Agrawal, A. El Abbadi, Limiting the spread of misinformation in social networks, in *Proceedings of the 20th International Conference on World Wide Web* (2011), pp. 665–674
12. A. Borodin, Y. Filmus, J. Oren : Threshold models for competitive influence in social networks. In: 6th International Workshop on Internet and Network Economics, 539–550.. Springer, Berlin (2010).
13. W. Chen, W. Lu, N. Zhang, Time-critical influence maximization in social networks with time-delayed diffusion process, in *Proceedings of the 26th AAAI Conference on Artificial Intelligence(AAAI-12)* (AAAI Press, 2012), pp. 592–598
14. E. Cohen, D. Delling, T. Pajor, R.F. Werneck, Timed influence: Computation and maximization. arXiv preprint [arXiv:1410.6976](https://arxiv.org/abs/1410.6976) (2014)
15. C. Song, W. Hsu, M.L. Lee, Targeted influence maximization in social networks, in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. CIKM'16* (2016), pp. 1683–1692
16. X. He, M. Gao, M.-Y. Kan, Y. Liu, K. Sugiyama, Predicting the popularity of web 2.0 items based on user comments, in *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)* (ACM, New York, NY, USA, 2014), pp. 233–242
17. B. Liu, G. Cong, D. Xu, Y. Zeng, Time constrained influence maximization in social networks, in *Proceedings of the IEEE 12th International Conference on Data Mining (ICDM'12)* (IEEE, 2012), pp. 439–448
18. J. Guo, Y. Li, W. Wu, Member: Targeted protection maximization in social networks. IEEE Trans. Netw. Sci. Eng. **7**(3), 1645–1655 (2019)
19. X. He, G. Song, W. Chen, Q. Jiang, Influence blocking maximization in social networks under the competitive linear threshold model, in *Proceedings of the 2012 SIAM International Conference on Data Mining* (SIAM, 2012), pp. 463–474
20. D. Kempe, J.M. Kleinberg, É. Tardos, Maximizing the spread of influence through a social network, in *9th Proceedings on ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, NY, USA, 2003), pp. 137–146
21. R.A. Rossi, N.K. Ahmed, The network data repository with interactive graph analytics and visualization, in *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI-15)* (2015), pp. 4292–93. <http://networkrepository.com>

Risk Detection of Android Applications Using Static Permissions



Meghna Dhalaria and Ekta Gandotra

Abstract Malware developers have targeted Android as it is the most widely used operating system for mobile devices and smartphones. Android protection is primarily based on user decisions to install applications by authorising their requested permissions. In this paper, we propose a system which detects the risk of an app based on static permissions. The proposed approach is tested on 3547 applications out of which 1747 are malware apps and 1800 are benign apps. An artificial neural network model is developed in order to predict the risk of an Android app. The model calculates the probability of malware and benign classes of data samples. The detection accuracy obtained by the proposed model is 96.7%. Based on probability of malware, the risk of an app is categorized into four factors, i.e. no, low, medium and high risk. A graphical user interface is also developed where the Android app can be uploaded to find its risk using the proposed model.

Keywords Android malware · Static malware analysis · Permissions · Risk detection

1 Introduction

Since Android has become the most common operating system (OS) for tablets and smartphones, its users have become the most vulnerable to security threats. One of the most significant challenges in mobile OS is security. Malicious applications (apps) attempt to attack Android device in order to steal confidential data, make fake calls and send SMS, etc. According to the MacAfee study, the growth rate of total and new malware is 121 million and 49 million, respectively, in 2020 [1]. The increase in growth rate of Android malware becomes dangerous to Android users. Due to lack of information and understanding, users do not determine whether an app is malicious or not. When downloading an app from the Android app store,

M. Dhalaria (✉) · E. Gandotra

Department of Computer Science and Engineering, Jaypee University of Information Technology, Wakanaghat, Solan, Himachal Pradesh, India

most Android users neglect or do not read the terms and conditions. Unfortunately, attackers take advantage of this fact and target mobile devices.

Due to the increase in Android malware, it has become difficult to manually handle the malicious samples. To overcome this limitation, it is necessary to build an efficient technique for better identifying risk of applications. Earlier signature-based approach was used for the identification of malware. This approach is based on matching the signature of the app in the database. The limitation of this method is that it is unable to detect unknown malware [2]. On a regular basis, malware developers create new malware to threaten the security of the system and the privacy of its users. The risk posed by malware necessitates for the development of effective methods. This evaluation aids in the provision of early warnings regarding a specific Android app, allowing immediate attention to be paid to it in terms of allocating resources. For better identifying the risk in apps, the concept of machine learning and deep learning is being used along with static and dynamic malware analysis. The static method checks the functionalities and maliciousness of an app by disassemble and examining its code without running the apps. This method uses less resources and quicker in analysing the code but fails against obfuscation techniques [3, 4]. To overcome this limitation, dynamic malware analysis approach is used. It has the capability to detect unknown malware and effective in tracking the characteristic of apps. Dynamic malware analysis examines the characteristic of the app while running the code in the virtual environment known as Sandbox. The limitation of this method is that some part of the code remains undetected at execution time [5–9]. This paper proposes a method to predict the risk of an Android app. An artificial neural network (ANN) model is developed which calculates the probability of malware and benign classes of data samples. Based on probability of malware, the risk of an app is categorized into four factors, i.e. no, low, medium and high risk. A graphical user interface is also designed where the Android app can be uploaded to find its risk using the proposed model.

The rest of the paper is structured as follows: The literature review is presented in Sect. 2. The proposed technique is described in Sect. 3. Section 4 explains the experimental results of the proposed approach. At last, Sect. 5 concludes the paper.

2 Literature Review

This section discusses the various approaches/techniques for risk analysis and detection of Android apps that have been reported in the literature.

Grace et al. [10] proposed an automated system named as RiskRanker to investigate whether an app is risky. The experiments were carried out on total 1,18,318 apps gathered from multiple Android markets. The findings show that RiskRanker was effective and scalable in policing Android markets of all kinds. Idress et al. [11] introduced PIndroid, a malware detection framework that uses ensemble learning methods. This research focuses on a system for detecting malware that incorporates the intents and permissions set by ensemble methods. The method is tested

on 445 clean and 1300 contaminated Android apps gathered from third-party and official sources. They came to the conclusion that the proposed framework could be used to classify Android apps. Sharma et al. [12] classified malware capabilities based on features discovered through static and dynamic malware review, as well as the malware naming convention used by antivirus (AV) vendors. The authors proposed a strategy to solve the problem of contradictions by evaluating the various capabilities of malware using fuzzy logic. According to the proposed FIS, 83% of malware samples were found to belong to the same group. For malware identification, Mariconti et al. [13] implemented MAMADROID, a static malware analysis-based framework. For malware detection, static features such as API calls and call graph are used. The findings were tested on 3.5 million malicious apps and 8.5 million benign apps. The proposed method yielded the F -measure of 0.99. Jang et al. [14] demonstrated Andro-Autopsy, an antimalware mechanism that protects mobile devices. The findings suggested that the proposed system is capable of detecting and classifying malware. Sharma et al. [15] proposed the RNPDrroid method for risk analysis based on permissions. On the M0Drroid dataset, which contains 400 Android samples with 165 attributes, the proposed approach is evaluated. For statistical analysis, they used the T -test and ANOVA. The results showed that at 5% level of significance, the calculated value of F is 517.3, which is higher than the tabulated value of F is 2.61. Gandotra et al. [16] proposed an automated method based on fuzzy logic for calculating the harm potential of malware programmes based on features obtained through automated analysis. Shrivastava et al. [17] suggested the SensDrroid system, which used sensitive analysis to evaluate the efficiency of Android permissions and intents. A sufficient number of apps from third-party and official app stores were included in the proposed system. According to the results, the proposed system was successful in classifying infected and clean apps with a higher detection rate.

This paper presents a method to predict the risk of an Android app. We create an ANN model that calculates the probability of malware and benign classes of samples. The risk of an app is classified into four categories based on the probability of malware, i.e. no, low, medium and high risk. A graphical user interface is also designed where the Android app can be uploaded to determine its risk using the proposed model.

3 Proposed Methodology

This section describes the architecture of RiskDetector for detecting the risk factor of apps while they are being installed on the device. First of all, we collect Android samples from different sources, including apkpure [18], apkmirror [19] and virusshare [20]. Based on their scanning performance, all applications are categorised as benign or malicious. Afterwards, we performed a static malware analysis to mine the static permissions. A system that detects an app's hazard at the time of installation is required for the user's mobile device's safety. As a result, app permissions are scrutinised in order to determine the risk factor. Figure 1 shows the

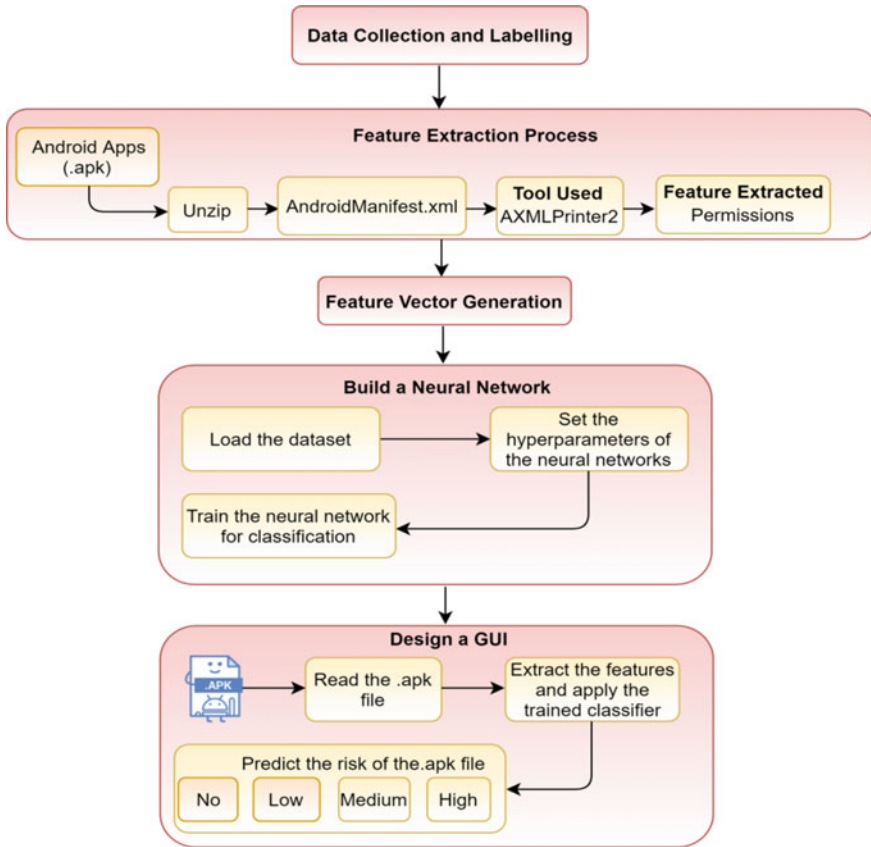


Fig. 1 Workflow of methodology used

architecture of RiskDetector.

Data Collection and Labelling

Data collection and labelling is the first steps in the proposed system. We collected 4400 Android apps from various sources, including virusshare, apkpure and apkmirror. Their hash value is computed to remove the duplicates. Subsequently, these are scanned using avira antivirus (AV) [21] to label as benign and malware. After removing duplicates and scanning, we are left with 3547 apps containing 1800 benign and 1747 malicious.

Feature Extraction Process

The second step is process of extracting features. In this, we mined static features using static malware analysis. Permissions are examined to identify the risk factor. To extract the features of the Android application package (.apk) file, we first unzip it to obtain AndroidManifest.xml. A total of 277 permissions are extracted from AndroidManifest.xml.

Build a Model Using ANN

ANN is also known as multilayer feed forward neural network which comprises of an input layer, hidden layer and the output layer. Every layer is made up of units and when inputs are given to these units it makes the input layer. The inputs from the input layer with some weights are then fed to the next layer, i.e. hidden layer. The output of the first hidden layer becomes input for the next hidden layer and so on. The output of the last hidden layer (i.e. weighted output) becomes input for the output layer which produces network prediction. Each unit imparts inputs to each unit in the next layer that's why it is known to be fully connected.

The inputs are passed through the input units, i.e. $\{x_1, x_2, \dots, x_n\}$. Then these inputs are multiplied by the respective weights to get the weighted sum, which is further added to bias θ_k associated with unit k . A non-linear activation function such as sigmoid, softmax is used to get net input. Its output O_k is equal to input I_k . The net input and output I_k and O_k , respectively, of unit k in the hidden and output layer are calculated as shown in Eq. (1)

$$I_k = \sum_j w_{jk} \times O_j + \theta_k \quad (1)$$

where w_{jk} represents the weight of the link from unit j in the previous layer to unit k , O_j denotes the output of unit j from the previous layer and θ_k represents the bias. The output of unit k is calculated as given in Eq. (2)

$$O_k = \frac{1}{1 + e^{-I_k}} \quad (2)$$

The sigmoid or logistic function is used. It is also known as squashing function because it transforms the large input values to a smaller range, i.e. from 0 to 1. To minimize mean square error (MSE) between network prediction value and actual target the weights are modified for each training tuple. Here, the changes are done in backward direction, i.e. from the output layer through every hidden layer to the first hidden layer. The error is propagating in a backward direction by updating biases and weights. For a unit k in the output layer, the error (err_k) is calculated as shown in Eq. (3)

$$err_k = O_k(1 - O_k)(T_k - O_k) \quad (3)$$

Here O_k represents the output of unit k and T_k denotes the target value. The error of hidden layer of unit k is computed as shown in Eq. (4)

$$err_k = O_k(1 - O_k) \sum_i err_i w_{ki} \quad (4)$$

Here w_{ki} represents weights of the link from unit k to a unit j . err_i denotes the error of unit i . The updated weights and bias are calculated using Eq. (5, 6) and (7, 8), respectively.

$$\Delta w_{jk} = l \times err_k O_j \quad (5)$$

$$w_{jk} = w_{jk} + \Delta w_{jk} \quad (6)$$

Here l represents the learning rate which lies between 0 and 1. Δw_{jk} denotes the change in weight.

$$\Delta \theta_k = l \times err_k \quad (7)$$

$$\theta_k = \theta_k + \Delta \theta_k \quad (8)$$

In this work, the classifier is trained using train-test split technique. It splits the dataset into the ratio of 80–20% in which first 80% of data is used for training purpose and rest 20% of data is used testing purpose. The efficiency of the model is examined using sensitivity or recall, precision, f -measure and accuracy. The results of test data in terms of recall, precision, F -measure and accuracy are shown in Eqs. (9–12):

$$\text{Sensitivity or Recall} = \frac{TP}{TP + FN} = 0.96 \quad (9)$$

$$\text{Precision} = \frac{TP}{TP + FP} = 0.96 \quad (10)$$

$$F - \text{measure} = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} = 0.96 \quad (11)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} * 100 = 96.7\% \quad (12)$$

Design Graphical User Interface (GUI)

After building a model, a GUI is designed to predict the risk of an app. It calculates the probability of malware and benign classes of test data using ANN model. Based on the probability of malware ($P(M)$) class of the test data, following criteria are defined to identify the category of the risk as shown in Algorithm 1.

Figure 2 displays the range for predicting risk of an app. It shows that if the probability of malware of the test sample range is in from 0 to <0.25 then the prediction output of the test sample falls under no risk. If the probability of malware of the test sample range is in between ≥ 0.25 and <0.5 then the prediction output of the test sample falls under low risk. If the probability of malware of the test sample range

Algorithm 1 Algorithm of predicting risk in Android app

Input: Probability of malware class ($P(M)$) of the test data
Output: Predicting risk of an app into four categories (i.e. high, medium, low and no risk).

```
01: if ( $P(M) = 0$  &&  $P(M) < 0.25$ )  
02: Output = No risk  
03: else if ( $P(M) \geq 0.25$  &&  $P(M) < 0.5$ )  
04: Output = Low risk  
05: else if ( $P(M) \geq 0.5$  &&  $P(M) < 0.75$ )  
06: Output = Medium risk  
07: else if ( $P(M) \geq 0.75$  &&  $P(M) = 1$ )  
08: Output = High risk  
09: end if
```

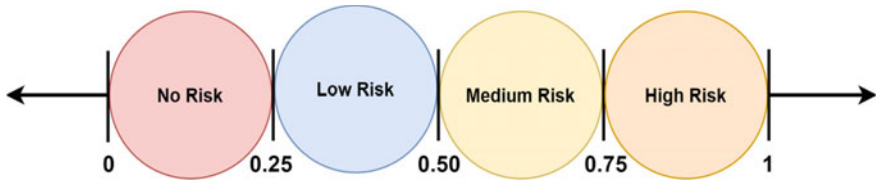


Fig. 2 Range for predicting risk of an app

is in between ≥ 0.5 to < 0.75 then the prediction output of the test sample falls under medium risk. If the probability of malware of the test sample range is in between ≥ 0.75 to 1 then the prediction output of the test sample falls under high risk. Based on this range, the Android apps are categorized into four risk factors, i.e. no, low, medium and high risk.

4 Experimental Results

This section describes the risk detection system based on static permissions. The web app is developed using Django which is a collection of python libs that allow to create a quality web application. Figure 3 shows the login page of the web application. The user must login with Google account to access the web app. The source of background image used in the web page (URL: <https://images.app.goo.gl/uGmCARga8NNyVXep6>).

After the successful login, the user must browse the .apk file to detect its risk factor (i.e. high, medium, low and no risk) of an Android app as shown in Fig. 4. It first extracts the static features (i.e. permissions) of the test sample. Then, it calculates the probability of malware and benign classes of test data using ANN model. Based on the probability of malware class, the risk of the app is identified as discussed in design graphical user interface subsection. Figure 5 shows the page displaying the message regarding the risk of test sample.



Fig. 3 Login page of web application

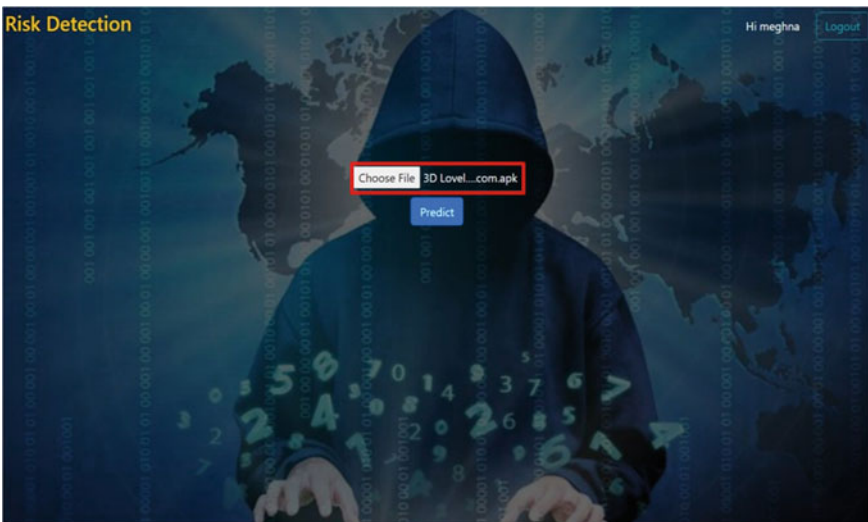


Fig. 4 Web page for browsing.apk file

5 Conclusion

The use of smartphones has increased dramatically in recent years, which has resulted in an increase in malicious applications. In mobile devices, the most difficult task is recognising and avoiding risk quickly. Before an app can be installed, it needs permission from the user's system. In this paper, we proposed a RiskDetector system

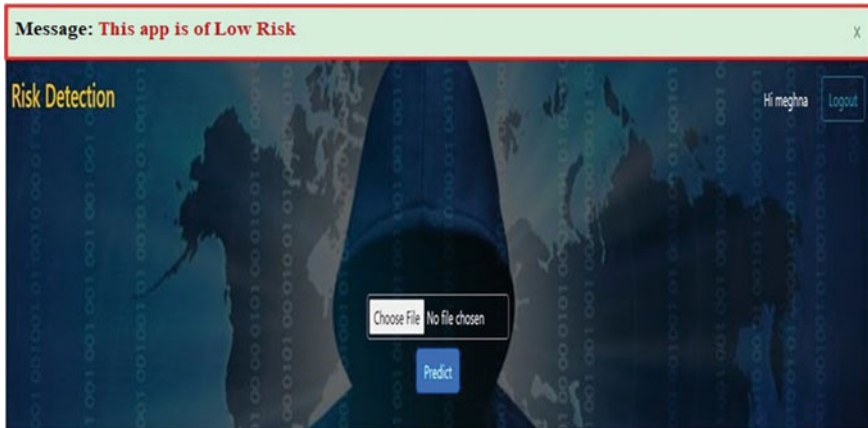


Fig. 5 Displaying prediction message of a test sample

for detecting the risk of an app based on static permissions. An ANN model is created which calculates the probability of malware and benign classes of data samples. This model was tested on 3547 Android apps containing 1747 malicious and 1800 benign. Based on the obtained probability of malware from the model, the risk of an app is categorized into four factors, i.e. no, low, medium and high risk. A graphical user interface is also designed where the Android app can be uploaded to find its risk using the proposed model. In future, fuzzy logic will be explored to provide more flexibility for analysing the risk of an app into different categories.

References

1. McAfee Labs, Threat Predictions Report, McAfee Labs, Santa Clara, CA, USA. (2020)
2. M. Dhalaria, E. Gandotra, S. Saha, Comparative analysis of ensemble methods for classification of android malicious applications, in *International Conference on Advances in Computing and Data Sciences* (Springer, 2019), pp. 370–380
3. M. Dhalaria, E. Gandotra, A framework for detection of android malware using static features, in *2020 IEEE 17th India Council International Conference INDICON* (IEEE, 2020), pp. 1–7
4. M. Dhalaria, E. Gandotra, Android malware detection using chi-square feature selection and ensemble learning method, in *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)* (IEEE, 2020), pp. 36–41
5. K. Tam, A. Feizollah, N.B. Anuar, R. Salleh, L. Cavallaro, The evolution of android malware and android analysis techniques. *ACM Comput. Surv. (CSUR)* **49**(4), 1–41 (2017)
6. M. Dhalaria, E. Gandotra, Android malware detection techniques: a literature review. *Recent Patents Eng.* **15**(2), 225–245 (2021)
7. E. Gandotra, D. Bansal, S. Sofat, Malware analysis and classification: a survey. *J. Inf. Secur.* (2014)
8. M. Dhalaria, E. Gandotra, CSForest: an approach for imbalanced family classification of android malicious applications. *Int. J. Inf. Technol.* **13**(3), 1059–1071 (2021)
9. M. Dhalaria, E. Gandotra, A hybrid approach for android malware detection and family classification. *Int. J. Interactive Multimedia Artif. Intell.* **6**(6) (2021)

10. M. Grace, Y. Zhou, Q. Zhang, S. Zou, X. Jiang, Riskranker: scalable and accurate zero-day android malware detection, in *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, pp. 281–294 (2012)
11. F. Idrees, M. Rajarajan, M. Conti, T.M. Chen, Y. Rahulamathavan, PIndroid: a novel android malware detection system using ensemble learning methods. *Comput. Secur.* **68**, 36–46 (2017)
12. A. Sharma, E. Gandotra, D. Bansal, D. Gupta, Malware capability assessment using fuzzy logic. *Cybern. Syst.* **50**(4), 323–338 (2019)
13. E. Mariconti, L. Onwuzurike, P. Andriotis, E. De Cristofaro, G. Ross, G. Stringhini, Mamadroid: Detecting android malware by building markov chains of behavioral models. arXiv preprint [arXiv:1612.04433](https://arxiv.org/abs/1612.04433) (2016)
14. J.W. Jang, H. Kang, J. Woo, A. Mohaisen, H.K. Kim, Andro-AutoPsy: anti-malware system based on similarity matching of malware and malware creator-centric information. *Digit. Investig.* **14**, 17–35 (2015)
15. K. Sharma, B.B. Gupta, Mitigation and risk factor analysis of android applications. *Comput. Electr. Eng.* **71**, 416–430 (2018)
16. E. Gandotra, D. Bansal, S. Sofat, Malware threat assessment using fuzzy logic paradigm. *Cybern. Syst.* **48**(1), 29–48 (2017)
17. G. Shrivastava, P. Kumar, SensDroid: analysis for malicious activity risk of Android application. *Multimedia Tools Appl.* **78**(24), 35713–35731 (2019)
18. Apkpure. <https://apkpure.com/>. Accessed Mar 2019
19. APKMirror. <https://www.apkmirror.com/>. Accessed Mar 2019
20. Virusshare. <https://virusshare.com/>. Accessed Mar 2019
21. Avira. <https://www.avira.com/>. Accessed Apr 2019

Cryptography-Based Efficient Secured Routing Algorithm for Vehicular Ad Hoc Networks



Deepak Dembla, Parul Tyagi, Yogesh Chaba, Mridul Chaba,
and Sarvjeet Kaur Chatrath

Abstract There exist certain challenges like high overhead, poor performance, and detection of malicious nodes in the vehicular ad hoc network. For improvement in security and performance, an algorithm is proposed and named as efficient secure routing algorithm (ESRA), which is based on a dual authentication scheme having a moderate level of time and space complexities. The proposed algorithm is implemented in two stages. At the first stage, the malicious nodes are detected depending upon the destination sequence number without using encryption and decryption. Authentication is checked using public key cryptography in the second stage, which provides less computational complexity. The comparison of the proposed algorithm with other secure routing protocols using the National Choi Tung University (NCTUns) simulator has been done. The proposed scheme in this research has the capability of preventing malicious attacks like tracking location, manipulation, impersonation, wrong information, Sybil, replay, and DOS, and it also supports traditional security needs and traceability. The main advantage of our proposed algorithm is that it uses a short key length leading to speedy encryption, and it consumes less power. Although there is a little disadvantage associated, it increases the size of encrypted text, yet security is not compromised. The result shows that throughput increases by 25% in the proposed algorithm, numbers of collisions are lesser, and packet drop is reduced by 15%. The results prove that this novel proposed algorithm is more effective in a sparse vehicular environment, is lightweight and secure, and

D. Dembla

Department of IT and Computer Application, JECRC University, Jaipur, India

P. Tyagi

Department of Electronics and Communication, JECRC College, Jaipur, India

Y. Chaba (✉)

Department of CSE, Guru Jambheshwar University of Science and Technology, Hisar, Haryana, India

M. Chaba

Netaji Subas University of Technology, New Delhi, India

S. K. Chatrath

University of Canberra, Canberra, Australia

finds applications in e-health care, smart ecosystem, and intelligent transportation systems, etc.

Keywords VANET · ECDSA · Authentication · Security · ECIES

1 Introduction

In modern scenarios, driving has become more critical due to the increasing number of vehicles on the highway. VANET provides a solution to the above issue by transferring information among vehicle to vehicle (V2V) and vehicle to roadside (V2R) [1]. There are two types of units for wireless communication like On-Board Unit (OBU) which is mounted on each vehicle and Road-Side Units (RSUs) located on the side of the road. The primary concern of this technology is to improve the overall safety of the vehicle. Security in VANETs routing is a more crucial issue because it suffers from many attacks, which can occur from the network. The working of the network is depreciated by these attacks, so many problems are created in the network. But in vehicular ad hoc networks, computation overhead, delay, and detecting malicious nodes in the field of routing are a common problem despite the research has defined numerous security improvement techniques. Right now, in VANET scenarios, no efficient and secured routing protocol provides sufficient security and efficiency in the network.

2 Related Work

A secured scheme that is based on privacy preservation has been proposed in [2]. This mechanism not just fulfills the security requirements but also protects the vehicles guaranteeing trust authority. The performance of this method indicates that it is more effective in terms of overhead cost for communication, but this method does not consider the private keys distribution and management. The authors have also done a comparative study [3] of symmetric schemes and public key schemes in vehicular ad hoc network scenarios. It proves that symmetric keys, are better than public keys in terms of performance because it has fewer overheads of computation and communication both. An identity-based management solution has been proposed which is an effective technique [4].

The authors [5] proposed a technique to secure the messages based on hop count with a digital signature and hash chains. PK sharing methodology is utilized in this protocol. It requires both hash chains and fingerprints for transfer messages, so it is computationally heavy. The proposed scheme does not prevent the attack initiated by the intermediate nodes in networks. Researchers have proposed an approach to protect VANETs from various interventions like distributed denial of service [6].

In 2019, a P-Gene method for security analysis has been suggested by using the technique of data hiding [7]. In this approach, every data item is latent and thereby securing the privacy of the data. The analysis based on simulation shows the method of securing the text aloofness. In 2020, the key monitoring has been investigated depending on the blockchain for the VANETs [8]. In the beginning, to naturally recognize the listing, amend, abrogation for the user's populace, the author has introduced a dynamic dispersed monitoring system for VANETs using blockchain (DB-KMM). The paper [9] has also analyzed the privacy of DB-KMM.

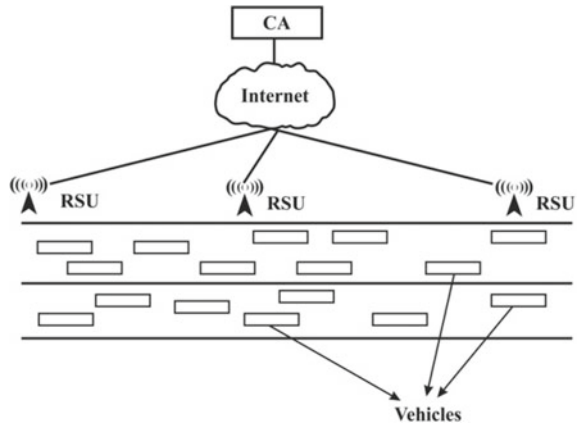
In addition to this, depending on bivariate polynomial, a featherweight mutual verification along with arranged protocol has been proposed. The paper has also analyzed the privacy of DB-KMM. The system has been introduced in the paper which eliminates the attack of a collision through RSU [10]. Besides this, the goal of the authors is also to use a few formal schemes for the verification and privacy of the system [11]

Although researchers have defined numerous security improvement techniques still in VANET scenarios, no efficient and secured routing protocol provides the highest security and efficiency in the network. Still, there is a lot of scope for the design of an efficient and secured routing algorithm. An efficient secure routing algorithm (ESRA) based on a dual authentication scheme having a moderate level of time and space complexities can be developed.

3 Proposed Model and VANET Scenario

In the proposed solution, we will use ECDSA in comparison with RSA for key generation and verification because it takes very little time for key generation in comparison with RSA. ECDSA algorithm is suitable in many applications like VANET, WSN, and smart card due to the presented performance and security. The detailed operation of ESRA is explained in the proposed algorithm. We have proposed (ESRA) algorithm, based on new dual security and authentication process which provides high security in a VANET environment when vehicles are communicating with each other. For providing security in dual mode, at the first level, we propose a scheme that detects malicious nodes based on destination sequence numbers with the help of a special data structure, i.e., heap. At the second level, to verify the authenticity, public key cryptography is used with the help of ECDSA and ECIES schemes. The dual security scheme implemented in this research is computationally efficient that supports secure communication from CA to RSU and RSU to OBU. The objective of the proposed algorithm is to block the unauthorized participation of malicious nodes with high-performance characteristics.

Fig. 1 VANET model with CA, RSU, and vehicles



3.1 System Model

Figure 1 depicts the system model. It consists of CA, RSU, and vehicles. Certificate authority (CA) is a fully trusted authority. At the beginning of system initialization, each vehicle and RSU has to register itself with CA to get certificates assigned to it. This entity performs the role of assigning anonymous certificates to vehicles and storing the same in the certificate database. Road-Side Unit (RSU) is installed at the intersections and curves of the path, which acts as a gateway to connect vehicle to vehicle. They are regularly monitored and managed by CA. RSU located at the side of the road is fixed and connected to the CA. Each vehicle communicates with the other taking the help of an On-Board Unit (OBU). Each vehicle generates a public key pair using ECDSA. Vehicles send an RREQ to all neighbors and which one has a route to destination sends an RREP to the source vehicle.

3.2 Working and Assumptions of System Model with Proposed Algorithm

Figure 2 shows the communication between vehicles through RREP message. In our proposed scheme, some important assumptions are considered for secure communication in VANET. These assumptions are as follows.

- CA is stronger than OBU and RSUs regarding computation, storage, and communication. All nodes have CA's public key at the time of registration.
- Before the time of registration, all nodes generate their private and public keys.
- Before starting communication, each node exchanges key pairs to each other.
- Each vehicle equipped with a positioning device (e.g., GPS) and secured hardware module (SHM) can get accurate time information, and SHM is tamper resistant.

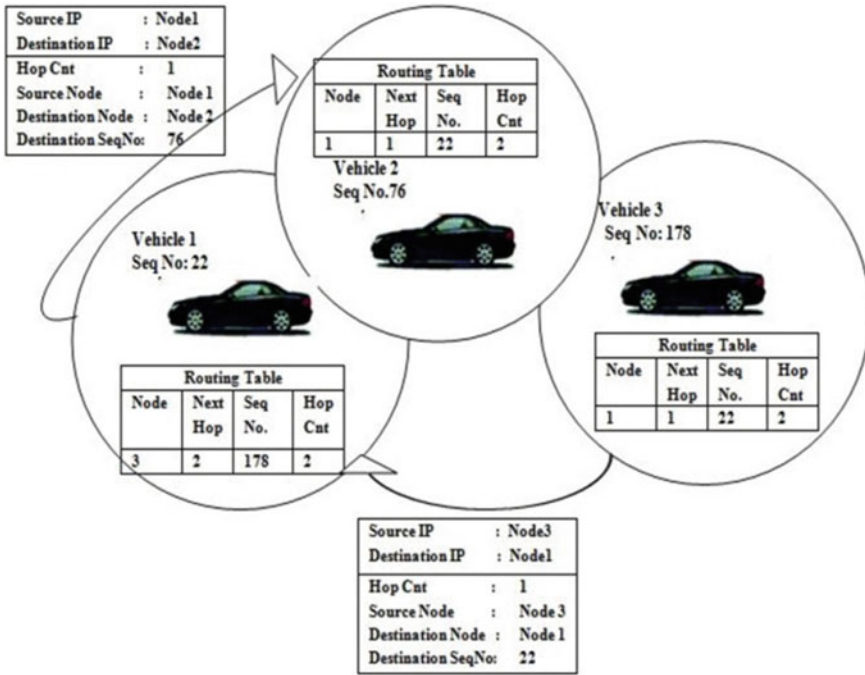


Fig. 2 RREP message in VANET

3.3 Proposed Algorithm (ESRA)

The ESRA operation is explained in detail in this section. ESRA prevents most of the attacks presented in the network using the cryptographic function. ESRA provides end-to-end authentication process. This protocol is more optimized in comparison with other secure ad hoc routing protocols. In ESRA, source node broadcasts a route discovery message, and the node replies to it. Only authorized nodes participate in each hop communication between source and destination, so routing messages are at the end to end authenticated. The following steps explain the proposed algorithm.

3.3.1 Certification by CA to Authorized Nodes

In ESRA, cryptographic certificates provide authentication and message integrity during the route discovery process. For this, it requires a certificate authority (CA), and all valid nodes have a public key CA. CA issues a certificate and maintains a key pair of each node. Nodes first registered share a key with CA and get a certificate. Before starting communication, each node must have a certificate from CA. Table 1 shows the notation used in our equations.

Equation 1 shows a certificate that is node *S* collected from CA as follows:

Table 1 Notations used in equations

Notations	Description	Notations	Description
CA	Certificate authorities	E	Time to live
$cert_A$	Certificate CA for node A	X	Destination
IP_A	IP of (A) node registered with CA	REP	Route reply
PK_A	Public key of node A	RDP	Route discovery packet
PR_A	Node A private key	RREQ _{ID}	A unique sequence number
T	CA timestamp	N	No. of nodes
PR_{CA}	CA private key	Cert _D	Route reply certificate
PK_{CA}	Public key of CA	IP_X	IP address of the destination
N_S	Nonce value	S_N	Encrypted sent by source node
t_0	Source node timestamp	Cert _A	Certificate of A
S_s	Sequence of source	PR_B	The private key of B
S_D	Sequence of destination	Cert _B	Certificate of B

$$CA \rightarrow S: cert_S = [IP_S, PK_S, t, e]PR_{CA} \quad (1)$$

The certificate includes the source IP address, source public key (PKS), the timestamp t at which the certificate was created, and e is the time to live for the certificate. Certificates are used to authenticate. All nodes must have valid certificates, and all certificates have an issue and expire time.

3.3.2 Secured Route

The main purpose is to check that the intended destination was reached. The source selects the return path based on trust in the destination. Route discovery begins from source node S to destination X , which is depicted in Eqs. 2 and 3. It broadcasts route discovery packet (RDP) to all its neighbors:

$$RDP: (RREQ_{id}, S_S, D_S, H_C), S_N: [N_S, t_0, e]PK_X \quad (2)$$

$$S \rightarrow \text{broadcast: } RREQ(S_N, [RDP, IP_X]PR_S cert_S) \quad (3)$$

The RDP includes RREQ_{id} (a unique sequence number), source sequence number (S_S), destination sequence number (D_S), and hop count (H_C), S 's certificate ($cert_S$), and all certificates signed with S 's private key. The content of RDP is not encrypted and the source signs it, so that the contents are viewed publicly. Each time source node monotonically increases the nonce and performs route discovery. When any node receives RDP message, a reverse path will be set up between the source and neighbor who send RDP message. The intermediate node uses S 's public key, to verify the signature and also check the validity of S 's certificate. Message signs by

the receiving node and the message append its certificate and broadcast the message to each of its neighbors. Spoofing attacks are prevented by the use of a signature. Suppose B is an intermediate neighbor that has received from S and rebroadcasts the request which is depicted in Eq. 4.

$$B \rightarrow \text{broadcast: RREQ}(S_N, [[\text{RDP}, \text{IP}_X]\text{PR}_S]\text{PR}_B, \text{cert}_S, \text{cert}_B) \quad (4)$$

After receiving the RDP, C is the neighbor of B's which certifies the signatures for S and then rebroadcasts the RDP using Eq. 5. Each intermediate follows the same steps as

$$C.C \rightarrow \text{broadcast: RREQ}(S_N, [[[\text{RDP}, \text{IP}_X] \text{PR}_S] \text{PR}_B] \text{PR}_C, \text{cert}_S, \text{cert}_B, \text{cert}_C) \quad (5)$$

3.3.3 Validated Route Setup

The destination X receives the message from the source that replies to the first RDP with a given nonce (NS). The path which has the less number of hop counts has lower latency, so RDP traveled through the path which has the least number of hops. In this case, a less number of hop paths are feasible to be preferred because of the reduction in delay. At each hop, messages are validated, so there is no chance to divert the traffic by malicious nodes. The destination node delivers a route reply (REP) after receiving RDP packet with the backward path to the source. The destination node revises (S_N), and the certificate is attached before forwarding the REP to the next hop. Let X send the first reply (REP) to D using Eq. (6).

$$X \rightarrow D: (S_N, ([\text{REP}, \text{IP}_S]) \text{PR}_X, \text{cert}_X) \quad (6)$$

Above equation shows, the IP address of S (IP_S), packet-type identifier (.REP), the certificate of X (cert_S), and the nonce sent by S_N . The intermediate node that gets route replies toward back to the predecessor which they got the route discovery packet. The REP of every node signs by a source with the reverse path, and Equation 7 shows the route reply from node D's to the next hop C.

$$D \rightarrow C: (S_N, [[\text{REP}, \text{IP}_S]\text{PR}_X]\text{PR}_D, \text{cert}_S, \text{cert}_D) \quad (7)$$

Verification of D's signature is done by C, then after that C node signs on the message attached certificate before forwarding to the next. Equation 8 depicts forwarding messages from node C to B.

$$C \rightarrow B : (S_N, [[\text{REP}, \text{IP}_S]\text{PR}_X]\text{PR}_C, \text{cert}_S, \text{cert}_S) \quad (8)$$

Every node checks the signature which avoids impersonation attacks and replays attacks. Source node A stores all route reply in a table for a constant period. After storing all route reply nodes, select the maximum sequence destination number (S_D) from the table if it is abnormal. In the process, entry from the routing table is deleted, else it will verify the nonce value and the destination's signature.

3.3.4 Route Maintenance of Proposed Algorithm

ESRA is an on-demand routing protocol. There is no traffic in establishing a path; then the path is deleted in the routing table. The node generates an error (ERR) message in case of the inactive route. Due to the high mobility of nodes, some routes are damaged, and then ERR messages are sent from the node to all active routes. If route breaks between source S and destination X, then ERR message is generated from node C to node B as follows:

$$C \rightarrow B: (S_N, [ERR; IP_S; IP_X]PR_C, cert_C) \quad (9)$$

The original message is passed to the source node. It is very hard to recognize whether links are active or not when ERR messages are fabricated for these types of links. However, this method prevents impersonation and enables non-repudiation using the signature on the message.

4 Implementation of Proposed Efficient Secure Routing Algorithm (ESRA)

There are no security parameters in the existing AODV [12] routing protocol. Efficient and secure communication is provided by the proposed algorithm (ESRA). AODV routing protocol coding is modified, and it is simulated in NCTUns [13]. RREQ and RREP functions of AODV routing protocol in VANET are modified as per our requirements. Miracle library is used for cryptographic functions [14].

4.1 Simulation Scenarios and Parameters

The simulation is completed for vehicles traveling at different time intervals. These scenarios are designed using a network simulator (NCTUns). The mobility model used here is the traffic light mobility model (TLM) which provides real-time scenarios, just traffic light signals them to be alive of neighboring vehicles. Simulations are done with five black hole nodes and 15 nodes in the territory of 1000×1000 m in 100 s. In this section, the execution of the proposed algorithm is done. The

following simulations show the performance of protocols, i.e., AODV and SAODV (secured AODV) with the proposed algorithm (ESRA) in a simulated environment. In real-time scenarios, always some malicious nodes exist in the network, and they always disturb network performance.

4.2 Result for Network Performance Metrics

Performance of various network parameters like throughput, packet collision, packet dropped, routing overhead, and packet delivery ratio for the proposed ESRA algorithm is shown in Figs. 3, 4 and 5.

Throughput: Due to the high-security mechanism used in the proposed algorithm, it completely blocked the malicious nodes. In this algorithm, ECDSA is used for the key generation which is more efficient in comparison with the RSA algorithm. So, the performance of the proposed algorithm is improved in terms of throughput. Figure 3

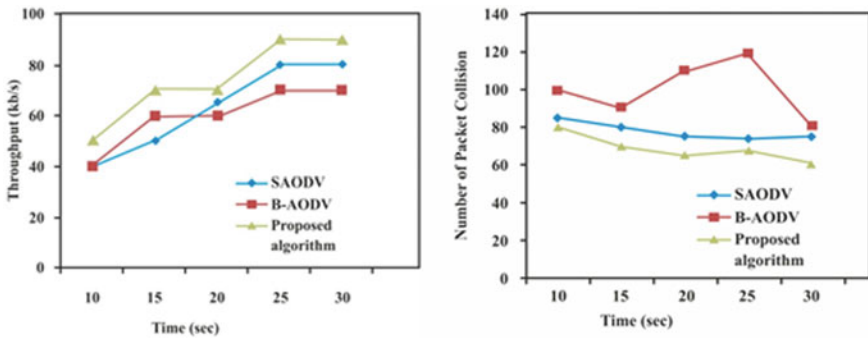


Fig. 3 Throughput and packet collision for proposed algorithm

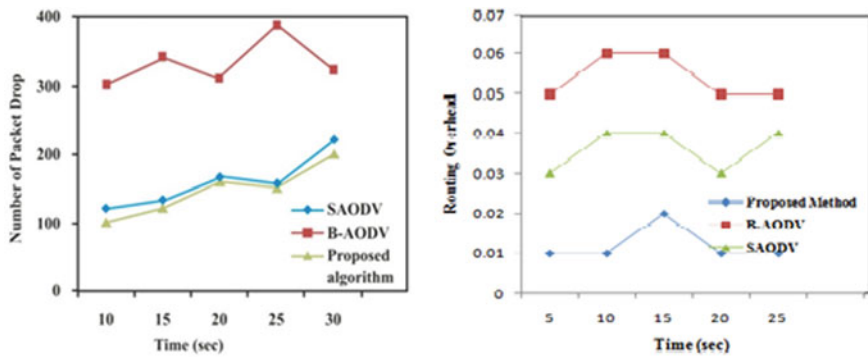
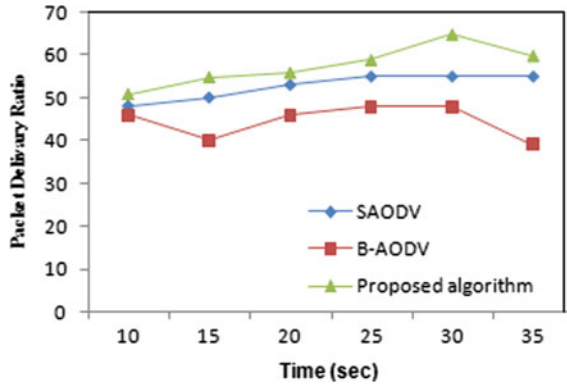


Fig. 4 Packet drop and routing overhead for proposed algorithm

Fig. 5 PDR for proposed algorithm



shows that this algorithm has the highest throughput approximately 25% higher than SAODV.

Packet Collision: Figure 3 shows that the proposed algorithm packet collision is low approximately 35% lower than the SAODV because this algorithm used a combination of ECDSA and ECIES algorithms to secure the data. Because it uses small keys, ECDSA performance has been proven to be efficient; thus, the computation cost is small compared with other cryptography algorithms.

Packet Drop: Packet drop is less in the proposed algorithm approximately 25% less than SAODV protocol as shown in Fig. 4 due to the use of lightweight hash algorithms to generate random functions for information transfer.

Routing Overhead: As shown in Fig. 4, the routing overhead of the proposed algorithm is less in comparison with standard SAODV and B-AODV, and overall overhead decreases.

Packet Delivery Ratio: Figure 5 shows PDR is high (10% higher than SAODV and 20% higher than B-AODV) because of this algorithm based on a cryptography scheme (ECDSA) which provides high security and less number of malicious vehicles.

Table 2 depicts a summary of the results. The simulation scenario was tested on NCTUNs for different performance parameters such as throughput, packet drop, packet collision, for AODV, SAODV (secure AODV), and our proposed algorithm, and it is evident from the result that in the proposed method, all these performance metrics show better results due to a smart check for the malicious node as well as better authentication keying used for the proposed algorithm.

4.3 Result for Security Performance Metrics

The key sizes for elliptic curve digital signature algorithm (ECDSA) vary from 165 to 573 bits, and for RSA, it varies from 1025 to 15,365 bits. In comparison with RSA, ECDSA requires less time for a key generation. From the last correlation,

Table 2 Summary of results

Performance parameters	AODV	SAODV (secured AODV)	ESRA (proposed algorithm)
Throughput (KB/s)	70	80	90
Packet collision	100	85	80
Packets dropped	240	200	180
Routing overhead	0.06	0.04	0.02
Packet delivery ratio	45	55	65

ECDSA takes 1.45 s while RSA took an aggregate of 680.07 s, so it is significantly faster than RSA. Because of the size of its small key, ECDSA performance has been proven to be efficient. So the computation cost of ECDSA is less in comparison with other cryptographic schemes. RSAs verification time was considerably faster than ECDSAs time, and it is increased as the size of key lengths increases. Consequently, it creates the impression that ECDSA has more advantages over RSA. Its small key sizes are useful in situations where resources, for example storage space, are constrained. Also, ECDSA has less time for key and signature generation, so it is much faster than RSA. Computational power is smaller for ECDSA. It provides less bandwidth and faster computation.

5 Conclusion

In this paper, we have proposed a new dual security scheme for improving the secured routing communications in the VANET environment, which has been implemented in the AODV routing protocol in VANET. For providing security in dual mode, at the first level, we have proposed a scheme, which detects malicious nodes based on destination sequence numbers with the help of a special data structure, i.e., heap. At the second level, public key cryptography is used to provide the authenticity and confidentiality of the message. As compared to other existing techniques, the dual security scheme implemented in this research is computationally efficient in terms of time and space complexity and supports secured communication from CA to RSU and RSU to OBU. Using NCTUns simulator, the performance of the proposed efficient secure routing algorithm (ESRA) protocol is compared with different protocols such as B-AODV (black hole attack) and SAODV (secured AODV). The proposed algorithm is more efficient in a sparse environment, and results prove that it has around 25% higher throughput in comparison with other protocols. Packet drop is also low, and the number of collisions has been reduced. The proposed scheme in

this research has the capability of preventing malicious attacks like tracking location, manipulation, impersonation, wrong information, Sybil, replay, and DOS, and it also supports traditional security needs and traceability. The main advantage of our proposed technique is that it uses a short key length leading to speedy encryption, and it consumes less power.

References

1. S. Zeadally, R. Hunt, Y. Shyan Chen, A. Irwin, A. Hassan, Vehicular ad-hoc networks (VANETS): status, results, and challenges. *Telecommun. Syst.* **50**(4), 217–241 (2012)
2. H. Zhong, J. Wen, J. Cui, S. Zhang, Efficient conditional privacy-preserving and authentication scheme for secure service provision in VANET. *Tsinghua Sci. Technol.* **21** (2016)
3. S. Ibrahim, M. Hamdy, A comparison on VANET authentication schemes: public key vs. symmetric key (2016)
4. K.M. Mahesh Kumar, N.R. Sunitha, R. Mathew, M. Veerayya, C. Vijendra, Secure ad-hoc on-demand distance vector routing using identity-based symmetric key management (2016)
5. X. Zhang, X. Bai, Research on routing incentive strategy based on virtual credit in VANET (2019)
6. Y. Wang, H. Zhong, Y. Xu, J. Cui, G. Wu, Enhanced security identity-based privacy-preserving authentication scheme supporting revocation for VANETs. *IEEE Syst. J.* **14** (2020)
7. P.S. Kumar, L. Parthiban, V. Jegatheeswari, Context-aware privacy, and security using P-Genes based on pseudonym in VANETs, in *Proceedings of IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (2019)*, pp. 1–5
8. Mihai, N. Dokuz, M.S. Ali, P. Shah, R. Trestian, Security aspects of communications in VANETs, in *Proceedings on 13th International Conference on Communications (2020)*, pp. 277–282
9. R. Sugumar, A. Rengarajan, C. Jayakumar, Trust-based authentication technique for cluster-based vehicular ad hoc networks (vanet). *Wireless Netw.* **24** (2018)
10. X. Hu, W. Tan, C. Yu, C. Ma, H. Xu, Security analysis of certificate-less aggregate signature scheme in VANETs (2019)
11. P. Tyagi, D. Dembla, Performance analysis and implementation of a proposed mechanism for detection and prevention of security attacks in routing protocols of the vehicular ad-hoc network (VANET). *Egyptian Inf. J.* **18**, 133–139 (2017)
12. M.C. Chuang, J.F. Lee, Team: Trust-extended authentication mechanism for vehicular ad hoc networks. *IEEE Syst. J.* **8** (2014)
13. S. Dokurer, Y.M. Erten, C. Erkin Acar. Performance analysis of ad-hoc networks under blackhole attacks (2007)
14. H.A. Esmaili, M.R.K. Shoji, H. Garage, Performance analysis of AODV under black hole attack through use of OPNET simulator. *World Comput. Sci. Inf. Technol. J.* **1**(2), 493–52 (2011)

An Efficient Feature Fusion Technique for Text-Independent Speaker Identification and Verification



Savina Bansal, R. K. Bansal, and Yashender Sharma

Abstract Speaker identification and verification is an important research area that finds applications in forensics voice verification, mobile banking and security authentication for access control. Various techniques for feature extraction are available in the literature. In this work, a speech feature fusion extraction technique based on fusion of time domain, frequency domain and cepstral domain features has been proposed. Supervised machine learning classification algorithms are used for speaker feature classification. Performance of proposed technique has been evaluated on two open-source speech datasets. Performance metrics of training time and accuracy (validation and test) are measured with the help of confusion matrix. The results indicate that even with smaller training datasets, the average accuracy achieved is 2.97 and 8.97% better and training time 1.95 and 2.03 s less as compared to MFCC and (MFCC + delta + delta delta) MFCC + Δ + Δ^2 , respectively.

Keywords Speech feature extraction · Feature fusion · Speaker identification · Supervised machine learning

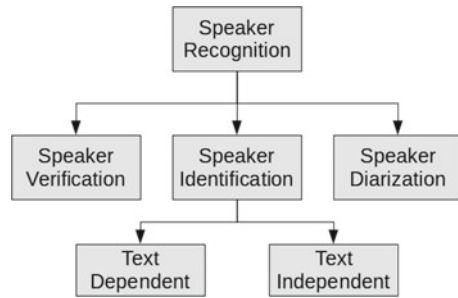
1 Introduction

Speaker identification and verification is a challenging area. It covers a wide range of utilities like authentication, surveillance, forensics speaker recognition, speech recognition, multi-speaker tracking, and personalized interfaces [1]. Speaker identification and verification comes under the area of speaker recognition. It is the method of knowing the distinctiveness of the speaker through speech uttered [2]. Speaker recognition system works on the principle that each talker's speech is idiomatic like fingerprints and thus can be used to identify speaker or validate his/her claim. These systems in general analyze characteristics or features in speech that are unique among speakers (Fig. 1).

S. Bansal · R. K. Bansal · Y. Sharma (✉)

Department of ECE, GZSCCET, MRSPTU Bathinda, Bathinda, Punjab 151001, India

Fig. 1 Classification of speaker recognition applications



Depending upon the application, speaker recognition is broadly divided into three classes [2, 3]. *Speaker identification (SI)* is used to identify a particular speaker out of the pool of many speakers. It is a method of discovery of speaker that delivers an assumed utterance. *Speaker verification (SV)* is the method of verifying the identity of a person using speech signals. *Speaker diarization (SD)* is the method of segmenting the input speech rendering to speaker identity.

Feature extraction and feature classification are the major steps involved in speaker recognition. Certain features that are essential for proof of identity of utterer is extracted from the speech data. While in feature classification, the features of an unidentified individual are taken and equated with the feature database of registered speakers to detect the original speaker. High inter-speaker and low-intra speaker variations, and easily measurable features are some important traits of an ideal speaker recognition system. Keeping in mind the traits of an ideal speaker recognition system and the need for less complex training data, a new fusion-based hybrid feature extraction technique (FFHT) is proposed. The fusion of multi-domain features results in lesser training time with comparable quality due to high variance in intra-speaker features set. Supervised machine learning classifiers such as linear discriminant analysis, ensemble of subspace discriminant classifiers, and wide neural network are used for feature classification.

2 Related Works

With the integration of machine learning-based set of rules and signal processing systems, audio signal processing has fully-fledged especially in the area of signal analysis and classification. Performance of any ML system hinge on mostly on the features which create training and testing datasets. Thus, feature extraction plays an important role for the reason that its ability to significantly affect performance of speaker classification model.

Hanifa et al. [3] provided a review on techniques and challenges of speaker recognition in the last decade. Their work presents a system and construction of speaker

recognition along with feature extraction and classifiers. Bai et al. [4] provided a wide-ranging summary of the deep learning-based speaker recognition and also examined the association amid diverse sub tasks, including speaker verification, identification, and diarization. Various feature extraction techniques from time domain, frequency domain, time–frequency domain, cepstral domain, phase domain, and deep features are reviewed in [5] and [6]. Murty and Yegnanarayana [7] combined residual phase information and MFCC features from 149 male speaker utterances using NIST 2003 dataset. Auto-associative neural network model was used as classifier with approximately 90% classification accuracy. Fong et al. [8] did a relative study to categorize utterers using time domain statistical features and machine learning classifiers. Researchers gained accuracy of around 94% with the multi-layer perceptron classifier. Ali et al. [9] projected a speaker identification model for recognizing 10 diverse speakers using Urdu language dataset. Their work merged deep learning-based and MFCC features using support vector machine (SVM) algorithm and claimed 92% classification accuracy. Soley-manpour et al. [10] explored clustering-based cepstral features united with an ANN classifier to sort 22 speakers from the ELDSR dataset and accomplished 93% classification accuracy. Selva Nidhyanathan et al. [11] projected a set of discriminative features to categorize utterances of 50 utterers from MEPCO speech dataset. They take out RASTA-cepstral features to categorize speaker utterances. Features were entered into a GMM-UBM classifier and reached 97% classification accuracy. Mohammadi et al. [12] presented fusion of four systems based on different speech features like MFCC, IMFCC, LFCC, and PNCC to advance verification accuracy under clean and noisy speech conditions. The evaluation was done on Gaussian mixture model that results in enhancing the accuracy of speaker verification system and reduces the equal error rate in some cases. The work by Jahangir et al. [13] proposed a fusion of cepstral and time-based features that were classified using DNN machine learning classification algorithm on LibriSpeech dataset. The comparative results indicate that DNN outperformed the other classification algorithms with a general accuracy of 92.9%. Bansal et al. [14] presented fusion-based hybrid feature extraction technique which incorporates multi-domain features. ANN model is used for speaker recognition. Their proposed technique achieved better accuracy over contemporary speech feature extraction techniques.

3 Proposed Methodology

Speech feature extraction is an important step in speaker identification and verification. There are numerous speech feature extraction techniques available to choose from. The proposed FFHT employs a fusion of different speech feature extraction techniques on the basis of their domain. Proposed FFHT includes zero crossing rate (from time domain features), root mean square energy, chroma feature, spectral centroid, spectral roll off and spectral bandwidth (from frequency domain features), and Mel frequency cepstral coefficients (from cepstral domain features). Feature fusion set file is then used in machine learning classifiers as training and testing

datasets. Confusion matrix is used for validation and test accuracy along with training time as performance evaluation metrics to compare with other feature extraction techniques like MFCC [15] and MFCC + Δ + Δ^2 [15] as detailed below.

3.1 Feature Extraction

Speech signal varies unceasingly, due to articulately movements. So, signal is divided into minor frames, wherein each signal frame is considered stationary, also known as short time processing. Respective frame is characterized by a spectral feature vector. Feature extraction algorithms were used in this work are briefly discussed here:

3.1.1 Time Domain Features

The modest method to examine any signal is in its original form. Time sequence signals evolves with time. By visualizing a signal in time domain, key characteristics of a signal can be analyzed. However, in real time, audio signals are non-stationary.

Zero Crossing Rate (ZCR) is a time domain feature that represents the quantity an audio signal crosses the zero-amplitude level or horizontal-axis during one second interval. It provides an approximation of the central rate of recurrence in the signal.

3.1.2 Frequency Domain Features

Frequency domain analysis is of extreme reputation in audio signal processing. Frequency domain features are given below:

Root Mean Square Energy (RMSE) represents the square root of mean squared amplitude for a specified and short time window.

Chroma Features (CF) are linked to the observation of the pitch.

Spectral Centroid (SC) indicates the center of the mass of the audio spectrum. It defines the intensity of a sound signal.

Spectral Roll off (SR) point is the 95% of the power spectral distribution. It is also known as the measure of skewness of the spectral shape.

Spectral Bandwidth (SB) is the 2nd order statistical rate that helps to discriminate tonal sounds (with low bandwidths) from the noise-like sounds (with high bandwidths).

3.1.3 Cepstral Domain Features

A cepstrum is attained by the inverse FT of the logarithm of signal spectrum. There are complex, power, phase, and real cepstrum. Analysis of the cepstrum is called as cepstrum analysis.

Mel Frequency Cepstral Coefficients (MFCC) is used as a perceptual weighting technique that duplicates how we perceive sounds such as music and speech. Cepstrum gives information about how those frequency components change. Mel frequency scale and cepstral analysis, make MFCC quite useful in speaker recognition.

3.2 Feature Classification Using Supervised ML Classifiers

A classifier in machine learning categorizes data into specified set of a class or classes. Supervised classifiers are fed with labeled training datasets, which they use to learn and to classify new data according to predefined classes. Three types of supervised machine learning classifiers are used in this work, as discussed below:

Linear Discriminant Analysis

Discriminant analysis is a common classification method due to its fast and precise abilities. It assumes that dissimilar classes produce information based on dissimilar Gaussian distributions and create linear boundaries among classes [16].

Ensemble-Based Subspace Discriminant

Ensemble learning helps in improving the machine learning results by combining several models. For the subspace ensemble classifier, it is deemed that each subspace classifier also has its own salient regions in its corresponding subspace [17].

Wide Neural Network

A neural network is basically an interconnection of layers. Layers are made up from basic units called as perceptron, which consists of an input terminal, processing unit, and an output terminal. WNN models are having a fast prediction speed, medium memory usage, and medium model flexibility. The model used here is a feed forward, fully connected neural network for the classification.

3.3 Proposed Feature Fusion Technique

The proposed FFHT is fusion of zero crossing rate (from time domain features), root mean square energy, chroma feature, spectral centroid, spectral roll off and spectral bandwidth (from frequency domain features), and Mel frequency cepstral coefficients (from cepstral domain features). 20 MFCC bins and 1 bin each from ZCR, RMSE, CF, SC, SR, SB thus comprising 26 feature set proposed technique. Mean values of frequency domain features are used in their respective bins. The pseudo-code for proposed technique is given below (Table 1).

Table 1 Pseudo-code of proposed feature fusion technique

Proposed feature fusion technique

- **Result: Seven** features comprising total of 26 feature set
- Input: Speech signal $y_i(k)$, sampling rate sr, time frame t
- Create.csv file with “filename, ZCR, RMSE, CR, SC, SR, SB, MFCCs” headers
- $ZCR = \sum(|\text{diff}(y_i(k) > 0)|)/\text{length}(y_i(k))$
- $RMSE = \mu(\sqrt{(\frac{1}{T} \sum_k (|y_i(k)|)^2)})$
- Convert $y_i(k)$ into frequency domain
- CR = mean (logarithmic STFT of the sound signal)
- SC = $\mu((\sum_{i=b}^c f_i S_i)/(\sum_{i=b}^c S_i))$, where f_j is frequency corresponding to i , S_i is spectral value at i , b , and c are band edges
- spec_roll = r such that $\sum_{j=b}^r S_i = 0.95(\sum_{i=b}^c S_i)$, where S_i is spectral value at i , b , and c are band edges
- SR = mean (spec_roll)
- spec_bw = mean (second order statistical value)
- MFCC(n) = DCT $_n$ coefficients, where $n = 1-20$

4 Performance Analysis

To estimate the performance of proposed technique, diverse performance metrics as used in literature, like confusion matrix, validation and test accuracy, and training time are used.

Confusion matrix gives summary of all the predictions made by the classifier for training or testing given in the tabular representation and used to calculate the performance of a classification model.

Accuracy is used for evaluating classification models. Informally, accuracy is the ratio of right predictions made by a model to the total number of predictions. *Validation accuracy* is the accuracy achieved on validation dataset (part of training set which is used to evaluate performance while tuning the hyperparameters). *Test accuracy* is the measure of model’s performance on new data it has not seen before.

Training time is the time taken by a classifier model to optimally classify training set samples. Training time needs to be less for making a better and efficient classification model.

For testing the goodness of our fusion model two open-source datasets are used for speaker identification and verification. **Dataset 1: Pitch Tracking Database from Graz University of Technology (PTDB-TUG)** [18] provides a microphone and laryngograph signals of 20 English native speakers (female and male speakers, 10 each). The database consists of approximately 4720 recorded sentences. The recording studio of the Institute of Broadband Communications at Graz University of Technology is used for all the recordings. **Dataset 2: Free ST American English Corpus** [19] is provided by Surfing Tech. Here, recordings are done in indoor environment using a cellphone comprising sounds from 10 speakers, each speaker is having 350 sounds on average.

The FFHT demonstrated lesser training time, due to high variance in intra-feature sets and comparable accuracy as to the other feature extraction techniques, however using a much smaller dataset for the system’s training purpose. Training time comparison of proposed technique along with MFCC and MFCC + Δ + Δ^2 for dataset-1 on LDA, ESD, and WNN classifiers shown in Fig. 2. Analysis indicates that proposed technique takes average training time of 4.49 s as compared to 6.68 and 6.95 s of MFCC and MFCC + Δ + Δ^2 , respectively.

Comparison of proposed technique with MFCC and MFCC + Δ + Δ^2 on the basis of training time for dataset-2 shown in Fig. 3, which indicate that proposed technique takes average training time of 6.56 s, whereas MFCC and MFCC + Δ + Δ^2 techniques takes 8.26 s and 8.16 s, respectively. For linear discriminant analysis and ensemble-based subspace discriminator, proposed technique outperforms other techniques but takes approximately 0.77 s more than MFCC for wide neural network classifier.

Figure 4 shows average accuracy comparison of proposed FFHT along with MFCC and MFCC + Δ + Δ^2 for dataset-1. Results indicate comparable average accuracy between proposed technique and MFCC, 95.09% and 95.03%, respectively. Proposed technique exhibits 10.43% high accuracy than MFCC + Δ + Δ^2 technique (84.66%), however using a much smaller dataset for the system’s training purpose.

Comparison of proposed technique along with MFCC and MFCC + Δ + Δ^2 on the basis of average accuracy for dataset-2 is shown in Fig. 5. Proposed FFHT

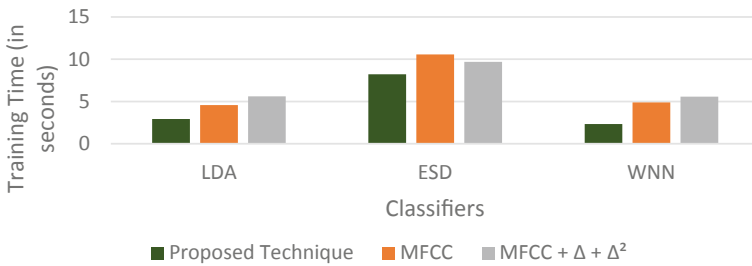


Fig. 2 Training time comparison for dataset-1

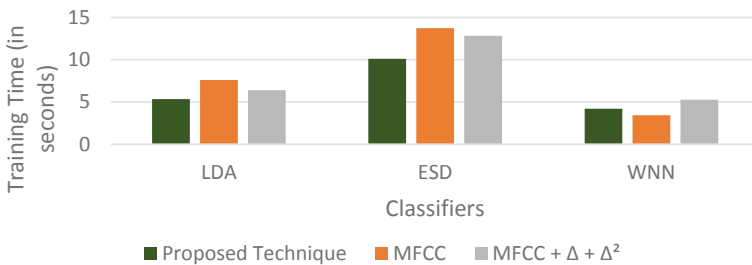


Fig. 3 Training time comparison for dataset-2

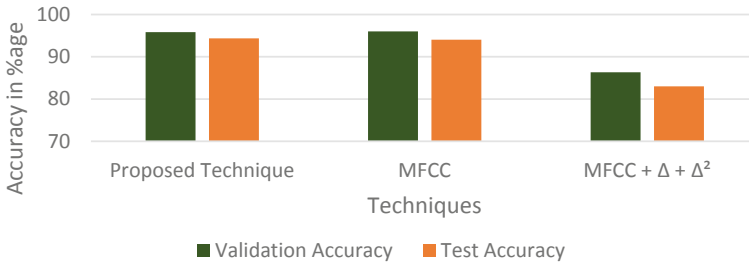


Fig. 4 Average accuracy comparison for dataset-1

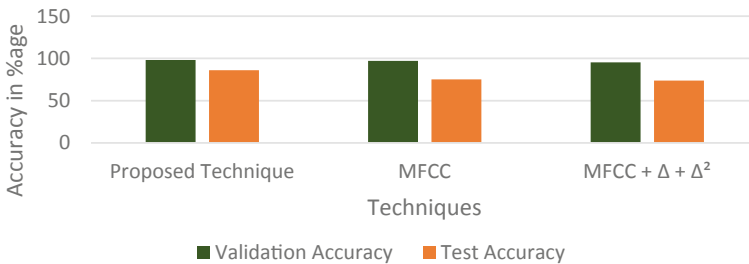


Fig. 5 Average accuracy comparison for dataset-2

outperforms available compared algorithms, on average accuracy 92.09%, 86.21%, and 84.58%, respectively using a much smaller dataset for the system's training purpose. Validation accuracy remains comparable with MFCC but in case of test accuracy, proposed technique performs much better than MFCC and MFCC + $\Delta + \Delta^2$.

Validation data confusion matrix of proposed FFHT for linear discriminant analysis in case of dataset-1 shown in Fig. 6. It gives summary of all the predictions made by the classifier. It also shows the detailed results of truly predicted classes and false predicted classes.

5 Conclusion

A novel feature fusion technique for speaker identification and verification is hereby proposed and evaluated. The performance of proposed technique is analyzed with respect to available feature extraction techniques like MFCC and MFCC $\Delta + \Delta^2$. Supervised machine learning classifiers are used for feature classification. Initial results demonstrated better accuracy and lesser training time on account of high variance in intra-feature sets. Furthermore, results of the proposed work have been achieved with lesser number of training samples per speaker. The proposed technique has also resulted in reduction of model complexity coupled with lower amount of

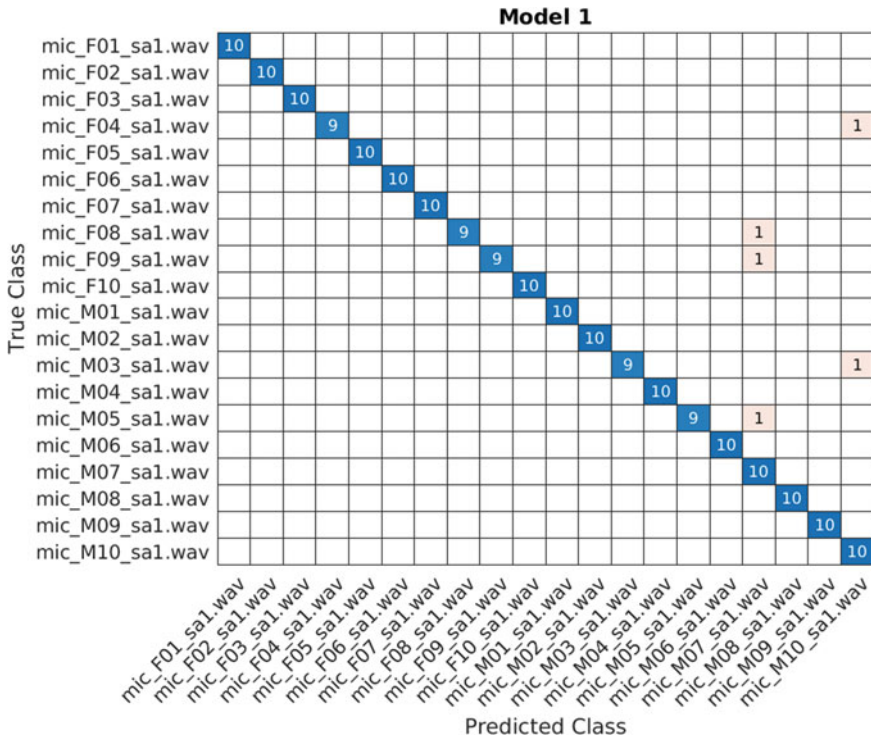


Fig. 6 Validation data confusion matrix

training data. The initial results are clear indication of better accuracy and fast training time, more exhaustive results shall be presented in our future work.

References

1. H. Garg, R.K. Bansal, S. Bansal, Improved speech compression using LPC and DWT approach. *Int. J. Electron. Commun. Instrum. Eng. Res. Dev. (IJECIRD)* **4**(2), 155–162 (2014)
2. Z. Zhang, Mechanics of human voice production and control. *J. Acoust. Soc. Am.* **140**, 2614–2635 (2016). <https://doi.org/10.1121/1.4964509>
3. R.M. Hanifa, K. Isa, S. Mohamad, A review on speaker recognition: technology and challenges. *Comput. Electr. Eng.* **90**, 107005 (2021). <https://doi.org/10.1016/j.compeleceng.2021.107005>
4. Z. Bai, X.-L. Zhang, Speaker recognition based on deep learning: an overview. *Neural Netw.* **140**, 65–99 (2021). <https://doi.org/10.1016/j.neunet.2021.03.004>
5. G. Sharma, K. Umopathy, S. Krishnan, Trends in audio signal feature extraction methods. *Appl. Acoust.* **158**, 107020 (2020). <https://doi.org/10.1016/j.apacoust.2019.107020>
6. F. Alías, J.C. Socoró, X. Sevillano, A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. *Appl. Sci.* **6**(5), 143 (2016). <https://doi.org/10.3390/app6050143>

7. K.S.R. Murty, B. Yegnanarayana, Combining evidence from residual phase and MFCC features for speaker recognition. *IEEE Signal Process. Lett.* **13**(1), 52–55 (2006). <https://doi.org/10.1109/LSP.2005.860538>
8. S. Fong, K. Lan, R. Wong, Classifying human voices by using hybrid SFX time-series preprocessing and ensemble feature selection. *BioMed Res. Int.* **2013**(720834) (2013). <https://doi.org/10.1155/2013/720834>
9. H. Ali, S.N. Tran, E. Benetos et al., Speaker recognition with hybrid features from a deep belief network. *Neural Comput. Appl.* **29**, 13–19 (2018). <https://doi.org/10.1007/s00521-016-2501-7>
10. M. Soleymanpour, H. Marvi, Text-independent speaker identification based on selection of the most similar feature vectors. *Int. J. Speech Technol.* **20**, 99–108 (2017). <https://doi.org/10.1007/s10772-016-9385-x>
11. S. Selva Nidhyanthan, R. Shantha Selva Kumari, T. Senthur Selvi, Noise robust speaker identification using RASTA–MFCC feature with quadrilateral filter bank structure. *Wireless Pers. Commun.* **91**, 1321–1333 (2016). <https://doi.org/10.1007/s11277-016-3530-3>
12. M. Mohammadi, H.R. Sadegh Mohammadi, Robust features fusion for text independent speaker verification enhancement in noisy environments, in *2017 Iranian Conference on Electrical Engineering (ICEE)* (2017), pp. 1863–1868. <https://doi.org/10.1109/IranianCEE.2017.7985357>
13. R. Jahangir et al., Text-independent speaker identification through feature fusion and deep neural network. *IEEE Access* **8**, 32187–32202 (2020). <https://doi.org/10.1109/ACCESS.2020.2973541>
14. S. Bansal, R.K. Bansal, Y. Sharma, ANN based efficient feature fusion technique for speaker recognition, in *International Conference on Emerging Technologies: AI, IoT and CPS for Science & Technology Applications* (2021). <http://ceur-ws.org/Vol-3058/Paper-063.pdf>
15. M.A. Hossan, S. Memon, M.A. Gregory, A novel approach for MFCC feature extraction, in *2010 4th International Conference on Signal Processing and Communication Systems* (2010), pp. 1–5. <https://doi.org/10.1109/ICSPCS.2010.5709752>
16. E. Alexandre-Cortizo, M. Rosa-Zurera, F. Lopez-Ferreras, Application of Fisher linear discriminant analysis to speech/music classification, in *EUROCON 2005—The International Conference on “Computer as a Tool”* (2005), pp. 1666–1669. <https://doi.org/10.1109/EURCON.2005.1630291>
17. S. Sun, C. Zhang, Subspace ensembles for classification. *Physica A* **385**(1), 199–207 (2007). <https://doi.org/10.1016/j.physa.2007.05.010>
18. G. Pirker, M. Wohlmayr, S. Petrik, F. Pernkopf, A pitch tracking corpus with evaluation on multipitch tracking scenario. *Interspeech*, 1509–1512 (2011). Available Online <https://www2.spsc.tugraz.at/databases/PTDB-TUG/>
19. ST-AEDS-20180100_1, Free ST American English Corpus. Available Online <https://www.openslr.org/45/>

Identifying Forged Digital Image Using Adaptive Over Segmentation and Feature Point



Nemani Nithyusha, Rahul Kumar Chaurasiya, and Om Prakash Meena

Abstract Wide availability of free image editing tools has made it very convenient to edit digital images. Image editing with wrong intentions is becoming a serious problem. The existing detection method deals with feature point equating and adaptive over segmentation in this work. Suggested conspire coordinates reality-based forgery and block-based detection techniques. Image tempering with bad intentions is becoming more common in the news in media field, patient data in medical field, and several other fields. This paper proposes an adaptive over segmentation-based detection of forged images. To begin, suggested algorithm slices the input picture into well separated and asymmetrical blocks suitably. At that point, the points extricated from scale invariant feature transform (SIFT) are again extricated from individual block. The features such obtained are called block features (BFs). To find labeled feature points, these BFs are equated with one another; this technique roughly shows the speculated fraud locales called surmised forgery locales. To identify the forgery locales correctly, the paper suggest an algorithm called forgery local extrication. Using the algorithm, the super pixels are replaced by the feature points and afterward combines similar local color features into the feature blocks to produce the unified locales. Lastly, to produce the detected forgery locales, we apply the morphological operation to the unified locales. The analyzed outcome and comparative analysis demonstrates that the proposed detection plan accomplishes prominent detection outcomes.

Keywords Digital image forensics · Adaptive-segmentation · Image terming · Forgery detection

N. Nithyusha · R. K. Chaurasiya (✉) · O. P. Meena

Department of Electronics and Communication Engineering, Maulana Azad National Institute of Technology, Bhopal, India

e-mail: rkchaurasiya@manit.ac.in

1 Introduction

From the good old days, pictures are for the most part acknowledged as a proof of event of the past occasions. Advanced picture is a piece of this present reality that is created after numerous cycles of picture age. The advancement of the web has introduced the unbelievable turn of events and upgrades in the lofty. At present the procedures introduced has made the human race likable and protective, at any rate security to primary reports has a spot with the checked individual is stayed as pushed in the high level picture planning a territory.

Well according to the definition of an image, image is a 2 dimensional (2D) printing that has a similar appearance to some subject usually a physical object or a person but some believe that images speak the truth of the incident that took place or say the situation. Earlier days manipulation of an image captured by a traditional film camera was not possible for everyone it was only possible for those who had knowledge in manipulating images professionally, whereas nowadays it is possible and easy for everyone to manipulate the images. Storing of digital images was also not possible for everyone in the earlier days but now it possible for each and everyone. Not only storing it is the same case with sharing of large number of images.

At present, web and different applications broadly utilizes the computerized pictures. Copying the pictures without leaving any traces is exceptionally simple with the progressed tools. Issues exists with the advancement of techniques that are identified by the authenticity of the picture. The only solution for the image tampering is digital forensic. In the majority of the occurrence, image is being copied on the twin picture, i.e., alluded as replicating.

Before proposing the current approach, a brief survey was done on the existing methods on image forgery detection. In existing wavelet-based method involves transformation of image using discrete wavelet transform (DWT) and/or discrete cosine transform (DCT). One could detect images, but it is a time taking process. This is because the image has to be divided into many parts and in those parts, we have to compare each part whether it has original pixel value or not [1]. Traditionally, block-based forgery recognition is utilized to identify forgery picture. However, this approach confronts a few downsides. For example, overlapping rectangular blocks have to be separated by the input pictures, which would be computationally costly. This is because the area of the picture increases and it is less proficient as it requires more computation time [2].

Swaminathan et al. [3] introduced a technique to approximate both in-camera and post-camera operation fingerprints for identifying the coherence of photographs. A tampering detection algorithm is used in this paper. This paper develops a new methodology for the forensic analysis of digital camera images. The main aim of the paper was to produce good scalability for identification of unseen distortions. The absence of camera-imposed fingerprints from an input image indicates that the input test image is not a camera output and is possibly generated by other image.

Sevinc Bayram et al. [2] suggested a method for the detection doctoring in digital image. Doctoring typically involves a sequence of elementary image processing

operations, such as rotation scale and other transition invariant. In this paper, the authors suggested lexicographically sorting algorithm. Being robust to lossy jpeg compression, scaling, and rotation is the advantageous part of the method. The main disadvantage is its high computationally complexity.

To overcome the limitations of existing methods and to avoid block-based forgery, we suggest adaptive over segmentation (image blocking method). The method isolates the input picture into well separated blocks suitably with the assistance of algorithms those are simple linear interacting clustering (SLIC) algorithm to slice the input picture into unpredictable blocks. Further, DWT algorithm is utilized to dissect the time-period and frequencies of the super pixel. Next, the image block shaped to the block feature (BF) extrication strategy, where the BFs are separated by utilizing scale invariant feature transform (SIFT) algorithm as it had consistent and good execution contrasted with other extrication techniques. Additionally, the cycle of BF equating is done that utilized SLIC for computing super pixel and DWT for discovering super pixel from one block and studying other for different blocks [4].

At the point when the features are separated and equated then we become acquainted with what locales the input picture has been falsified. At last, forgery locale extrication algorithm and morphological operation are implemented to identify forgery locales precisely. A systematic flow chart of the proposed forgery detection method for digital images is described in Sect. 3.

To summarize, the algorithms of the proposed work are as follows:

- Adaptive over segmentation
- BF extrication
- BF matching
- Forgery locale extrication

The applications of forgery detection can be found in education purpose (college logos, certificates), currency, commercial purpose (branded clothes, objects), and digital forensic labs [3, 5].

The rest of the paper is structured as follows: Sect. 2 describes the traditional DWT-based method of forgery detection, along with its drawback. Section 3 presents the proposed methodology. The experimental results are presented in Sect. 4. The study is concluded in Sect. 5.

2 Traditional DWT-Based Method and Its Drawbacks

Before proposing a new method, we have implemented the traditional DWT-based method for image forgery detection. The description and drawbacks of the method are covered in this section as follows.

Only time or frequency component of a signal may not provide sufficient information. However, the wavelet transform of a signal represents both time and frequency components. Wavelet transform is equipped for giving the time and frequency information all the while, subsequently giving a time frequency portrayal of the picture.

The DWT deteriorates a specific signal into various set of signals, where individual set is a time series of coefficients depicting the time growth of the signal in the corresponding frequency band. DWT has inbuilt features that isolates the image into equivalent pixels. The DCT addresses a picture as a sum of sinusoids of varying magnitudes and frequencies. The DCT2 function processes the 2D DCT of a picture [6].

In DWT-based forgery detection method, a color image is first converted into gray scale image. Then, the DWT is used to find pixels in the image overlapping. Therefore, by using lexicographical sorting (setting matrices into rows and columns), a copy and move forgery is detected. We used DWT (rather than DCT) as a low-level segmentation method and found some drawbacks of it.

2.1 Drawbacks of the DWT-Based Method

Several pixel overlapping levels exists, and it is not easy to detect forgery.

- Unknown parts are not detected.
- Can be applied to only low-level segmentation.
- Due to the time variant property, image blurriness may influence the performance.

2.2 Motivation

Based on the experimental results, we analyzed the drawbacks of the DWT-based method. This has motivated us to propose a new method based on adaptive over segmentation and feature point equating. The advantages of the proposed method are highlighted below:

- Reduces pixel overlapping
- Reduce non overlapping blocks
- Has better time invariant property
- Find exact foreground image with more accuracy

3 Proposed Method

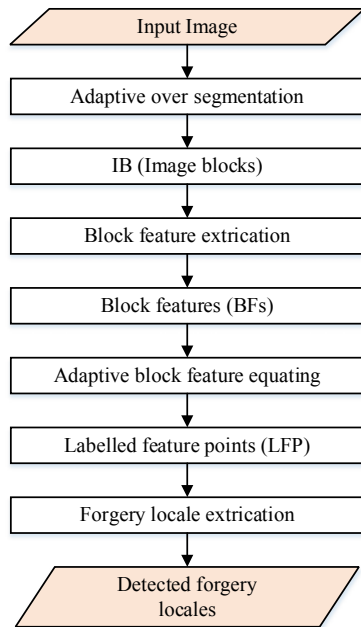
This section presents a detailed description of the proposed method.

3.1 Over Segmentation

In this phase, a feature point equating and adaptive over segmentation is described in depth in order to find forgery detection of an image. A flow diagram of the suggested image forgery detection strategy is depicted in Fig. 1. First, in order to slice the input image into well separated and asymmetrical blocks called image blocks (IB), adaptive over segmentation strategy is introduced. Then, scale invariant feature transform (SIFT) or super pixel technique is implemented in individual blocks. This extricate the SIFT feature points which are known as BF.

Additionally, these BFs and the feature points are equated with one another successfully in addition to attain labeled feature points (LFP), which precisely stimulate the surmised forgery locales. Lastly, the forgery local extrication method is implemented to detect the forgery locale from the input picture as stated by the extricated LFP.

Fig. 1 Block diagram of proposed system



3.2 Algorithms Used in the Proposed System

3.2.1 Louvain Method

The Louvain method is an algorithm to detect communities in large network [7]. It is a hierarchical clustering algorithm that repetitively unifies communities into a single node and carries out the modularity clustering on the condensed graph. This method works in 2 phases: (a) modularity and (b) community aggregation.

As shown in the Fig. 2, the Louvain method identifies the nodes that belong to similar community by optimizing the modularity and construct a new network by aggregating the community, when there is no change in labeling the process stops. Iteratively repeat the process of modularity optimization in first phase and community aggregation in second phase are the basic ideas of the Louvain method.

(a) Modularity (M)

Modularity tries to detect communities. Modularity measures relative of edges inside networks as for edges outside networks. Graphs with a high modularity score will have many connections within a community, but only few pointing outwards to other communities. Modularity is a scaled between -0.5 to 1 . In Fig. 2, first we indicate the colors in the image and then we aggregate them into their neighbor communities. After iterating through all the nodes, it is required merged few nodes together and form some communities. This becomes the new input for the algorithm. Modularity calculation is described in Eq. (1).

$$M = \frac{1}{2n} \sum_{x,y} [W_{xy} - \frac{z_x z_y}{2n}] \delta(k_x, k_y) \tag{1}$$

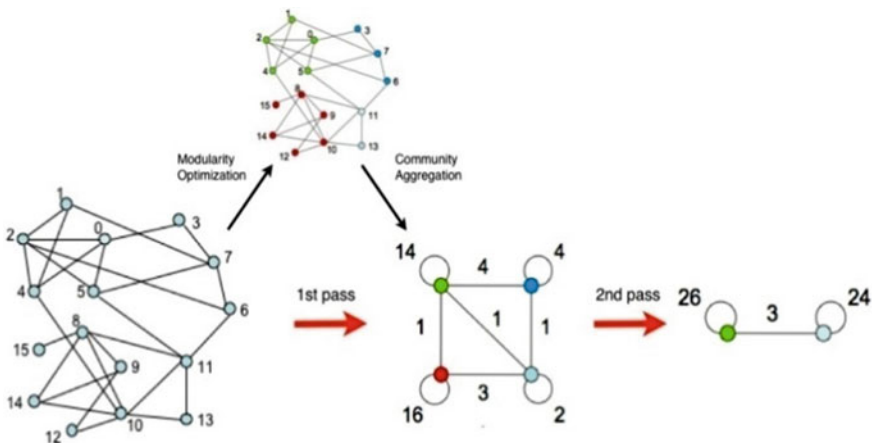


Fig. 2 Louvain method overview

where W_{xy} is the weight of the edge between x and y , z_x, z_y are the sum of weights of the vertices attached to the vertex X and vertex Y , also called as degree of the nodes. Variables k_x, k_y are the communities to which vertex X , vertex Y are assigned and n is the number of links.

(b) *Community Aggregation*

In community aggregation, small communities are established by optimizing modularity on all nodes, then individual community is arranged into one node and the repetition occurs. Community aggregation treats the networks with the same marker as a single node in the network and summates the waited adjacency matrix by summing over all the weights linking two communities. The procedure is repeated until there is no change in modularity increase caused by unifying any two communities.

3.2.2 Adaptive Over Segmentation Algorithm (Super Pixel Technique)

The process chart of the suggested adaptive over segmentation algorithm is shown in Fig. 3a. Fundamentally, when an input image is given, we employ the DWT to secure the coefficients of the high and low frequencies of the input picture. Then, the block size is computed to determine initial size of the block (S). Further, we compute simple linear iterative clustering algorithm where the computer S (initial size) to slice input picture to attain the IB. The suggested super pixel technique can lead to good results with compared to other forgery detection techniques as the equivalent time decreases the analytical expenses.

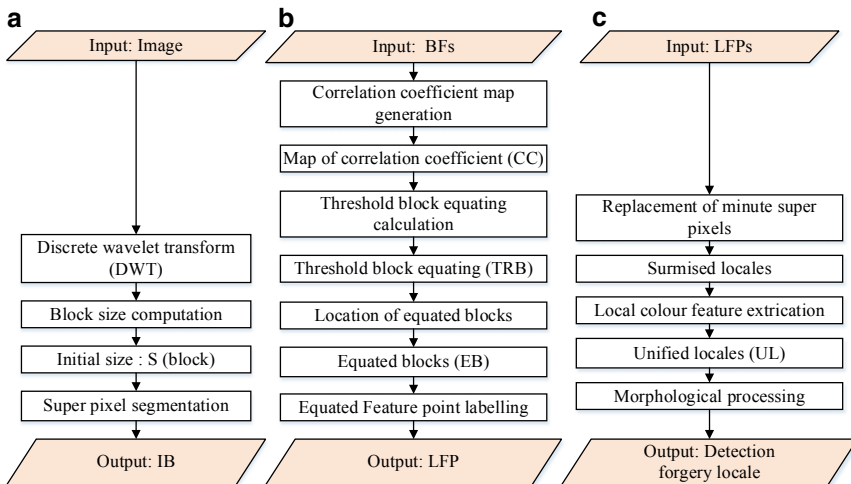


Fig. 3 Block diagram of **a** Adaptive over segmentation algorithm, **b** block feature equating algorithm, **c** forgery locale extrication algorithm

3.2.3 Block Feature Extrication Algorithm

Here, we extricated BFs from the individual blocks of the image. The standard block-based forgery exposing extricated similar features of the same length as the features of the block called BFs. This precisely uses the pixels of the picture blocks as the features of the block called BFs; yet the features of the image return the content of the image blocks by abandoning out the spot data. Therefore, we extricated features from individual picture block, and these points must be strong for different distortions. For example, distortions like rotation, compression of jpeg, and image scaling. Here, we used SIFT feature extrication approach to extricate the points called feature points from individual picture block. Hence, individual BFs contain asymmetrical block locale data and the extricated SIFT feature points.

3.2.4 Block Feature Equating Algorithm

It is required to individually spot the equated blocks after obtaining the BFs. As per existing block-based methods, domineering that they have the same shift vector, if there are many equating pairs in the same common position, the block equating progress achieves a certain block pair. The shift vector surpasses a user-identified threshold, and only when the blocks are equated can put up shift vector are identified as locales that are falsified.

3.2.5 Forgery Locale Extrication Algorithm

Forgery locales have the locations, which are extricated as the LFPs, and must still spot the forgery locales. As we know the input images are very well segmented by super pixels, replacing of the LFP with small super pixels is suggested to procure the surmised locales (SL) that are summation of labeled minute super pixels. Moreover, the neighbors of the SLs are the local color feature of the super pixels that are measured. If surmised locale is similar to the color feature, we unify the adjacent super pixels into communicating surmised locales that generates the unified locales (UL). Lastly, an operation called morphological is enforced to unified locales to recognize the forgery locales.

4 Experimental Results

The proposed strategy is first contrasted and the best in class interactive co-segmentation techniques. On earlier standard datasets to accomplish an equitable comparison, the suggested technique and the interactive co-segmentation technique use the similar scrawls in all experiments. In the analysis, we gather a variety of image groups from familiar image databases such as Microsoft dataset and ICOSEG

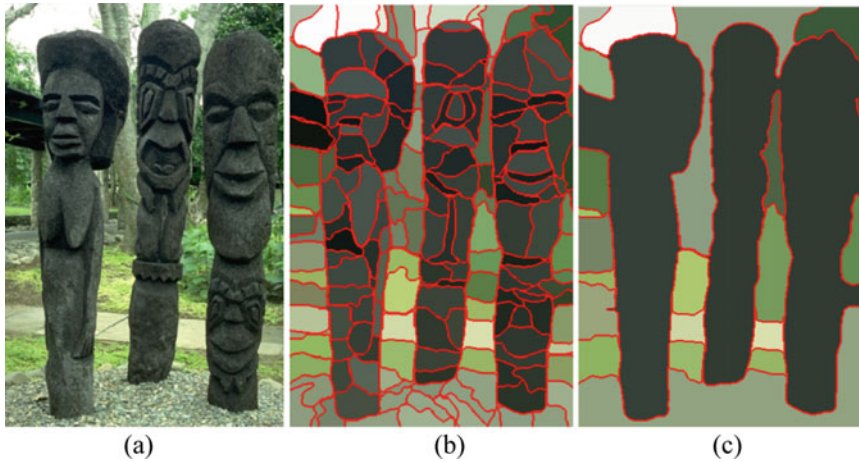


Fig. 4 a Original image, b super pixels image, c segmented image

dataset those two databases are very prominent for image co-segmentation analysis [8].

The experiments of forgery detection were carried out in Matlab (version R2015a). Matlab repeats wealthy toolbox can be quickly prototype an algorithm before it carries out the enlargement resources to executing the algorithm in another language like Java, C++ or Python. Linear system package (LINPACK) and Eigen system package (EISPACK) of the Matlab were also used in the process. Matlab furnishes four functions which permits to effortlessly produce basic matrices.

The experimental results obtained for the proposed process is depicted in Fig. 4. An original image, corresponding super pixels image, and the segmented image are depicted in Fig. 4a–c, respectively. Further, the results for detecting forgery in digital images are shown in Fig. 5. Figure 5a shows a forged image, and 5b highlights a patch detected by the proposed method.

For evaluating the effectiveness of the proposed work, we tested it on 100 original and same number of copy-move forged images. We measured the sensitivity, specificity, and accuracy of discrimination between the original and forged images. Table 1 presents the evaluation methodology of measuring metrics, and Table 2 presents the results. We can clearly observed that the proposed algorithm performed better than the DWT-based method explained in Sect. 2. It was also better than SIFT and SURF methods.

The DWT-based, SIFT-based, and SURF-based method do not take a comprehensive approach. Where as reported in Fig. 1, the proposed work is based on an exhaustive approach. Hence, the results were batter by the proposed method.

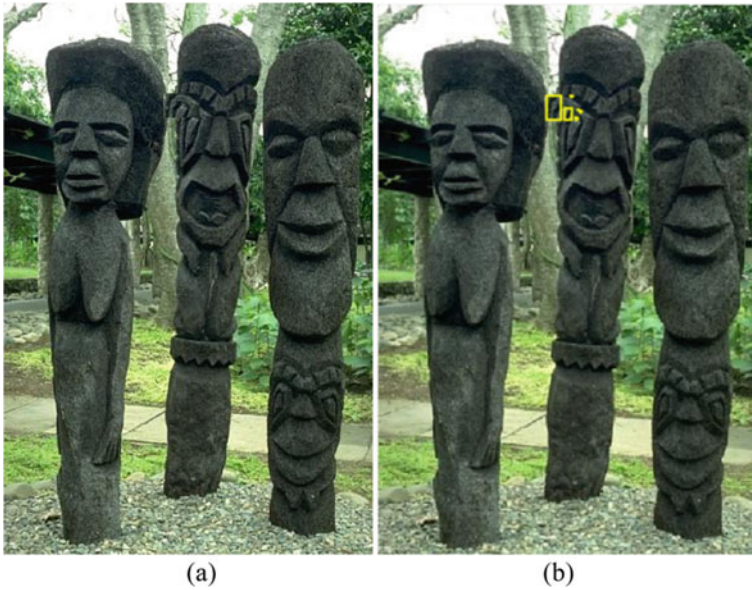


Fig. 5 a Forged image, b forged patch detected by the proposed algorithm

Table 1 Performance measurement parameters

Sensitivity	$\frac{TP}{TP+FN} \times 100\%$
Specificity	$\frac{TN}{TN+FP} \times 100\%$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN} \times 100\%$

TP true positive, true negative, false positive, false negative

Table 2 Comparative results

Method	Sensitivity	Specificity	Accuracy
DWT-based [2]	64	70	67
SIFT-based [9]	62	74	68
SURF-based [10]	68	78	73
Proposed	80	82	81

5 Conclusion

In this work, an automatic pixel equating is suggested with the use of forgery layer, updating mechanism and achieving copy-move detection results. Discovering of digital falsified images with copy-move operations are inspiring to recognize [11]. The paper focuses on recognizing falsified locales using feature point equating and adaptive over segmentation. Isolating picture into well separated, asymmetrical and

symmetrical blocks utilized by SLIC strategy. Extricating feature points from individual asymmetrical and symmetrical utilize SIFT and SURF strategies and feature equating algorithm detects the labeled feature points. A morphological strategy is applied in order to identify the exact tampered spot. Compared with SIFT, SURF technique abandon to recognize the falsified spot in the greater spot. Consequently, SIFT strategy offers good outcomes compared with SURF strategy. Our framework can shortly hold blocks of any shape size or distance from the image border lines. Even so, the limitation is that it fails for simply structured images. Apart from above mentioned limitations, one major issue of these detection techniques is the limited scope of utilization. In spite of burgeoning research in the field of image forgery detection, no detection method can be used as a solution for detecting all kind of forgeries. Thus, there is a high chance to expand a powerful, advanced forgery detection method that could remove previous drawbacks. Moreover, researchers may continue these algorithms to identify forgery in video clips.

References

1. G. Li, Q. Wu, D. Tu, S. Sun, A sorted neighborhood approach for detecting duplicated regions in image forgeries based on DWT and SVD, in *Proceedings on IEEE International Conference of Multimedia Expo*, July 2007, pp. 1750–1753
2. S. Bayram et al., An efficient and robust method for detecting copy move forgery, in *IEEE International Conference on Acoustics, Speech, and Signal Processing* (2009), pp. 1053–1056
3. Swaminathan et al., digital image forensics via intrinsic fingerprints. *IEEE Trans. Inf. Forensics Secur.* 1556–6013 (2008)
4. X.-C. Yuan, C.-M. Pun, X.-L. Bi, Image forgery detection using adaptive oversegmentation and feature point equating. *IEEE Trans. Inf. Forensics Secur.* 1705–1716 (2015)
5. B.P. Yadav, An automatic recognition of fake Indian paper currency note using matlab. *Int. J. Eng. Sci. Innov. Technol.* 3(4) (2014)
6. V.S. Vijayalakshmi, Comparative study of splicing based image forensic detection using KNN, fuzzy and SVM classifiers. Master's thesis, Visvesvaraya Technological University (2015)
7. R. Campigotto, A generalized and adaptive method for community detection. French National research agency (2014)
8. W. Wang, Higher order image co-segmentation. *IEEE Trans. Multimedia* 18(6) (2016)
9. I. Amerini, L. Ballan, R. Caldelli, A. Del Bimbo, G. Serra, A SIFT-based forensic method for copy-move attack detection and transformation recovery. *IEEE Trans. Inf. Forensics Secur.* 6(3), 1099–1110 (2011)
10. B.L. Shivakumar, S.S. Baboo, Detection of region duplication forgery in digital images using SURF. *IJCSI Int. J. Comput. Sci. Issues* 8(4, 1), 199–205 (2011)
11. X. Bo, W. Junwen, L. Guangjie, D. Yuewei, Image copy-move forgery detection based on SURF, in *2010 International Conference on Multimedia Information Networking and Security (MINES)* (2010), pp. 889–892
12. X. Pan, S. Ly, Region duplication detection using image feature matching. *IEEE Trans. Inf. Forensics Secur.* 5(4), 857–867 (2010)

A Granular Access-Based Blockchain System to Prevent Fraudulent Activities in Medical Health Records



Megha Jain, Dhiraj Pandey, and Krishna Kewal Sharma

Abstract Electronic health record (EHR) systems provide patient health information. EHR faces data security, integrity, and management challenges. Records can be modified by different stakeholder, as it can be used by different users in more than one form. Medical records management is an appropriate application for enabled blockchain records that can be stored, tracked, and managed. In the healthcare sector, a novel system has been presented here using the blockchain concept. Initially, the purpose of our proposed framework is to introduce blockchain innovation for EHR and, in addition, to provide safe electronic record functionality by characterizing granular access rules for the proposed system's customer. The presented framework has been deployed using blockchain technology and checked for its adaptability, safety, and other several necessary support in the system. So, the blockchain technology is used to create an electronic health record system that must be protected from manipulation and misuse.

Keywords Healthcare system decentralization · Ethereum · Consensus · Scalability · Blockchain · Smart contract · Electronic health record

1 Introduction

The electronic health record (EHR) is for the most part characterized to be the assortment of patients' electronic well-being data (e.g., as electronic clinical records—EMRs). EMRs can fill in as an information hotspot for EHR essentially from medical services suppliers in the clinical establishments. The patient health record contains individual medical care data, for example, those got from wearable gadgets claimed and constrained by patients. Data gathered as a feature of patient health record can be accessible to medical care suppliers, by clients (patients). Figure 1 shows medical record system which gives a large group of advantages to clinical practices [1].

M. Jain (✉) · D. Pandey · K. K. Sharma
JSS Academy of Technical Education, Noida, India

Sanskriti University, Mathura, India



Fig. 1 Medical record system [2]

The existing method of monitoring is facing a major corruption problem in several world as highlighted in several reports. Major components and system architecture of a typical blockchain scheme which is best suited to be deployed on any platform to help the EHR system for the transparency of records are explored here. In summary, a novel platform to support EHR system has been discussed here on blockchain technology, to prevent fraudulent activities in the system [2].

The paper is structured as follows. Section 2 explains blockchain technology in health care. Section 3 presents the related study of blockchain technology in the healthcare sector, and Sect. 4 discusses the proposed framework. The details of result have been discussed in Sect. 5. Section 6 discusses comparison with existing technology, and Sect. 7 concludes the whole approach as suggested.

2 Blockchain Technology with Health Care

To make the foundation frameworks straightforward, a systematic literature review has been done on many areas where blockchain technology is applied. EHR improves the quality of care, like fast admittance to quiet records and increment treatment adequacy, and suggests potential treatment choices, and gateways give patients admittance to their clinical subtleties, improve understanding doctor correspondence, and improve preventive consideration [3]. Figure 2 shows electronic health record system using blockchain. Patients whose information is re-appropriated or put away in these EHR frameworks for the most part lose control of their information and have no chance to get of realizing who is getting to their information and for what sort of purposes (e.g., alteration of individual protection).

In this article, we explain how blockchain technology could be used to stop corruption and secure our records. We proposed a framework to secure medical records and

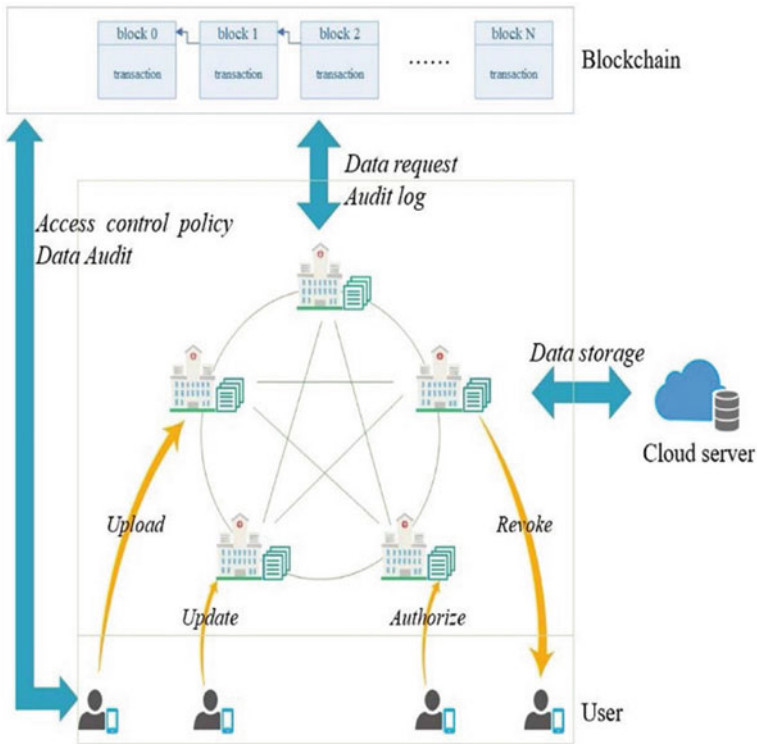


Fig. 2 Electronic health record system [2]

make it robust. Blockchain is a secure decentralized distributed platform to provide temper proof and shareable transactions. With this framework, a user would have the option to watch their records completely through a foundation to the recipient and past. When there is transparency, the user has to worry less about corruption [4].

3 Related Study

To make the foundation frameworks straightforward, a systematic literature review has been done in many areas. A summary of a detailed review of major encryption works in direction of various approaches toward healthcare management is presented below [5].

3.1 Healthcare Approach

Various approaches toward healthcare management have been suggested by researchers. Uddin et al. discuss that the increased usage of the Internet of Things (IoT) in the day-to-day life of human beings has also led to the concept of remote patient monitoring, and a patient-centric agent using the blockchain technology is commendable. Agbo et al. [6] worked extensively on the research paper, Blockchain Technology in Healthcare: Security, because the data of health care of a patient not only contains the records of his or her health and treatments but also personal information like contact address, contact number, and other sensitive information such as social security number. Matthias Mettler and MA [7] studied the different areas and various fields where blockchain can be used and how it can be used in other non-financial sectors. The major areas in which blockchain can be implemented successfully are the areas of smart healthcare systems, to fight the counterfeit drugs in pharmaceutical companies, for digitally signing of emails as well as contracts of legal purposes. Kumar et al. [8] discussed in the field of healthcare record securing and efficient data accessibility by using blockchain technology. Blockchain as a decentralized and circulated innovation can assume a key function in giving such medical care administrations. Tanwar et al. [9] proposed medical aid frameworks which are described as being exceptionally remarkable technology. The utilization of blockchain in medical aid frameworks assumes a basic part of the current medical services market. Fotiadis et al. [10] proposed BHEEM: A blockchain approach-based system for securing medical records in 2018. They have proposed a blockchain framework which gives secure data access to health records by patients, doctors, and outsiders while securing the patient private data.

3.2 Healthcare Information System

Electronic medical records (EMRs) meet the opposition to converse patient's privacy and share healthcare medical records among researchers when they need it. El-Yafouri and Klieb [11] discussed the software technology in medical sectors using electronic medical record adoption model, the healthcare information, and management system society model for considerate effect on organizations approving hospital information system and explained different types of levels of adoption [12]. Casino et al. [13] proposed that blockchain-based applications provide a logical writing survey over different areas which commit to research on blockchain and its applications and become more developed; their applications are required to enter more areas. Konstantinidis et al. [14] completed a survey on blockchain innovation generally called the mechanical premise on which bitcoin is made. This innovation has made elevated standards, as exchanges of every sort are executed in an extremely decentralized manner, without the necessity of an outsider. Jayasinghe et al. [15] use

cases and work may be a preliminary to limit various gaps by utilizing the combination information from use cases by monitoring the network. Kosba et al. [16] survey blockchain technology and summarize that it can be a progressive innovation which do a survey on blockchain which has great possibilities in settling problems with the transaction, improving scalability, upgrade security, and founded trust and protection.

4 Proposed Work

System design for electronic health record on blockchain technology has been suggested in this section. The main idea behind developing a secure and transparent platform using blockchain is to mainly make the existing EHR system to be more precise, accurate, secure, and transparent. Objectives have been set to develop a secure, decentralized, and distributed platform for medical records using blockchain. The goal is to create a system that is private, secured, confidential, compact, and free from any altering which uses blockchain for electronic health records [17].

Figure 3 depicts the system framework of blockchain-based medical service. The main activity is a standard blockchain exchange. The subsequent activity, addressed in ran blue lines, addresses interior exchanges. The third sort of activity, addressed in orange, is an eth call, which is utilized when information should be shipped off a keen agreement, yet should be kept in touch with the blockchain. The last sort of activity,

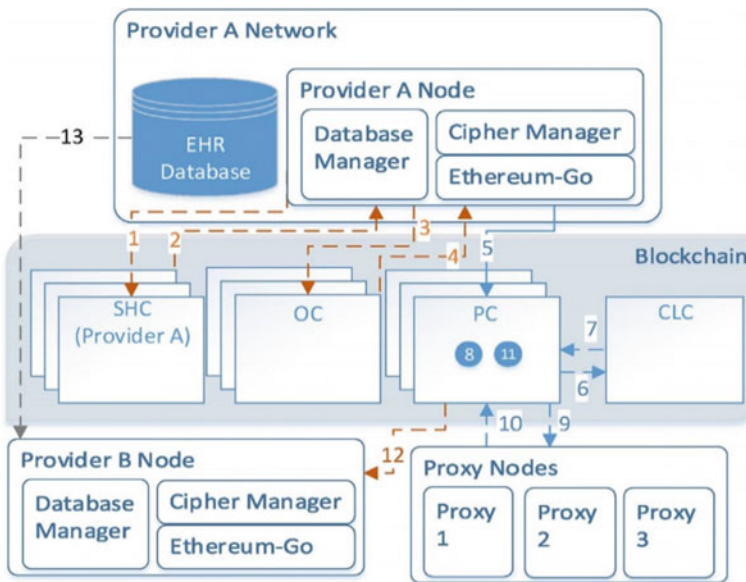


Fig. 3 System design of blockchain-based medical service framework

addressed in dim, is a non-blockchain activity. Steps for a patient accessing their record from electronic health record on blockchain technology have been described here [18].

Steps for a patient access their record

1. Node A sends the patient ID to patient service history records.
2. From service history record return to ownership address.
3. The node sends the filename of the requested record and Ethereum address of the patient to the ownership address.
4. The ownership records check to confirm that Ethereum address has permission.
5. If the node gets permission, their symmetric key is sent to the ownership records.
6. The ownership sends the encrypted symmetric key and database access information to the node.
7. The cipher manager decrypted the symmetric key using the private key of the node and then decrypted the query link with the symmetric key.
8. The database manager related to query link retrieves the encrypted document the from EHR database.
9. The cipher manager decrypts the record with the symmetry key.
10. The re-encrypted symmetry key is sent to permission records.
11. The permission records add the re-encrypted symmetric key to its database.
12. The permission record sends permission record address to node B.
13. Over HTTP, node A sent encrypted query link to node B. Node B decrypts the link and retrieves the record.

On-blockchain actions are steps 2, 3, 4, and 7, while off-blockchain actions are steps 1, 5, 6, 8, and 9. Hashing connections, update database, and sending notification are examples of off-blockchain behavior. The cost can be very low depending on how the off-blockchain modules are implemented [19].

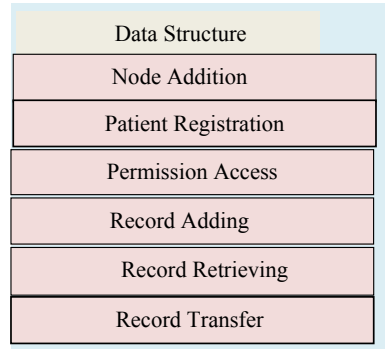
5 Implementation Details and Results

In this section, node addition, patient registration, permission access, record adding, record retrieving, record transfer, and data structure of our blockchain-based EHR sharing protocol are all described. The process of blockchain-based EHR system is shown in Fig. 4.

1. Node Addition

The way toward adding a hub starts by having citizen hubs approve that the public ID suits the mentioned characterization. Since adding a persistent has the least degree of consents on the blockchain and on the grounds that just a mathematical ID for patients is shipped off electors, patients will be added to the framework with little approval [20].

Fig. 4 Detailed structure of a blockchain-based EHR



2. Patient Registration

Enlisting a patient is one illustration of setting up a relationship between two distinct hubs. This cycle would be finished each time another patient visits a supplier. Utilizing the service history reports each time another relationship is framed. On the off chance that the patient is not a hub in the framework, the ‘Adding a Node’ cycle would be finished first. In the wake of affirming hub status, the supplier sends the appropriate data in an exchange to their service history [21].

3. Permission Access

The permission will at that point affirm that an adjustment in consent is conceivable prior to giving a solicitation to the patient. On the off chance that the patient did not have responsibility for record, those with possession would be sent the solicitation. The permission will at that point update its nearby information base and return a positive or negative notice to the supplier [22].

4. Record Adding

The cycle for adding a record starts with interior encryption in a supplier hub. It ought to be expected that the supplier and patient have effectively settled a relationship and have a common ownership. When a supplier hub makes another record, it will be moved to the database administrator, and a question connect to the EHR database will be made [23].

5. Record Retrieving

Recovering a record is a non-burdening measure on the grounds that no exchanges are required. The cycle starts by the patient finding the ownership for the supplier who stores the record. The patient at that point gives a demand for the record. In the event that the patient has consent to get to the record, the scrambled symmetric key is returned. When the patient decodes the key, they may unscramble the question connect they ought to have put away in their code manager, access the record in the supplier’s EHR data set, and unscramble the record [24].

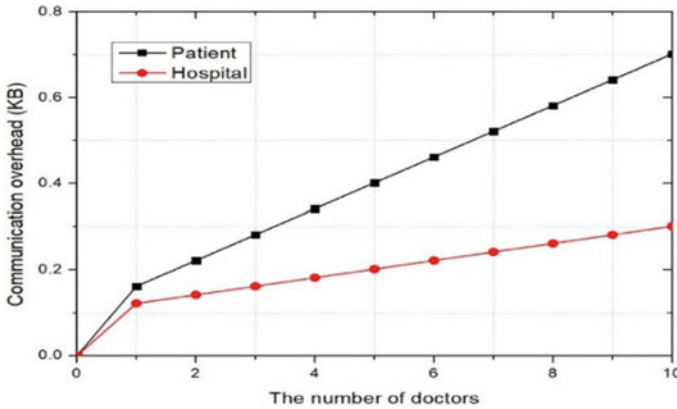


Fig. 5 Communication cost on the patient and hospital

6. Record Transfer

Smooth moving of records is vital for any EHR, the board framework. It utilizes intermediary re-encryption to adjust the requirement for availability while keeping up security portrays the interaction of one supplier sending a record to another. It ought to be noticed that the measure for moving a record could in fact happen by recovering a record, decoding, and shipping off another gathering [25].

Our scheme consists of six phases: node addition, patient registration, permission access, record adding, record retrieving, and record transfer. So, finally doctors can create EHR data and place the address of EHR data on the healthcare blockchain. When a doctor generates an EHR, they broadcast the data's address to the blockchain. Patients and other users in our scheme are verifiers. They can only access the data on the blockchain and verify the attributes. Figure 5 shows communication overhead on the patient and hospital. The hospital communication costs are divided into two categories. One is to schedule appointments with patients, and the other is to nominate doctors. The patient communication costs are divided into two categories. One is to make a hospital appointment, and the other is to delegate doctors. We present the communication overhead for the patient, hospital, and doctor, demonstrating TP-EHR entities with low communication costs.

Information security and privacy are fundamental priorities for the system. Blockchain with its properties of secure recording, anonymity, transparency, and distributed implementation will further strengthen the decentralized operation of proposed system. A multifaceted approach to security for our proposed EHR system includes mainly blockchain encryption and smart contracts. These smart contracts can be placed directly on the blockchain as transactions providing not only assurances of validity but an audit mechanism as well.

Table 1 Comparison of traditional methods with the proposed framework

Properties	Traditional method	Proposed framework
Structure	Centralized	Decentralized
Security	Less	High
Transparency	No	Yes
Risk	High	Low
Control	Not provide control to user	Provide control to user

6 Comparing Proposed Framework with Existing Framework

The conventional method used by traditional EHR system has a major drawback. It does not allow a user to have any control over the medical record. This results in lots of scams. The proposed method addresses this problem and provides user to control over system. A comparison between existing methods with the proposed method is shown in Table 1.

A novel framework based on the concept of blockchain will not only make the process faster but also will eliminate any chance of mismanagement or fraud. In the proposed scheme, the user can always track their record with the help of the unique hash attached to the block. This helps to solve the problem of transparency which was prevalent in several developed earlier management systems. Records reliability is also one of the major limitations of earlier work. The proposed system ledger is decentralized, so the manipulation is difficult and prevents corruption.

7 Conclusion

The design of the security of medical records for any EHR system during the dissemination of information between different parties is still a crucial problem. In this paper, a new extension has been designed to secure medical records using the blockchain approach. This new approach gives the insurance of integrity of all records and a robust EHR scheme. A framework has been designed to develop a secure, decentralized, and distributed platform for medical records using blockchain. This approach helps to monitor the lack in the transparency of medical professionals, and it allows entry to the EHR system only to the legitimate number of users. We focused on improving the security aspects of health records data, and in turn, it opens up new directions of other research to meet other security requirements. In future work, other security models that will be more generalized and cater to the need for health data security as well as identifying the untrusted hosts will be useful for critical application development for any medical purpose. This would help to save computing time

as well as help us to prevent cheating. Several new application areas to implement transparency toward patients as well as toward medical professional and allied areas will be benefited from the proposed work.

References

1. A. Azaria, A. Ekblaw, T. Vieira, A. Lippman, MedRec: Using blockchain for medical data access and permission management, in *Proceedings—2016 2nd International Conference on Open and Big Data, OBD 2016*, Sept 2016, pp. 25–30. <https://doi.org/10.1109/OBD.2016.11>
2. S. Cao, G. Zhang, P. Liu, X. Zhang, F. Neri, Cloud-assisted secure eHealth systems for tamper-proofing EHR via blockchain. *Inf. Sci. (Ny)*. **485**, 427–440 (2019). <https://doi.org/10.1016/j.ins.2019.02.038>
3. L. Chen, W.K. Lee, C.C. Chang, K.K.R. Choo, N. Zhang, Blockchain based searchable encryption for electronic health record sharing. *Futur. Gener. Comput. Syst.* **95**, 420–429 (2019). <https://doi.org/10.1016/j.future.2019.01.018>
4. Q. Xia, E.B. Sifah, A. Smahi, S. Amofa, X. Zhang, BBDS: Blockchain-based data sharing for electronic medical records in cloud environments. *Information* **8**(2) (2017). <https://doi.org/10.3390/info8020044>
5. R. Guo, H. Shi, Q. Zhao, D. Zheng, Secure attribute-based signature scheme with multiple authorities for blockchain in electronic health records systems. *IEEE Access* **6**, 11676–11686 (2018). <https://doi.org/10.1109/ACCESS.2018.2801266>
6. C. Agbo, Q. Mahmoud, J. Eklund, Blockchain technology in healthcare: a systematic review. *Healthcare* **7**(2), 56 (2019). <https://doi.org/10.3390/healthcare7020056>
7. R. Tonelli, in *IEEE International Conference on Software Analysis, IWBOSE'19 : 2019 IEEE 2nd International Workshop on Blockchain Oriented Software Engineering (IWBOSE'19)*, IEEE Computer Society, Institute of Electrical and Electronics Engineers, Hangzhou, China, 24 Feb 2019
8. T. Kumar, V. Ramani, I. Ahmad, A. Braeken, E. Harjula, M. Ylianttila, Blockchain utilization in healthcare: Key requirements and challenges, in *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)* (2018), pp. 1–7. <https://doi.org/10.1109/HealthCom.2018.8531136>
9. S. Tanwar, K. Parekh, R. Evans, Blockchain-based electronic healthcare record system for healthcare 4.0 applications. *J. Inf. Secur. Appl.* **50**, 102407 (2020). <https://doi.org/10.1016/j.jisa.2019.102407>
10. D.I. Fotiadis et al., in *Biomedical and Health Informatics and the Body Sensor Networks Conferences : 4–7 Mar 2018, Treasure Island Hotel—Las Vegas, Nevada, USA*
11. IEEE Computer Society et al., in *IEEE 2018 International Congress on Cybermatics ; 2018 IEEE Conferences on Internet of Things, Green Computing and Communications, Cyber, Physical and Social Computing, Smart Data, Blockchain, Computer and Information Technology : iThings/GreenCom/CPSCoM/SmartData/Blockchain/CIT 2018 : Proceedings : Halifax, Canada, 30 July–3 Aug 2018*
12. K.N. Griggs, O. Ossipova, C.P. Kohlhos, A.N. Baccarini, E.A. Howson, T. Hayajneh, Healthcare blockchain system using smart contracts for secure automated remote patient monitoring. *J. Med. Syst.* **42**(7) (2018). <https://doi.org/10.1007/s10916-018-0982-x>
13. F. Casino, T.K. Dasaklis, C. Patsakis, A systematic literature review of blockchain-based applications: current status, classification and open issues. *Telemat. Inf.* **36**(2018), 55–81 (2019). <https://doi.org/10.1016/j.tele.2018.11.006>
14. I. Konstantinidis, G. Siaminos, C. Timplalexis, P. Zervas, V. Peristeras, S. Decker, Blockchain for business applications: a systematic literature review. *Lecture Notes Bus. Inf. Process.* **320**, 384–399 (2018). https://doi.org/10.1007/978-3-319-93931-5_28

15. D. Jayasinghe, S. Cobourne, K. Markantonakis, R.N. Akram, K. Mayes, Philanthropy on the blockchain, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 10741 LNCS, pp. 25–38. https://doi.org/10.1007/978-3-319-93524-9_2
16. A. Kosba, A. Miller, E. Shi, Z. Wen, C. Papamanthou, Hawk: The blockchain model of cryptography and privacy-preserving smart contracts, in *Proceedings—2016 IEEE Symposium on Security and Privacy, SP 2016*, Aug 2016, pp. 839–858. <https://doi.org/10.1109/SP.2016.55>
17. A. Al Omar, M.S. Rahman, A. Basu, S. Kiyomoto, MediBchain: A blockchain based privacy preserving platform for healthcare data, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017, vol. 10658 LNCS, pp. 534–543. https://doi.org/10.1007/978-3-319-72395-2_49
18. A. Shahnaz, U. Qamar, A. Khalid, Using blockchain for electronic health records. *IEEE Access* 7, 147782–147795 (2019). <https://doi.org/10.1109/ACCESS.2019.2946373>
19. N. Rifi, E. Rachkidi, N. Agoulmine, N.C. Taher, Towards using blockchain technology for IoT data access protection
20. H. Zhao, Y. Zhang, Y. Peng, R. Xu, Lightweight backup and efficient recovery scheme for health blockchain keys, in *Proceedings—2017 IEEE 13th International Symposium on Autonomous Decentralized Systems, ISADS 2017*, May 2017, pp. 229–234. <https://doi.org/10.1109/ISADS.2017.22>
21. X. Yue, H. Wang, D. Jin, M. Li, W. Jiang, Healthcare data gateways: found healthcare intelligence on blockchain with novel privacy risk control. *J. Med. Syst.* 40(10) (2016). <https://doi.org/10.1007/s10916-016-0574-6>
22. W. Wang et al., A survey on consensus mechanisms and mining strategy management in blockchain networks. *IEEE Access* 7, 22328–22370 (2019). <https://doi.org/10.1109/ACCESS.2019.2896108>
23. D.C. Nguyen, P.N. Pathirana, M. Ding, A. Seneviratne, Blockchain for secure EHRs sharing of mobile cloud based e-health systems. *IEEE Access* 7, 66792–66806 (2019). <https://doi.org/10.1109/ACCESS.2019.2917555>
24. P. Zhang, D.C. Schmidt, J. White, G. Lenz, Blockchain technology use cases in healthcare, in *Advances in Computers*, vol. 111 (Academic Press Inc., 2018), pp. 1–41
25. G. Yang, C. Li, A design of blockchain-based architecture for the security of electronic health record (EHR) systems, in *Proceedings of the International Conference on Cloud Computing Technology and Science, CloudCom*, Dec 2018, vol. 2018, pp. 261–265. <https://doi.org/10.1109/CloudCom2018.2018.00058>

Suspicious Activity Detection in Surveillance Applications Using Slow-Fast Convolutional Neural Network



Mitushi Agarwal, Priyanka Parashar, Aradhya Mathur, Khushi Utkarsh,
and Adwitiya Sinha

Abstract With the expansion in crime and abnormal human activities, extensive need has arisen to upgrade the cyber-physical security mechanism with artificial intelligence. Our research aims at proposing a model that can be installed in the CCTVs in any area and inform about the human activities happening in that particular area. We have developed a deep learning model, using a variant of slow-fast algorithm to detect and classify the unusual activity happening in the areas, which becomes tough and tedious with manual effort, thereby engaging large workforce. Our solution approach focuses on alleviating the problem using convolutional neural network. Our model uses a slow pathway and a fast pathway, wherein for analyzing stable or static content of the video, a low temporal resolution, with slow pathway is used. Also, another high temporal resolution with fast pathway is used for analyzing the active or dynamic data of video. These two pathways are combined by lateral connections to deliver global detection of the video. The experimental results show effectiveness of our method in detecting unusual movement and identification of suspicious objects in video recordings that may include both real-time live streams and prerecorded videos.

Keywords Cyber-physical security · Artificial intelligence · Deep learning · Slow-fast CNN · Classifier · Real-world surveillance

1 Introduction

There is a huge rise in the number of strange events taking place these days. This necessitates security to upgrade in real-world organizations, which require a constant need to monitor people and their interactions [1]. However, this eventually requires heavy engagement of manual workforce. Therefore, the challenge that comes forward is the demand for an automatic and intelligent surveillance system that helps in the analysis of such videos.

M. Agarwal · P. Parashar · A. Mathur · K. Utkarsh · A. Sinha (✉)

Department of Computer Science and Engineering and Information Technology, Jaypee Institute of Information Technology, Noida-62, Uttar Pradesh, India

Numerous associations have introduced closed-circuit televisions (CCTVs) for the consistent checking of individuals and their collaborations. A large portion of the memory spaces of the business are involved in big data. The execution of CCTV cameras in all areas because of security purposes and utilization of CCTV cameras is basic; however, it devours more memory spaces to store data. Security is utilized for robbery recognizable proof, viciousness recognition, unapproved people entering, and criminal behavior in a locale. Thus, for all unusual movement's security assumes a significant job, so security must be actualized in the area of more privacy. Utilizing CCTV videos in the past day's strategy to discover the robbery happenings and different exercises, this is a dreary cycle and tedious job. The crowd may emerge in occupied roads, games, music shows, and fights, among others. Although individuals in a crowd regularly move in an organized way, little unsettling influences may prompt a frenzy circumstance and perhaps grievous results.

As technology grows faster, crime rates and strategies also continue [2, 3]. One of the biggest crimes facing almost the entire world is street crime and theft. However, there is no wise way to identify or find something or someone. A common way to watch is by observing different long videos carefully from each one of the CCTV recordings. It is tough to find unusual activities using this CCTV footage. Image quality, movement, and objects are detected by CCTV cameras. The basic challenge of the CCTV surveillance module is to automatically and intelligently study videos and detect unusual and unpleasant incidents in high-speed areas and aid better prevention of people in that area.

Therefore, our project comes forward as an attempt to provide a solution to such a problem as it is a smart surveillance model that can detect unusual activity automatically. The model can be installed in a CCTV camera for real-time surveillance and detect any suspicious activity that occurred in that particular environment. The main objective of this model is to detect unusual activity in video surveillance for identification of unusual activities in surveillance systems where unusual activities could be abuse, arrest, assault, burglary, explosion, road accident, shoplifting, and other such related applications.

The slow-fast network follows divide and conquer where each pathway leverages its strength in video modeling. One pathway processes the clips at slow rate, while the other fast pathway operates the same raw video clips, but at much higher frame rate. Such speed allows better understanding of various kinds of movements in the video. The main benefit of the approach comes by the efficiency gained by reducing the channel capacity of the fast pathway which boosts its temporal modeling ability and makes it lightweight and cost-effective. Our method doesn't compute optical flow and learns from end-to-end from raw data which makes the slow-fast network empirically more effective. Therefore, the overall system results with less computational complexity and higher accuracy than any other compute-heavy approaches.

2 Related Work

There are several ongoing researches in the area of automated real-time detection of unusual activities. The authors in [1] highlighted an assortment of circumstances, for example, public shows and matches for analyzing unusual events. In ordinary conditions, the mass moves in an efficient way, however, alarm circumstances may result in calamitous outcomes. They proposed a computer vision technique to recognize movement design changes in human groups which can be identified with an uncommon occasion. The novel method can distinguish worldwide changes, by assessing 2D movement histograms in time, and neighborhood impacts, by recognizing bunches that present comparable spatial areas and speed vectors. Xiao in [4] presented a united slow-fast audio-visual network. It is an extended model of slow-fast model with integrated audio and visual both and performs recognition.

Feichtenhofer in [5] presents a model for video recognition, namely slow-fast model. The model involves two things, firstly pathway for slow, secondly pathway for fast. By reducing its storing capacity, the slow pathway can be made light; still, it can be used for video recognition by learning data. In another research in [6] conducted research to spot street crime, the amount of moving target simulates objective overall performance, speed, and volume of video speed. The authors used is Snatch 1.01 dataset; for preprocessing, the dataset is divided into frames, and then, features are extracted for this purpose; the VGG19 algorithm had been applied in this paper, and it is concluded that the results of the algorithm contrast the features from the original video. The proposed system surpass the state-of-the-art systems with 81% accuracy and 0.025 frames per second detection time.

Ahir in [7] states that the papers focal point is on a deep learning approach to detect suspicious activities using convolutional neural networks (CNN) from images and videos. They studied the previous approaches present and offered an alternative approach to detect suspicious activities happening in public places. They used CNN for finding if the activity was suspicious. The ResNet architecture was used to build the CNN model. They tried ResNet-18, ResNet-34, and ResNet-50 approaches. According to their research, they concluded that the ResNet-50 works the best for the task. Al-Zawi in [8] explains and defines all the concepts and problems related to CNN, and how it works also they will explain how to state the parameters that affect CNN effectiveness. The supreme layer in CNN is the convolution layer, and it takes the maximum time in the network. The network performance also depends on the number of levels within the network. However, the number of levels increases the time required to train and test the network. The authors have illustrated the power of CNN in applications, like face detection and image, video recognition, and voice recognition.

Amudha in [9] has proposed a model which can stop the crime before it happens. The real-time CCTV footage is tracked and analyzed. The dataset used are CAVIAR dataset, KTH dataset, some YouTube videos, and some videos taken from their campus. The dataset has been trained on the LSTM model. The accuracy they could achieve is 76% for 20 epochs and could be increased if trained for more iterations. In

yet, another research conducted in [10] includes an in-depth study that begins with object recognition, action recognition, crowd analysis, and finally the discovery of violence in a crowded area. The paper discusses in-depth learning start-up technology that is involved in a variety of video analytics methods. The techniques like YOLO, SVAS, IBSTM, KLT, GMM, SVM, and many other techniques used for video and crowd analysis are explained in detail. In [11], Zheng performs an exhaustive review of deep learning-based discovery frameworks. It gave a brief about CNN then focuses on typical generic object detection architectures and some useful pointers to improve the accuracy of the detection. The research provided a review of a comprehensive learning-based discovery framework dealing with a few different problems, such as occlusion, clutter, and low adjustment, with varying degrees of conversion on RCNN.

Peng in [12] explore explicitly current advances in the field of acquisition and classification based on computer vision, along with a comparison of these methods. The object detection methods associated with color or shape like color index, SVM, neural network, and Hough transform are studied, and their advantages and disadvantages are given. Krizhevsky in [13] trained a large, deep convolutional neural network to categorize the images in the ImageNet LSVRC-2010 into the 1000 different classes for better performance in object recognition. They concluded that a large, using only supervised learning, D-CNN is capable of achieving great results on a challenging dataset. In another work [14], they have proposed the spatiotemporal slow-fast self-attention network for action recognition. They have proposed a module that can extract four features in video information: spatial information, temporal information, slow action information, and fast action information using the self-attention mechanism GAN (SAGAN). They have used AVA dataset and have shown frame AP improvement in 28 categories. Their network is based on a fast RCNN algorithm. Finally, motivated from the existing research, we have implemented a CNN-based slow-fast model for abnormal activities detection to enhance the performance of video surveillance.

3 Proposed Model

There is a huge increase in unusual activities like robbery, accidents, and assaults, and hence, the need of the hour is to increase the security in our areas. In order to detect the anomalies in the people's behavior, the cameras need to be monitored constantly which requires a large workforce and constant attention. Therefore, it creates a need to develop such a system that can automatically detect the same without the manual labor. Hence, we need to accurately detect the unusual activities in our environment that includes both homes and workplaces. Our method involves the implementation of slow-fast model in surveillance systems motion that improves action classification and action detection by simultaneously extracting information from video at both slow and fast frame rates (Fig. 1).

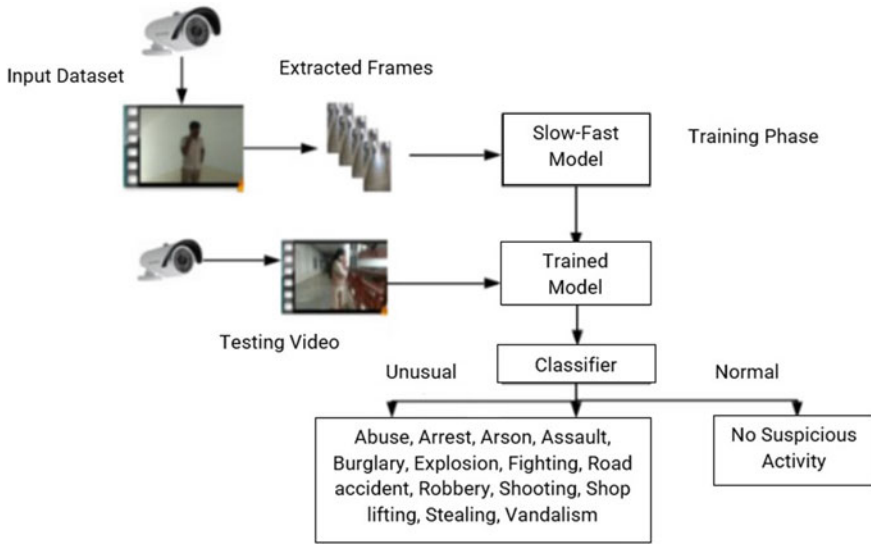


Fig. 1 Flow diagram for the proposed model using slow-fast convolutional networks

The model focuses on two pathways with one focusing on extracting the spatiotemporal features using the slow pathway, and the other pathway detects the rapidly changing motion characteristics thus detecting each and every activity happening in the area accurately.

3.1 Dataset Description

DCSASS dataset is used for the evaluation of the model. It is a standard dataset. It contains surveillance camera’s videos that contains both abnormal and normal behaviors. The videos are segmented in 13 labels that includes like assault, road accidents, burglary, robbery, shoplifting, explosion, stealing, arrest, vandalism, arson, fighting, abuse, and shooting [15]. Each video is marked with two labels based on the type of contents, including normal (as 0) and abnormal (as 1). Finally, the distribution of this database is highlighted in Figure 2. It contains a total 16,853 videos, wherein 9676 are labeled as normal videos, and 7177 are labeled as unusual. The dataset distribution for different labels is illustrated in Figure 3.



Fig. 2 Dataset with multiple labels describing different unusual activities

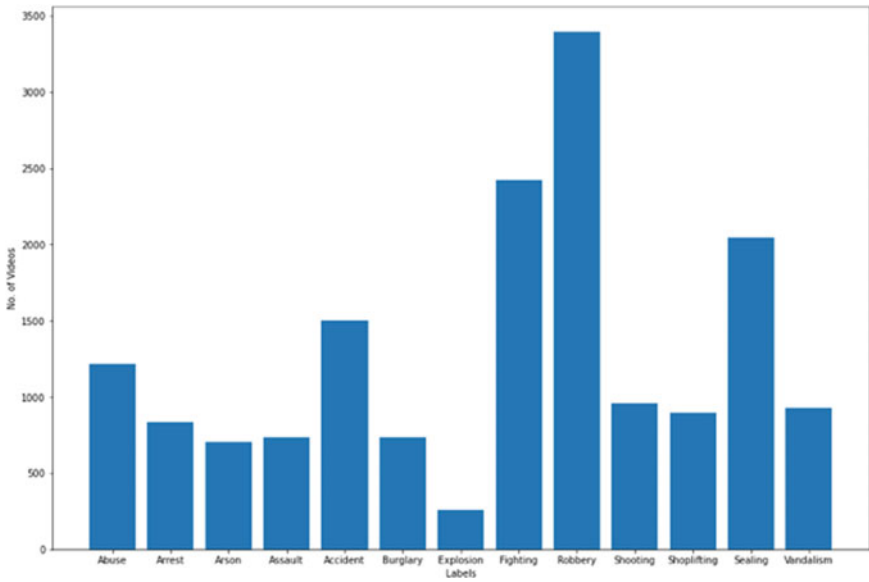


Fig. 3 Dataset distribution for different unusual activities

3.2 Data Exploration and Preprocessing

The unusual activity detection system for video surveillance is an intelligent open computer vision-based surveillance system. For the implementation of this automated model effectively, we have surveyed various research papers in order to come forward with most accurate and efficient results. The following is the solution approach for the developed model. The preprocessing of data is done as follows:

- Step 1: involves data cleaning where the dataset which contains 9676 normal videos and 7177 as abnormal videos needs to be sorted firstly. So, we clean the

data by sorting the videos according to their labels into normal and abnormal categories and discarding all the duplicate files.

- Step 2: involves splitting the video files into sequence of frames. As the video files are nothing but a collection of a set of images. This set of images are known as frames which are further combined to obtain the original video.

The process of image classification is done by extracting features using feature extractors like CNN from the images and further classify them according to the extracted features. While in the video classification, an extra step is included, wherein, firstly, frames are extracted from the video, and then, the same procedure is followed as that of image processing.

3.3 Training Using Slow-Fast CNN

The preprocessed dataset now undergoes the training phase which has been performed using the slow-fast algorithm. While dealing with video scenes, one thing to remember is that it usually contains information in two distinct parts:

- Firstly, it includes the static areas, wherein the either the frame remains the same or changes very slowly.
- This is followed by rapidly moving dynamic areas that indicate the moment which is in progress.

Video classification is done typically using the 3D convolutional neural networks which consists of several layers, and after each stage, the receptive field on the filter of input increases; it extends the 2D image-based network to 3D. An extended version of this algorithm is used in video recognition that improves action detection and classification by simultaneously extracting the information from both slow and fast frame rates from the video. This model is known as slow-fast that uses two pathways:

- Slow Pathway: This focuses on processing the spatial temporal features such as colors, textures, and objects that can be viewed at low frame rates.
- Fast Pathway: It focuses on the rapidly changing movements that are often easily recognized in videos that operate at higher frame rate.

In the slow-fast model, both the slow and fast pathways use a 3D ResNet model which captures many frames at a time and runs 3D convolutional operations on the pathways. For analyzing the stable or static content of the video, slow-fast uses a low temporal resolution, slow CNN, i.e., slow pathway. While side by side another high temporal resolution, fast CNN, i.e., fast pathway is running which is used for analyzing the active or dynamic data of the video. These two pathways are combined by lateral connections. The slow-fast network follows divide and conquer where each pathway leverages its strength in video modeling. One pathway processes the clips at slow rate, while the other fast pathway operates the same raw video clips, but at much higher frame rate. Such speed allows better understanding of various kinds of movements in the video. The main benefit of the approach comes

by the efficiency gained by reducing the channel capacity of the fast pathway which boosts its temporal modeling ability and makes it lightweight and cost-effective. Our method doesn't compute optical flow and learns from end-to-end from raw data which makes the slow-fast network empirically more effective. Therefore, the overall system results with less computational complexity and higher accuracy than any other compute-heavy approaches.

4 Experimental Results and Discussion

We have performed the comparative analysis between the RNN model and the slow-fast model to obtain the accurate and efficient results. In the first half, we have applied RCNN to our dataset to obtain the results.

After training, we have obtained the accuracy of 97.46% of the first layer and validation-accuracy of 11.9%. This is evident from the trend plotted between the accuracy and validation accuracy in the graph illustrated as Fig. 4. After training with fourth layer, we can see that the accuracy rises at a constant rate with the accuracy of 96.19% and Val-accuracy of 19.05%. We can see the graph between the accuracy and Val-accuracy of layer 4 in Figure 5. In the second half, we have applied the slow-fast model to obtain better results on our dataset after RCNN. Figure 6 shows the model accuracy graph plotted between accuracy and Val-accuracy. After training, we have obtained the accuracy of 98.39% and Val-accuracy of 85.02% with loss of 5.29%, and Val-loss of 0.63% of Figure 7 shows the model loss graph plotted between loss and Val-loss.

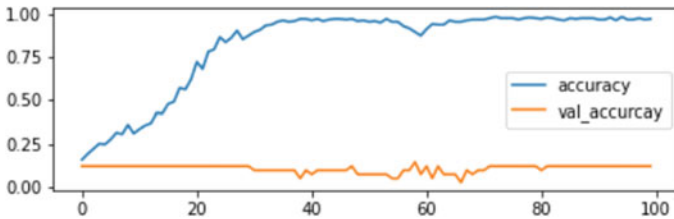


Fig. 4 Model accuracy of RCNN model in the first layer

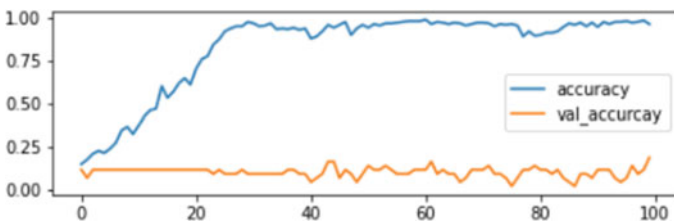


Fig. 5 Model accuracy of RCNN model in the fourth layer

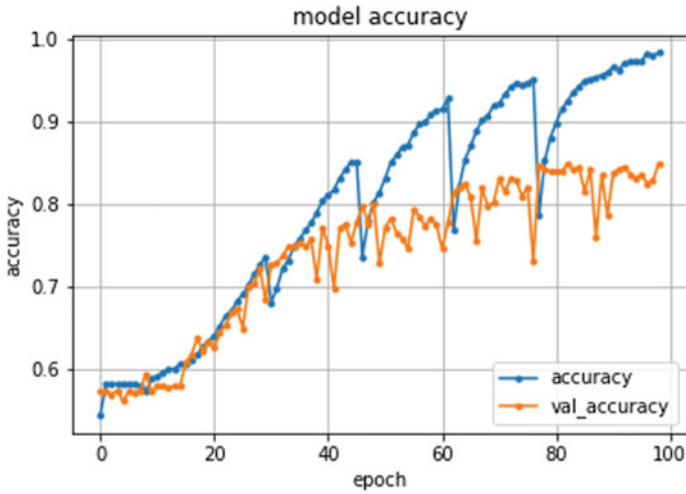


Fig. 6 Model accuracy of slow-fast CNN model

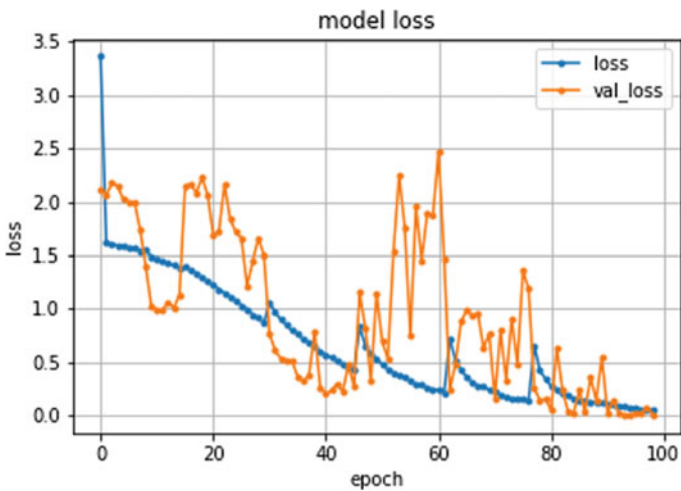


Fig. 7 Model loss of slow-fast CNN model

After performing the comparative analysis between RCNN and slow-fast model, we can clearly see that SlowFast performs better. Therefore, we have applied SlowFast algorithm to our model for testing the videos. Shown below in Figures 8 and 9 are the few results that were obtained after passing two real-time CCTV video footages from the SlowFast model. The two activities that were detected includes road accidents and assault in Figures 8 and 9, respectively.



Fig. 8 Prediction result: road accidents



Fig. 9 Prediction result: assault

5 Conclusion and Future Scope

We have presented an unusual activity detection method using the slow-fast CNN-based approach. The slow-fast presents a completely unique and new approach for video detection and classification, taking advantage of the structure of real-world scenes; this approach works on two pathways, namely a slow pathway that takes into consideration of the spatiotemporal features and the fast pathway taking into the consideration of only the motion characteristics, thus detecting each and every minute action possibly happening in the scene. Thus, a fast and intelligent method to check these surveillance cameras is at most required. It would help in cutting down a lot of work to be done by people struggling to monitor it and would help it taking faster actions during those situations by integrating these with alarms and other

important actions like informing the police. After model training, we have obtained the accuracy of 98.39% and validation-accuracy of 85.02% with loss of 5.29% and validation loss of 0.63%. Since, the model has two pathways, and by operation on the fast pathway with a higher frame rate, being very lightweight, becomes cost-effective. This provides efficient results for action classification and detection for video sequences that can be further used in CCTV surveillance.

The model can be further used to fulfill the need of the current open challenge by integrating it with the audio interface. For instance, a woman running and shouting for help in an area could be detected as unusual activity and immediately assisted. Model can also be integrated with the facial recognition system and can be helped in criminal identification. The model is environment specific, which becomes as one of its limitations, and therefore, expanding the dataset would further help in minimizing the limitation.

References

1. R. Mehran, A. Oyama, M. Shah, in *Abnormal Crowd Behavior Detection Using Social Force Model*, 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE (2009, June), pp. 935–942
2. S. Gera, A. Sinha, A machine learning-based malicious bot detection framework for trend-centric twitter stream. *J. Discrete Math. Sci. Crypt.* (2021), pp. 1–10
3. P. Kumar, A. Sinha, Information Diffusion Modeling & Analysis for socially interacting networks. *Social Network Analysis & Mining*, Springer **11**(11), 1–18 (2021)
4. F. Xiao, Y.J. Lee, K. Grauman, J. Malik, C. Feichtenhofer, Audiovisual slowfast networks for video recognition (2020). arXiv preprint [arXiv:2001.08740](https://arxiv.org/abs/2001.08740)
5. C. Feichtenhofer, H. Fan, J. Malik, K. He, in *Slowfast Networks for Video Recognition*. Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6202–6211
6. S. Letchmunan, F.H. Hassan, S. Zia, A. Baqir, Detecting Video Surveillance Using VGG19 Convolutional Neural Networks
7. R. Gugale, A. Shendkar, A. Chamadia, S. Patra, D. Ahir, Human Suspicious Activity Detection using Deep Learning (2008)
8. S. Albawi, T.A. Mohammed, S. Al-Zawi, in *Understanding of a Convolutional Neural Network*. 2017 IEEE International Conference on Engineering and Technology (2017, August), pp. 1–6
9. C.V. Amrutha, C. Jyotsna, J. Amudha, in *Deep Learning Approach for Suspicious Activity Detection from Surveillance Video*. In 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) (2020, March). IEEE, pp. 335–339
10. G. Sreenu, M.S. Durai, Intelligent video surveillance: a review through deep learning techniques for crowd analysis. *J. Big Data* **6**(1), 1–27 (2019)
11. Z.Q. Zhao, P. Zheng, S.T. Xu, X. Wu, Object detection with deep learning: a review. *IEEE Trans. Neural Networks Learn. Syst.* **30**(11), 3212–3232 (2019)
12. J. Wu, B. Peng, Z. Huang, J. Xie, in *Research on Computer Vision-Based Object Detection and Classification*. International Conference on Computer and Computing Technologies in Agriculture (Springer, Berlin, Heidelberg, 2012, October), pp. 183–188
13. A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2012)

14. M. Kim, T. Kim, D. Kim, in *Spatio-Temporal Slowfast Self-Attention Network for Action Recognition*. 2020 IEEE International Conference on Image Processing (ICIP) (2020, October). IEEE, pp. 2206–2210
15. W. Sultani, C. Chen, M. Shah, in *Real-World Anomaly Detection in Surveillance Videos*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018), pp. 6479–6488

Licence Plate Recognition System for Intelligence Transportation Using BR-CNN



Anmol Pattanaik and Rakesh Chandra Balabantaray

Abstract Automatic licence plate recognition (ALPR) is useful in a variety of applications, and several methods have been suggested for the same. Light, shadow, background complexity, vehicle speed, and other variables can easily affect the conventional location recognition algorithm, resulting in failure to meet the application of real scenes. The use of deep learning (DL) accelerates the licence plate recognition algorithm and also make it capable of extracting more intricate details, greatly increasing detection and recognition accuracy. In this paper, a novel technique for robust recognition of licence plates (LPs) is presented. The proposed novel deep learning-based automatic licence plate recognition model works with bounding rectangle-based (BR) segmentation and convolutional neural network-based (CNN) recognition called BR-CNN model to enhance ALPR accuracy. The proposed BR-CNN model works on three primary stages to identify and recognise licence plate number, namely licence plate detection, bounding rectangle segmentation, and recognition of licence plate characters with CNN. During the preliminary phase, the LP detection process is carried out with the help of a connected component analysis (CCA) model. The LP image is then segmented using the bounding rectangle technique. Finally, with the aid of the CNN model, the characters in the LP are recognised. The obtained outcome simulation result ensures that BR-CNN model outperformed the other models.

1 Introduction

In the field of intelligent transport system (ITS) and graphical processing unit (GPU), licence plate recognition (LPR) that does not involve human intervention or control is crucial [1]. With the rise in car usage comes the emergence of new problems in the

A. Pattanaik (✉) · R. C. Balabantaray
BBSRInternational Institute of Information Technology, Bhubaneswar, Odisha, India
e-mail: C119001@iiit-bh.ac.in

R. C. Balabantaray
e-mail: rakesh@iiit-bh.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022
P. Verma et al. (eds.), *Advances in Data Computing, Communication and Security*,
Lecture Notes on Data Engineering and Communications Technologies 106,
https://doi.org/10.1007/978-981-16-8403-6_60

659

car process, such as car theft, traffic accidents, road congestion, severe environmental pollution and so on. This is because ALPR has a wide range of uses, including parking lot control, restricted area protection, traffic law enforcement, congestion pricing and automatic toll collection. LPR methods differ depending on the working situation. Most current algorithms, on the other hand, only operate well under managed conditions or with sophisticated image capture systems. As a result, many strategies, such as permanent lighting, low vehicle speed, assigned routes and a static context, can be enforced with strict laws. Despite the fact that two decades have been spent developing LPR systems to meet the needs of different contexts, there are still a number of obstacles to overcome in order to achieve high detection and recognition rates. So, the question of how to enhance the accuracy and recognition rate of licence plate detection in a dynamic setting is of great interest to researchers.

Numerous researchers have begun to concentrate on LPR, which is concerned with the localisation, segmentation and recognition of LPs [2]. Thus, effective placement of the LP system requires meticulous attention to detail, whereas a single individual must perform a task in a coordinated manner throughout prolonged dissection of a single component. This paper introduces the BR-CNN model, a DL-based ALPR model that uses bounding rectangle-based segmentation and CNN-based recognition. Our study's primary contributions are summarised below:

1. Implementation of Harr Cascade and CCA model for localising and detecting LPs.
2. Using the bounding rectangle segmentation method, demonstrate a strategy for segmenting LP images.
3. Finally, employ a CNN model to perform LP character recognition.
4. Validate the proposed model's performance on the test dataset.

The remainder of the paper is laid out as follows: Section 2 provides an overview of related work. Section 3 summarises the integrated model and provides an overview of each component. Section 4 continues with experimental verifications, and Section 5 contains conclusion.

2 Related Work

Numerous algorithms have been suggested for plate detection. Some of these algorithms work by looking for image edges, such as horizontal and vertical edges [3–5] use the Canny edge detector to localise plates. Other algorithms detect plates by locating their boundaries utilising the Hough transform [6, 7], which is a time-intensive method that requires a large amount of memory [8]. Such a tool is incapable of detecting plates with no discernible boundaries. Additionally, plate detection has been performed using wavelet analysis [9, 10]. Wavelet-based methods detect plate candidates by using high-frequency coefficients. Because these coefficients correlate to edges, they share many of the limitations of edge detection techniques. Certain

detection algorithms combine mathematical morphology with connected component analysis [11, 12]. For plate detection, these algorithms require images with a moderate to high contrast. Character segmentation has long been a large focus for research due to the abundance of techniques based on morphological operations [13] and connected component analysis (CCA) [14, 15]. To get a binary picture of the plate in such ways, a suitable thresholding strategy must be applied prior to any further processing. Plate binarization can be achieved using thresholding methods such as Niblack [16], SAUVOLA [17], Wolf and Jolion [18] and OTSU [19]. Many different classification tools and techniques have been used for character recognition in the past, including artificial neural networks (ANN), support vector machines (SVM), Bayes classifier, k-nearest neighbour and so on. Feature detection is performed to supply classifiers with the data they need. Many methods have been suggested for feature extraction, including character skeleton [20], active areas [21], HOG [22], horizontal and vertical projection [23] and the multiclass AdaBoost approach [24]. The character recognition system in some methods, such as SIFT [25] and SURF [26], is focussed on key point's localisation.

3 The Proposed BR-CNN Model

This section explains how our proposed ALPR solution works. Earlier stages of the process involve the localisation and recognition of licence plates using the Haar cascade and CCA models. The characters found in the LP are then split using the bounding rectangle algorithm, which involves separating the characters on the licence plate. Finally, the CNN-based character recognition algorithm is used to identify the characters in LP. Figure 1 shows a visual representation of the BR-CNN model's core functioning principle.

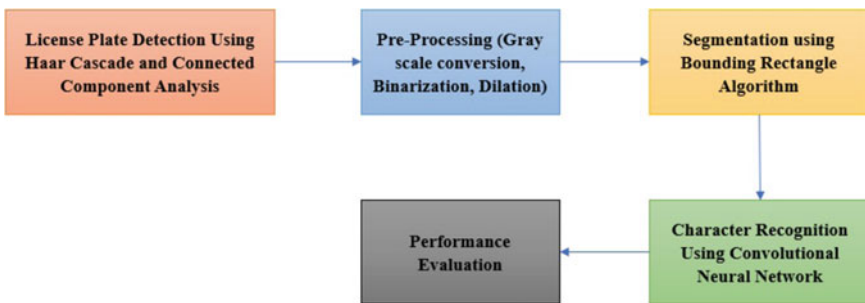


Fig. 1 Overall working principle of proposed BR-CNN model

3.1 Licence Plate Localisation and Detection Process

This part of the algorithm deals with locating the LPs is necessary before proceeding to the next step. It is concerned with detecting and localising the LP. This stage receives a car image as input and produces a portion of the image containing the potential licence plate as output. A licence plate can appear in any portion of the image.

3.1.1 Haar Cascade Classifier-Based Localisation

Face detection pioneered the use of a Haar-like cascade classifier. In the default window, a range of Haar-like features is extracted. Haar cascade is a machine learning-based approach for extracting or highlighting an image's region of interest. It is a classifier that is capable of discriminating between the trained object and the rest of the picture. In its simplest form, a Haar cascade is an XML file containing the object's feature set.

To train the Haar cascade, a large number of positive and negative samples are required to construct a database, where positive samples are simply images of the object to be trained, and negative samples are any random image, but care must be taken to ensure that negative samples do not contain any part of the object. A primitively trained Haar cascade is employed in the framework proposed in this paper. Haar cascade has been shown to be more effective for licence plate localisation than other algorithms such as directly using contours or even some loosely trained neural networks. After that, to get rid of the noise, use the Wiener filtering technique.

3.1.2 Licence Plate Detection

Connected component analysis is a widely used image processing technique that visualises an image and assigns units to pixels based on their relationship to one another. Following class assignment, each pixel is assigned a value based on the variable. Two types of LP forecasts are anticipated based on the LP data.

1. A white background with black characters
2. A black background with white characters

Two different detection models were employed here: To begin, CCA is used to recognise a white frame, and then, it is used to detect black characters. To get linked components of the same size at first, a few parameters are added based on the characters' previous data, such as pixel value of linked component, width greater than 12, height greater than 25, ratio of height-to-width is less than 2.5 or greater than 1.5 and so on. Then, certain non-character-related components are removed by means of creating alternative restrictions between two characters that use the character's LP position.

3.2 *Image Pre-processing and Segmentation*

In order to extract segments of symbols and text characters, image processing and segmentation are needed. We employ four different processing techniques that include greyscale conversion, binarization, dilation and segmentation to analyse these images in this paper. The first step in the process is to convert the image to greyscale, followed by the conversion to binary. Binarization will make both the symbols and characters and the background colours far clearer and more distinct. Picture dilation then enhances the characters by applying its processes.

1. **Greyscale conversion:** The coloured image is transformed to a greyscaled image, which means that instead of three channels (i.e. BGR), the image is reduced to a single eight-bit channel with values varying from 0 to 255, where 0 represents black, and 255 represents white.
2. **Binarization:** The greyscale image is now converted to a binary image using the threshold function. This function accepts only greyscale images and outputs a map of each pixel value to 0 or 255, depending on threshold limit. Any pixel with a value less than the set limit will be mapped to black, and any pixel with a value greater than the set limit will be mapped to 255 (white). The local Otsu approach, which is a conventional binary approach, is used here.
3. **Dilation:** Now, that the image is clean and free of boundary noise, we dilate it to fill in the missing pixels, which are pixels that should have a value of 1 but have a value of 0. It works by considering each pixel in the image individually and then its neighbouring pixels.
4. **Segmentation using Bounding Rectangle:** It involves localising image blobs through a contour algorithm applied to the dilated image. The bounding rectangle algorithm is employed to create a rectangle that encompasses each contour. As the kernel (x, y) size is increased, many characters become confined to the same rectangle box, which is undesirable. In this work, a kernel of $(3, 1)$ is found to be appropriated. Unlabelled text character images can be generated as an output using segmentation.

3.3 *CNN-Based Recognition Process*

CNN is a well-known deep learning model that is used to identify characters in segmented LPs. These layers are used to create CNN models with a variable number of blocks, as well as blocks that have been added or deleted [14].

1. **CONV Layer:** The quality of each neural network varies from one to the next since every pixel is not connected to the layer of weights and biases that follows. However, when the weights/biases are applied, the entire image is partitioned into smaller bits. These are referred to as filters or kernels, and they are convoluted with each smaller region in the input picture to produce feature maps. The hyperparameters of the convolution layer are filter count, local region scale,

stride and padding. These hyperparameters are tuned to achieve effective output for various sizes and genres of the input image.

2. **Pooling Layer:** Pooling layer is used to decrease the processing cost by reducing the spatial dimension of the image and the number of parameters. Max pooling is used in this case, with the $n \times n$ window being slid over the input with a stride value of s . The maximum value in the $n \times n$ region is taken into account for each position, resulting in a reduction in the input size. It exhibits translational invariance, allowing for the recognition of minute differences in position.
3. **FC Layer:** The final pooling layer's flattened output is used as an input to an FC layer in this case. It behaves similarly to a classical neural network, with each neuron from the previous layer connected to the current layer. As a result, the layer's parameter count is greater than the equivalent layer's. It is associated with the output layer, which is often referred to as the classifier.
4. **Activation Function:** Numerous activation functions are used in conjunction with various CNN architectures. It is possible to use the nonlinear activation functions ReLU, LReLU, PReLU and Swish. The nonlinear activation mechanism contributes to the training phase being accelerated. ReLUs are found to be more effective than other functions in this paper.

4 Performance Evaluation

4.1 Implementation Detail

The presented BR-CNN method was speeded up by using a PC with an i5, 8th generation processor and 16 GB of RAM. With OpenCV, the BR-CNN method is processed using Python. For our image classification, we used two distinct datasets. We used Indian licence plate dataset which contains 3200 images of various Indian vehicle licence plates in different illumination condition. We used this dataset to train our Haar cascade classifier on the original image of the vehicle in order to detect licence plates. For validation, we used 1100 car images divided into 36 distinct categories out of which 90% of images were acquired from Internet, and 10% of images was taken manually from streets and roads in different lighting conditions. Classes are specifically comprised of alphanumeric characters (0–9 and A–Z). We have compared our proposed method with the existing models. The experimental results demonstrate the performance's supremacy in both in contrast to conventional licence plate recognition systems in terms of accuracy and efficiency.

4.2 Classical Measurements

The research uses four assessment metrics: accuracy, precision, recall and F1 score, to conduct performance evaluation. The ratio of true expected values to all data

is used to determine accuracy in an experiment. Precision demonstrates the data’s resemblance to the expected value. The amount of data related to the calculated data is stated by recall or sensitivity. The *F1*-score, on other hand, is a precision commission and recall approach to assessing accuracy outcomes. The four matrices have the following formulae:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$F1 - \text{Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

where TP = True positive, TN =True negative, FP = False positive and FN = False negative. In our evaluations, the overall accuracy is calculated based on:

$$\text{Overall Accuracy} = (D \times S \times R)\% \tag{5}$$

where *D* =Plate detection rate, *S* = Plate segmentation rate and *R* = Character recognition rate (Fig. 2).

The relative analysis of BR-CNN and traditional approaches is shown in Table 1. In this case, the values shown in the table are shown as a comparison of precision, recall and F-score on the reported dataset. The table values indicate that the VGG CNN M 1024 model performed ineffectively for LP detection, with a precision of 0.935, a recall of 0.956, an F-score of 0.945. In addition, with a precision of 0.952, a recall of 0.967, an F-score of 0.959, it is clear that the ZF method generated slightly superior performance. At the time, the ResNet50 model had a precision of 0.959, a recall of 0.964, an F-score of 0.961. Accordingly, the VGG16 model achieved a

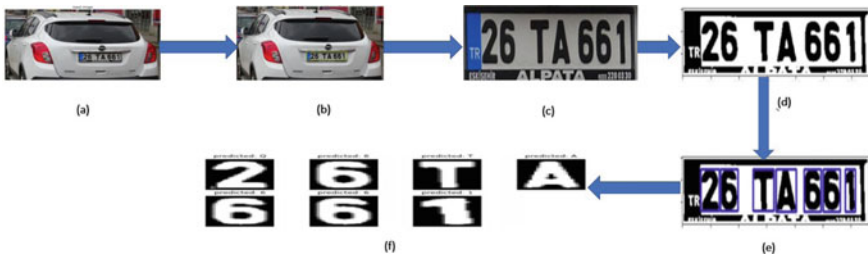


Fig. 2 a Original image. b LP localisation. c Extracted LP. d Pre-processed LP. e Segmented LP. f Predicted LP characters

Table 1 Result analysis of existing models with proposed BR-CNN for applied datasets

Model	Precision	Recall	F1-Score
ZF [27]	0.952	0.967	0.959
VGG_CNN_M_1024 [28]	0.935	0.956	0.945
ResNet50 [29]	0.959	0.964	0.961
VGG16 [28]	0.966	0.975	0.970
Proposed model	0.980	0.978	0.981

relatively moderate precision of 0.966, recall of 0.975, F-score of 0.970. As a result, the BR-CNN model outperformed conventional methods in terms of precision, recall, F-score, with an optimum precision of 0.980, recall of 0.978, F-score of 0.981.

Table 2 summarises some analyses of various ANPR systems. The table compares the plate characters, platform and methodology requirements of each system, as well as different assessments of each system. The accuracy analysis provided by the BR-CNN mechanism in comparison to previous models on the implemented dataset is summarised here. As a result of its higher overall accuracy of 0.978, the presented BR-CNN model obtained the best recognition efficiency. The table also compares detection, segmentation and recognition accuracy across three categories. The projected BR-CNN model performed significantly better than previous techniques at recognising all applied images.

Table 2 State-of-the-art ANPR systems compared with the proposed ANPR system

System	Plate detection	Character segmentation	Character recognition	Overall accuracy	Plate characters
Ref. [21]	0.971	0.983	0.978	0.935	English, Japanese
Ref. [22]	0.965	NR	0.891	0.860	English
Ref. [30]	0.973	NR	0.945	0.919	Persian
Ref. [27]	0.969	0.987	0.945	0.904	Persian
Ref. [26]	0.993	NR	0.966	0.960	Chinese, English
Ref. [31]	0.959	NR	0.923	0.90	English
Ref. [30]	0.973	NR	0.957	0.931	Persian
Ref. [32]	0.971	NR	0.964	0.936	English
Ref. [33]	0.987	1	0.976	0.963	Persian
Our system	0.993	1	0.985	0.978	English

^a NR: Not Reported

5 Conclusion

A new BR-CNN technique for successful detection and identification of LPs has been proposed in this paper. Three phases make up the current BR-CNN model. The first stage involves the use of Haar cascade and CCA models to perform LP localisation and detection. The LP image is then segmented using the bounding rectangle technique, and the characters in the LP are eventually identified using the CNN model. On the used dataset, the suggested BR-CNN model attained the highest possible overall accuracy of 0.978, according to the results of experimental analysis. The BR-CNN model's ability to understand multilingual LPs may be improved in future. The provided techniques, algorithms and parameter setting procedures, as well as our data set and related analyses, provide a comprehensive collection of solutions to the issues and challenges that may come up with integrating ANPR systems into various ITS applications.

References

1. R. Qian, R.T. Tan, W. Yang et al., in *Attentive Generative Adversarial Network for Raindrop Removal from a Single Image*. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (2018), pp. 2482–2491
2. J. Sun, W. Cao, Z. Xu et al., in *Learning a Convolutional Neural Network for Non-Uniform Motion Blur Removal*. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (2015), pp. 769–777
3. V. Abolghasemi, A. Ahmadyfard, An edge-based color-aided method for license plate detection. *Image Vis. Comput.* **27**(8), 1134–1142 (2009)
4. B. Hongliang, L. Changping, in *A Hybrid License Plate Extraction Method Based on Edge Statistics and Morphology*. Proc. IEEE 17th ICPR (Vol. 2, 2004), pp. 831–834
5. A. Mousa, Canny edge-detection based vehicle plate recognition. *Int. J. Sign. Process. Image Process. Pattern Recognit.* **5**(3), 1–8 (2012)
6. T.D. Duan, D.A. Duc, T.L. Du, in *Combining Hough Transform and Contour Algorithm for Detecting Vehicles' License-Plates*. Proc. IEEE Int. Symp. Intell. Multimedia, Video Speech Process (2004), pp. 747–750
7. K. Deb, A. Vavilin, K.H. Jo, in *An Efficient Method for Correcting Vehicle License Plate Tilt*. Proc. IEEE Int. Conf. GrC. (2010), pp. 127–132
8. S. Du, M. Ibrahim, M. Shehata et al., Automatic license plate recognition (ALPR): a state-of-the-art review. *IEEE Trans. Circuits Syst. Video Technol.* **23**(2), 311–325 (2012)
9. P. Kanani, A. Gupta, D. Yadav et al., in *Vehicle License Plate Localization Using Wavelets*. Proc. IEEE ICT (2013), pp. 1160–1164
10. R.T. Lee, K.C. Hung, in *Real-Time Vehicle License Plate Recognition Based on 1-d Discrete Periodic Wavelet Transform*. Proc. IEEE IS3C (2012), pp. 914–917
11. J.W. Hsieh, S.H. Yu, Y.S. Chen, in *Morphology-Based License Plate Detection from Complex Scenes*. Proc. IEEE 16th Int. Conf. Pattern Recognit. (Vol. 3, 2002), pp. 176–179
12. D. Llorens, A. Marzal, V. Palazon, in *Car License Plates Extraction and Recognition Based on Connected Components Analysis and HMM Decoding*. In Iberian Conference on Pattern Recognition and Image Analysis (Springer, Berlin, Heidelberg, 2005), pp. 571–578
13. J.C. Poon, M. Ghadiali, G.M. Mao, in *A Robust Vision System for Vehicle Licence Plate Recognition Using Grey-Scale Morphology*. Proc. IEEE ISIE (Vol. 1, 1995), pp. 394–399

14. Y. Wen, Y. Lu, J. Yan et al., An algorithm for license plate recognition applied to intelligent transportation system. *IEEE Trans. Intell. Transp. Syst.* **12**(3), 830–845 (2011)
15. C.N. Anagnostopoulos, I.E. Anagnostopoulos, V. Loumos et al., A license plate-recognition algorithm for intelligent transportation system applications. *IEEE Trans. Intell. Transp. Syst.* **7**(3), 377–392 (2006)
16. W. Niblack, *An Introduction to Digital Image Processing* (Strandberg Publishing Company, 1985)
17. J. Sauvola, M. Pietikäinen, Adaptive document image binarization. *Pattern Recogn.* **33**(2), 225–236 (2000)
18. C. Wolf, J.M. Jolion, F. Chassaing, in *Text Localization, Enhancement and Binarization in Multimedia Documents*. Proc. IEEE 16th Int. Conf. Pattern Recognit. (Vol. 2, 2002), pp. 1037–1040
19. N. Otsu, A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**(1), 62–66 (1979)
20. T. Ejima, The characteristic feature based on four types of structural information and their effectiveness for character recognition. *IEICE Trans.* **68**(4), 789 (1985)
21. S. Ghofrani, M. Rasooli, Farsi license plate detection and recognition based on characters features. *Majlesi J. Electr. Eng.* 44–51 (2011)
22. M.S. Sarfraz, A. Shahzad, M.A. Elahi, Real-time automatic license plate recognition for CCTV forensic applications. *J. Real-Time Image Proc.* **8**(3), 285–295 (2013)
23. M.H. Glauberger, Character recognition for business machines. *Electronics* **29**(2), 132–136 (1956)
24. M.M. Dehshibi, R. Allahverdi, Persian vehicle license plate recognition using multiclass Adaboost. *Int. J. Comput. Electr. Eng.* **4**(3), 355 (2012)
25. H. Bay, T. Tuytelaars, L. Van Gool, in *Surf: Speeded Up Robust Features*. Proc. ECCV (Springer, Berlin, Heidelberg, 2006) pp. 404–417
26. R. Azad, F. Davami, B. Azad, A novel and robust method for automatic license plate recognition system based on pattern recognition. *Adv. Comput. Sci. Int. J.* **2**(3), 64–70 (2013)
27. M.D. Zeiler, R. Fergus, in *Visualizing and Understanding Convolutional Networks*. European Conference on Computer Vision (Springer, Cham, 2014), pp. 818–833
28. K. Simonyan, A. Zisserman Very deep convolutional networks for large-scale image recognition (2014). arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556)
29. K. He, X. Zhang, S. Ren, in *Deep Residual Learning for Image Recognition*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016), pp. 770–778
30. Z.X. Chen, C.Y. Liu, F.L. Chang, Automatic license-plate location and recognition based on feature saliency. *IEEE Trans. Veh. Technol.* **58**(7), 3781–3785 (2009)
31. J. Jiao, Q. Ye, Q. Huang, A configurable method for multi-style license plate recognition. *Pattern Recognit.* **42**(3), 358–369 (2009)
32. J.M. Guo, Y.F. Liu, License plate localization and character segmentation with feedback self-learning and hybrid binarization techniques. *IEEE Trans. Veh. Technol.* **57**(3), 1417–1424 (2008)
33. R. Panahi, I. Gholampour, Accurate detection and recognition of dirty vehicle plate numbers for high-speed applications. *IEEE Trans. Intell. Transp. Syst.* **18**(4), 767–779 (2016)

VLSI Design and Materials

First Principle Study of Mechanical and Thermoelectric Properties of In-Doped Mg_2Si



Abdullah bin Chik and Lam Zi Xin

Abstract In this study, the mechanical and thermoelectric properties of Mg_2Si doped with In were successfully calculated using DFT package CASTEP, and semi-classical Boltzmann transport theory in relaxation time approximation BoltzTraP and Phono3py packages. The mechanical property calculations show that the doped compound is softer, less rigidity, and less brittle. The band structure and density of states calculations show that the doped compound showing metallic behavior with increased density of states at the Fermi level due to Si $3p$ states, Mg $2p$ states, and In $5p$ states. The thermoelectric property calculations show decreased magnitude of Seebeck coefficient, and conduction type changes from n to p type. The conductivity of In doped increases as well as the thermal conductivity, leading to reduced figure of merit at high temperature from 0.43 in undoped compound to 0.15 in In-doped compound. In conclusion, the In dopant has improved the mechanical properties of Mg_2Si , but at the same time reduced the thermoelectric properties of the compound.

Keywords Density functional theory · Thermoelectric materials · First principle study

1 Introduction

Thermoelectric materials and technology have been utilized in industrial applications since 1950s with thermoelectric material bismuth telluride. The thermoelectric materials were widely accepted in industrial practice, with applications in space missions, laboratory equipment, wearable nano-based technology, and medical tools

A. Chik (✉) · L. Z. Xin

Centre for Frontier Materials Research, Universiti Malaysia Perlis, 01000 Kangar, Perlis, Malaysia
e-mail: abdullahchik@unimap.edu.my

Center of Excellence Geopolymer and Green Technology (CEGeoGTech), Universiti Malaysia Perlis, 01000 Kangar, Perlis, Malaysia

Faculty of Chemical Engineering Technology, Universiti Malaysia Perlis, Taman Muhibbah School Complex 2, 02600 Jejawi, Perlis, Malaysia

[1]. Renewed interest in thermoelectric materials research came from the need to create an alternative energy system that is both environmentally friendly and efficient [2]. The efficiency of thermoelectric materials can be calculated using on the formula, $ZT = S^2\sigma T/\kappa$ [3], where ZT , S , σ , T , and κ are figure of merit, Seebeck coefficient, conductivity, temperature, and thermal conductivity, respectively. The materials with high figure of merit will have large Seebeck coefficient, large electrical conductivity, and low thermal conductivity.

Intermetallic compounds with higher thermoelectric properties were found with forming compounds of Mg with Si, Ge or Sn. Mg_2Si , for example, can be prepared to be either p type conduction by doping with Ag and Cu and n type conduction by doping with Sb, Al, and Bi [4] Doped Mg_2Si compounds also have been studied for its low thermal conductivity and high electrical conductivity, such as Sakamoto et al. [5] reported Mg_2Si -based devices made of Sb-doped and Al-doped Mg_2Si . However, these compounds still showing brittleness nature or low mechanical properties. In this study, we conduct a density functional theory investigation of In-doped magnesium silicide (Mg_2Si). In has been known to be highly ductile metal and p type element, and by doping with In at 12.5% concentration, we aim to study the mechanical properties and the thermoelectric properties $Mg_2Si_{0.875}In_{0.125}$.

2 Computational Method

The thermoelectric properties and elastic constant of In-doped Mg_2Si were investigated using density functional theory approach. The Mg_2Si was doped with In at Si site with concentrations of 12.5%. In order to achieve this concentration, a $2 \times 2 \times 2$ supercell of Mg_2Si (24 atoms) was generated, with a single In atom replaced one Si atom of the supercell. The calculations were performed using the first-principles software package CASTEP [6] based on the plane-wave pseudo-potential approach. We utilized the exchange correlation potential described by the generalized gradient approximation (GGA) with Perdew-Burke-Ernzerhof (PBE) scheme and a $3 \times 3 \times 3$ Monkhorst-Pack K -point grid for Brillouin-zone integration [7]. The plane-wave energy cut-off used in calculations was 489 eV. Initially, the geometry optimization procedure was carried out for the doped and undoped compound. After suitable geometry-optimized supercell was found, the self-consistence field calculations were carried out and followed by the calculations of electronic properties. The mechanical properties were investigated using elastic constant calculations. The convergence criterion for the force between atoms is set to be 0.01 eV/Å, for a maximum displacement is 0.001 Å, when the total stress tensor is reduced to the order of 0.05 GPa. The total self-consistent field (SCF) energy change is 5×10^{-6} eV/atom, and the energy between optimization steps is 5×10^{-5} eV/atom. The thermoelectric properties were calculated using BoltzTrap package [8], which utilized semiclassical Boltzmann transport equation, from the output of CASTEP package. The lattice thermal conductivity for both compounds was calculated using Phono3py [9] package from the output of DFT package.

3 Results and Discussion

The Mg_2Si crystal belongs to $\text{Fm}\bar{3}\text{m}$ space group and has a cubic antiferroite structure. In the primitive cell, there are two Mg atoms which are located at $\pm\mu$ ($\mu = (1/4, 1/4, 1/4)$ a, where a is the lattice constant), and one Si atom which occupies a face-centered cubic (fcc) site. The unit cell used in this study was taken from the Materials Project Web site [10], which is a free open database offering material properties. The geometry-optimized lattice constant of our compound is 4.498 Å. The $\text{Mg}_2\text{Si}_{0.875}\text{In}_{0.125}$ compound, after geometry optimization procedure, also shows $\text{Fm}\bar{3}\text{m}$ space group, but with slight increase in lattice constants, 4.539 Å, and cell volumes as shown in Table 1. The increase in lattice constant is due to the bigger ionic radius of In at 0.8 Å compared to Si ionic radius of 0.4 Å [11]. The calculation of elastic constants using Voight-Reuss-Hill approximation yields bulk modulus, shear modulus, and Young modulus is also shown in Table 1.

The experimental bulk modulus, shear modulus, and Young modulus of Mg_2Si are also included in Table 1. The obtained experimental data were in good agreement with our results, as reported by Schmidt et al. [12]. The In-doped Mg_2Si shows a reduced in bulk, shear, and Young modulus compared to undoped compound. The possible explanation is that In is a soft and ductile metal; the effect of substitutional In atoms have “softened” the lattice, and the lattice relaxation resulted in decrease of all modulus in this compound. Reduced in all modulus, show that the doped compound is less rigidity, more flexibility, more elasticity, and less brittle. Figure 1 shows the band structure of Mg_2Si doped with In with concentration of $x = 0.0$ and 0.125. The band structure of Mg_2Si shows that the compound is an indirect semiconductor with band gap of 0.201 eV, measuring from Γ to X , similar to results reported by Wang et al. [13]. The In-doped Mg_2Si show that the Fermi level overlap with the conduction and valence band at point Γ , exhibiting metallic behavior. Finally, the In-doped compound showing dense band lines approaching the Fermi level at Γ point.

Figure 2 shows the partial density of state plots for Mg_2Si and $\text{Mg}_2\text{Si}_{0.875}\text{In}_{0.125}$. The Mg partial density of states of undoped compound, shown in Fig. 2a, shows that the 2p and 3s states are the components of conduction band and valence. Both 3s and 2p states are in the range of 0.1–15 eV in conduction band, and about –9.5 to –7.3 eV and about –5.0 to –0.1 eV. Figure 2b shows the Si partial density of states for

Table 1 Lattice and mechanical properties of In-doped Mg_2Si

Properties	Mg_2Si (this work)	Mg_2Si (exp. [12])	$\text{Mg}_2\text{Si}_{0.875}\text{In}_{0.125}$
a (Å)	4.498	6.35	4.539
Cell volume (Å ³)	64.338	–	66.131
Bulk Modulus (GPa)	52.897	49.0	50.737
Shear Modulus (GPa)	45.819	48.92	31.387
Young Modulus (GPa)	106.661	116.7	78.065

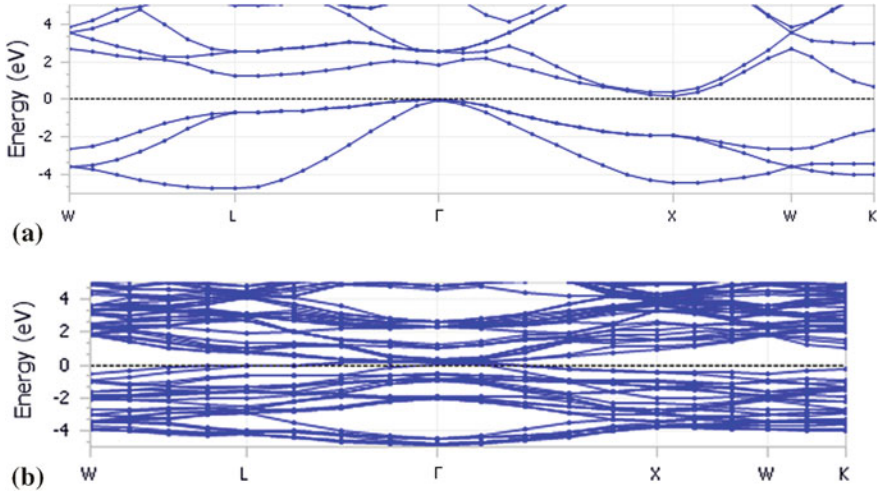


Fig. 1 Band structure of **a** undoped Mg_2Si , **b** 12.5% In-doped Mg_2Si , **c** 25.0% In-doped Mg_2Si

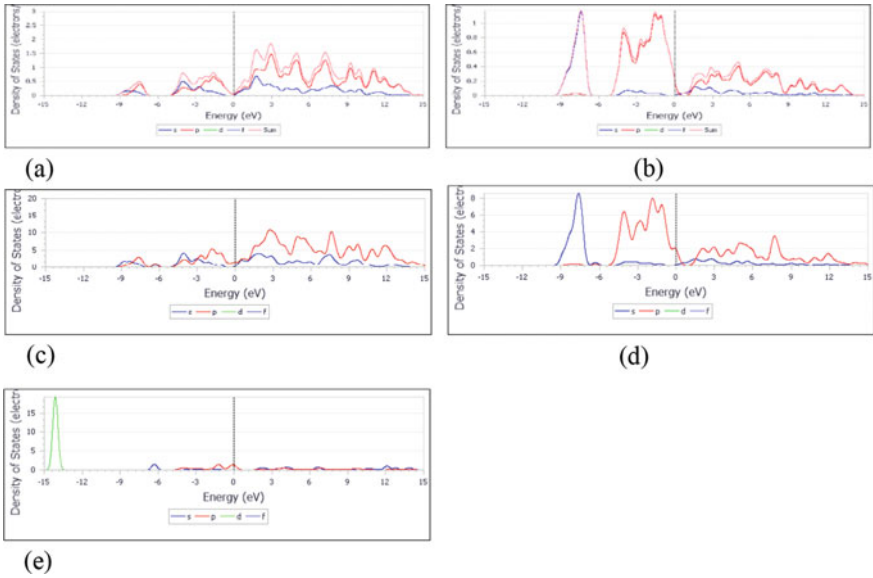


Fig. 2 Partial density of states of **a** Mg in Mg_2Si , **b** Si in Mg_2Si , **c** Mg in $Mg_2Si_{0.875}In_{0.125}$, **d** Si in $Mg_2Si_{0.875}In_{0.125}$, **e** In in $Mg_2Si_{0.875}In_{0.125}$

undoped Mg_2Si , exhibiting more $3p$ states compared to $3s$ states are in conduction band as well as in valence band. Compared to conduction band, more $3p$ states are in conduction band from range of about -9.5 to -6.8 eV and about 5.0 to 0 eV. For the doped compound, Fig. 2c shows the Mg partial density of states. In doped has shift the Fermi level upward closed to the upper $3s$ states in conduction band ranged from about 0 to 13.7 eV. The $2p$ states in doped Mg_2Si are filling the Fermi level changed the conduction type of $\text{Mg}_2\text{Si}_{0.875}\text{In}_{0.125}$ to metallic conduction. Figure 2d shows the Si partial density of states for $\text{Mg}_2\text{Si}_{0.875}\text{In}_{0.125}$, which describing that $3p$ states filling the Fermi level ranged from about -5.5 eV to 0.8 eV. The $3p$ states also is the majority states in both conduction and valence bands. The $3s$ states started to become majority states from about -7.0 to 9.5 eV. In partial density of states in In-doped compound, shown in Fig. 2e, shows that $5p$ states are filling the Fermi level. Both $5s$ and $5p$ states are shown to be components of the valence band and conduction band. The $4d$ states are deep inside the atom at about -13.5 to -15.0 eV. Overall, In-doped compound shows more states at Fermi level compared to undoped compound.

Figure 3 shows the plot of temperature dependence of thermoelectric properties of Mg_2Si and $\text{Mg}_2\text{Si}_{0.875}\text{In}_{0.125}$. Figure 3a shows the temperature-dependent Seebeck coefficient for Mg_2Si and In-doped compound. The Seebeck coefficient for Mg_2Si has negative values, indicating the compound is n -type material. However, for In-doped compound, the sign of Seebeck coefficient has changed to negative values, indicating material having p type carriers. The In has been known as a p -type material, and thus, the In-doped Mg_2Si has changed conduction type to the p -type materials. The magnitude of Seebeck coefficient, however, is seen decreased in In-doped compound. Equation 1 [14] shows the relation between the Seebeck coefficient and density of carriers:

$$S = \frac{8\pi^2 k^2}{3eh^2} mT \left(\frac{\pi}{3n} \right)^{2/3} \quad (1)$$

where S , k , e , h , and m are Boltzmann constant, carrier charge, Planck's constant, and effective mass, respectively. Analysis of BoltzTraP data shows that density of carriers per unit cell has increased in In-doped compound due to increase in density of states at Fermi level. Hence, with the increasing of density of carriers, the Seebeck coefficient will be decreased in In-doped compound, according to Equation 1. Figure 3b shows the temperature-dependent conductivity for Mg_2Si and $\text{Mg}_2\text{Si}_{0.875}\text{In}_{0.125}$. The conductivity of In-doped compound shows an increase in magnitude compared to undoped compound. According to Equation 2 [14], shown below, describing the relation between carrier density and conductivity:

$$\sigma = ne\mu \quad (2)$$

where σ , n , e and μ are conductivity, carrier density, carrier charge, and mobility, respectively, the increases in density as discussed in paragraph above, will lead to

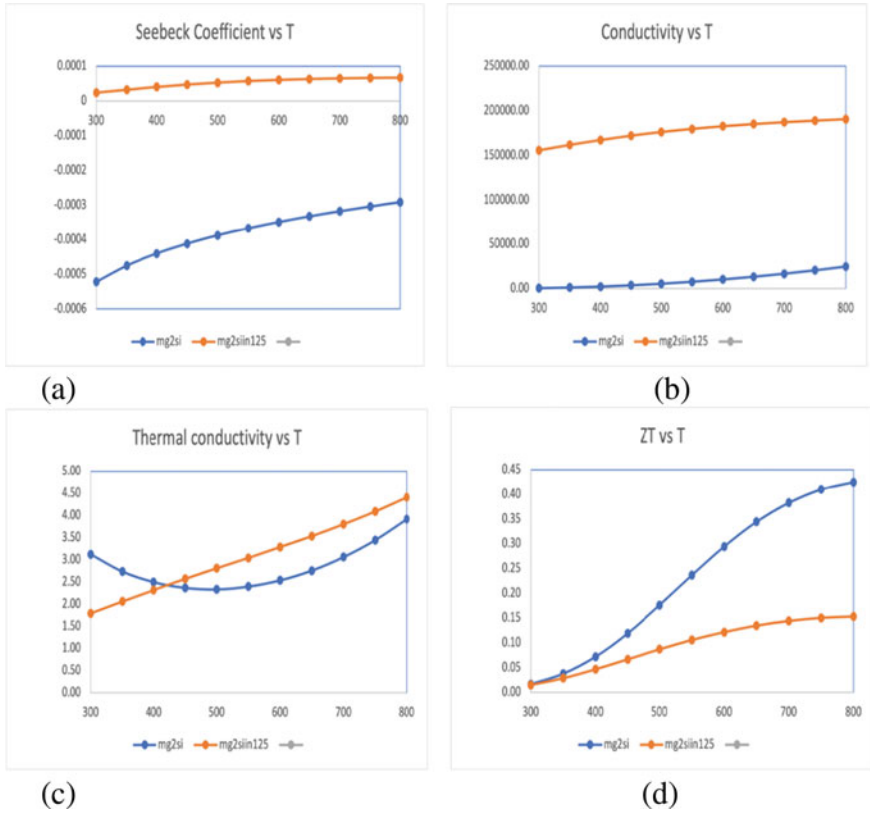


Fig. 3 Thermoelectric properties of Mg₂Si and Mg₂Si_{0.875}In_{0.125}: **a** Seebeck coefficient, **b** electrical conductivity, **c** thermal conductivity, and **d** figure of merit vs temperature

increase in overall conductivity. Figure 3c shows the temperature-dependent thermal conductivity of Mg₂Si and In-doped compound. The thermal conductivity of In-doped compound shows the thermal conductivity increases from temperature 400 K onward, compared to the undoped compound. According to Equation 3 [14] below:

$$k_{el} = L\sigma T \tag{3}$$

where k_{el} , L , σ , and T are electronic thermal conductivity, Lorentz number, conductivity, and temperature, respectively, the increases in conductivity as discussed in paragraph above, will actually lead to increase in electronic thermal conductivity and overall thermal conductivity. So due to In doping, the change of conduction type from semiconducting to metallic behavior caused the Fermi level overlapping the conduction and valence bands. This creates higher density of states at Fermi level and thus increasing the density of carriers for the In-doped compound. Hence, the conductivity and thermal conductivity dramatically improved in In-doped compound.

Finally, Fig. 3d shows the temperature dependent of figure of merit, ZT, for Mg_2Si and $\text{Mg}_2\text{Si}_{0.875}\text{In}_{0.125}$. The graph figure of merit vs temperature for Mg_2Si shows that the compound exhibit high ZT at 800 K with value of 0.43, similar to reported values in literature [15]. The figure of merit for In-doped compound shows a dramatic decrease compared to undoped compound from 0.43 to 0.15 at 800 K. The In dopant in Si site has decreases the Seebeck coefficient magnitude throughout temperature range, and increases the conductivity and thermal conductivity, resulting lower ZT.

4 Conclusion

In this study, the mechanical and thermoelectric properties of Mg_2Si doped with In were successfully calculated using CASTEP, BoltzTraP, and Phono3py packages. The mechanical property calculations show that the doped compound is softer, less rigidity, and less brittle. The band structure and density of states calculations show that the doped compound showing metallic behavior with increased density of states at the Fermi level. The thermoelectric property calculations show decreased magnitude of Seebeck coefficient and conduction type changes from *n* to *p* type. The conductivity of In doped increases as well as the thermal conductivity, leading to reduced figure of merit at high temperature. In dopant has improved the mechanical properties of Mg_2Si but at the same time reduced the figure of merit of the compound.

Acknowledgements The authors would like to thank Center of Excellence Frontier Materials Research, Center of Excellence Geopolymer and Green Technology (CEGeoGTech), the Faculty of Chemical Engineering Technology, Universiti Malaysia Perlis.

References

1. S. Yokhasing, K. Chaarmart, M. Rittiruam, K. Matarat, T. Seetawan, *Mater. Today: Proc.* **5**, 14074–14078 (2018)
2. K. Kaur, S. Dhiman, R. Kumar, *Indian J. Phys.* **91**, 1305–1317 (2017)
3. Y.L. Chen, J.G. Analytis, J.-H. Chu et al., *Science* **325**(5937), 178–181 (2009)
4. M. Akasaka, T. Iida, A. Matsumoto, K. Yamanaka, Y. Takanashi, T. Imai, N. Hamada, *J. Appl. Phys.* **104**(1), 013703 (2008)
5. T. Sakamoto, T. Iida, S. Kurosaki, K. Yano, H. Taguchi, K. Nishio, Y. Takanashi, *J. Elec. Mat.* **40**(5), 629–634 (2010)
6. V. Milman, B. Winkler, J.A. White, C.J. Pickard, M.C. Payne, E.V. Akhmatkaya, R.H. Nobes, *Int. J. Quantum Chem.* **77**, 895 (2000)
7. J.P. Perdew, J.A. Chevary, S.H. Vosko, K.A. Jackson, M.R. Pederson, D.J. Singh, C. Fiolhais, *Phys. Rev. B* **46**, 6671 (1992)
8. G.K.H. Madsen, D.J. Singh, *Comput. Phys. Commun.* **175**, 67 (2006)
9. A. Togo, L. Chaput, I Tanaka, *Phys. Rev. B* **91**, 094306 (2015)
10. A. Jain, S. P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, *APL Mater.* **1**(1), 011002 (2013)
11. R.D. Shannon, *Acta Cryst.* **A32**, 751 (1976)

12. R.D. Schmidt, E.D. Case, J. Giles et al., *J. Electron. Mater.* **41**, 1210–1216 (2012)
13. H. Wang, W. Chu, H. Jin, *Comput. Mater. Sci.* **60**, 224–230 (2012)
14. H. Alam, S. Ramakrishna, *Nano Energy* **2**, 190–212 (2013)
15. Y.Z. Zhang, Y.H. Han, Q.S. Meng, *Mater. Res. Innov.* **19**(S1), 265 (2015)

Design and Analysis of Area and Energy-Efficient Quantum-Dot Half Adder Using Intercellular Interaction Technique



Neeraj Tripathi , Mohammad Mudakir Fazili , Abhishek Singh , Shivam , and Suksham Pangotra 

Abstract The current trends in micro- and nano-electronics are incompatible with the CMOS-based VLSI technology. Quantum-dot cellular automata (QCA) a new technology based on applications of quantum physics enables nano-fabrication of digital logic circuits. This new paradigm enables ultra-low power and high-speed operation. In this paper, an area- and energy-efficient QCA circuit for half adder is presented. The proposed design is simulated using QCA Designer software. The simulated waveform is in agreement with the truth table. Strong polarization of the output signals is reported. The critical efficiency parameters of interest are cell count, cell area, latency, area-delay product, and energy dissipation. The presented design uses 17 cells occupying an area of $0.018 \mu\text{m}^2$. A 20% area efficiency is achieved using the proposed QCA layout. A latency of 1 clock phase is incurred in the design. The energy dissipation analysis of the proposed circuit is performed using QCA Pro software. The total energy dissipation of the circuit is reported to be 26.92 meV at 0.5 kink energy.

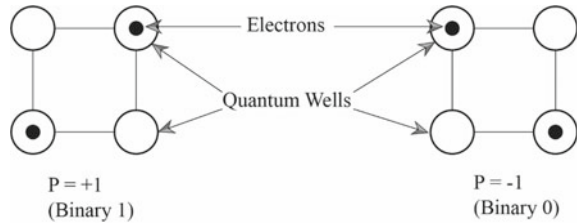
Keywords Nano-electronics · Quantum-dot cellular automata · Quantum physics · Half adder · Energy dissipation

1 Introduction

The need for more computational power has led to the research of more robust and efficient micro- and nano-electronics. Moore's law has been the guiding principle for the prediction of increase in transistor density. But now, as the size of the CMOS gate has reached the size of a molecule, i.e., 10 nm , we now encounter quantum effects. Also, the short channel effects hinder further miniaturization. Beyond 10 nm fabrication, the atoms of two neighboring CMOS gates interact with each other and thus results in distorted results [1, 2]. The solution to all these problems was proposed by Lent et al. in [3]. In [3], the researchers proposed that using quantum phenomenon

N. Tripathi · M. M. Fazili (✉) · A. Singh · Shivam · S. Pangotra
Shri Mata Vaishno Devi University, Katra, JK 182320, India

Fig. 1 Two possible states of QCA cell



like tunneling and superposition, quantum circuits that are more efficient than current CMOS-based circuits can be developed. Quantum-dot cellular automata (QCA) technology was proposed as an alternative to CMOS technology [4]. QCA technology is an application of quantum physics. In QCA technology, charge is trapped in metal islands called quantum wells, and to propagate information, tunneling energy is provided through clocking [5].

QCA technology is different from conventional CMOS fabrication technology. CMOS technology uses a combination of transistors to realize logic gates and then uses these logic gates to realize functions. But in QCA technology, QCA cells are used to realize logic functions based on their cell–cell interaction. The polarization of one cell affects the polarization of neighboring cells in a radius of 65 nm. The polarization is propagated using clocking energization. The two electrons reside in the QCA cell in two diagonally opposite quantum wells [6]. There exist two possible states for electron pairs to reside in a quantum cell that can be observed in Fig. 1.

In this paper, a fresh QCA circuit design for half adder is proposed based on cell–cell interaction instead of the majority voter approach. Prior designs [7–13] use the majority voter approach to realize the half adder function. The presented circuit is efficient on critical design parameters. The design metrics of interest are cell count, cell area, area efficiency, latency, area-delay product, complexity, and energy dissipation. The QCA half adder design [7] uses 21 cells and consumes 0.02 μm^2 area. However, the presented design uses 17 cells and consumes only 0.018 μm^2 area. Based on the design metrics, the presented design is found to be 20% area-efficient compared to [7].

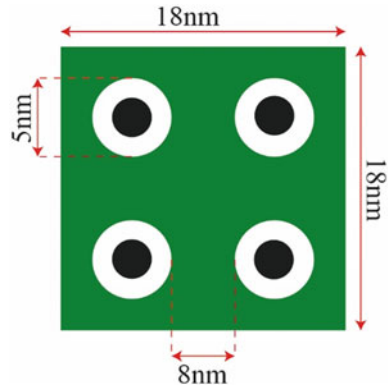
In Sect. 2, a brief discussion on QCA concepts is presented. Section 3 presents the proposed design and examines its results. Discussions and comparative analysis are presented in Sect. 4.

2 Background

2.1 Quantum-Dot Cellular Automata

The information processing in QCA technology is based on coulombic interaction between the electrons within a cell and the cells surrounding it. A QCA cell consists

Fig. 2 Features of a QCA cell



of four quantum wells separated by quantum barriers. Two electrons reside in the two diagonally opposite wells to reduce coulombic interaction. Binary information is represented by a QCA cell using the two allowable positions of electrons. If the two electrons reside in quantum well 1 and 3, it represents logic ‘0’, and if the electrons reside in quantum well 2 and 4, it represents logic ‘1’. The propagation of information takes place using the lowering of quantum barriers so that the electrons can change their position within a cell based on the polarization of its neighboring cells. The radius of effect of a cell is assumed to be 65 nm. A quadrupole moment is induced in neighboring cells because of the electrostatic interaction. In QCA technology, if two cells are placed adjacent to each other, they will tend to align their polarization as a result of the electrostatic interaction of electrons between cells. Therefore, QCA cells can act both as switching devices and interconnects [6, 14]. Figure 2 depicts the features of a QCA cell.

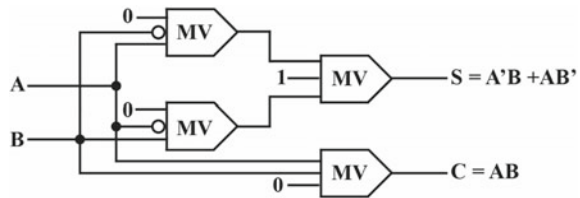
The electrostatic potential energy (U) is calculated using Eq. 1. Here, $F = 23.06 \times 10^{-20}$ J nm and r_{ij} are the distance between the electron i and j in nm. The stability of electron is determined by observing the lowest energy state.

$$U_{ij} = \sum_{ij} \frac{q_i q_j}{4\pi \epsilon_0 \epsilon_r r_{ij}} = F \sum_{ij} \frac{1}{r_{ij}} \tag{1}$$

2.2 Half Adder

A half adder circuit adds two input bits and produces two output bits sum and carry. The half adder maps input vector $I(A, B)$ to output vector $O(S = A \oplus B, C = AB)$. Prior half adders QCA designs [7–13] use the conventional method of majority logic implementation depicted in Fig. 3. Using majority logic implementation, the half adder QCA design is saturated at 21 cells occupying 0.02 μm^2 area in Ref. [7]. But, using the intercellular interaction technique, the further reduction in cell area is

Fig. 3 Majority logic implementation of half adder circuit



possible. The proposed design employed the intercellular interaction technique and realized half adder using only 17 QCA cells occupying $0.018 \mu\text{m}^2$ of area. A 20% efficiency in terms of the cell area is reported.

2.3 Design Modeling Software

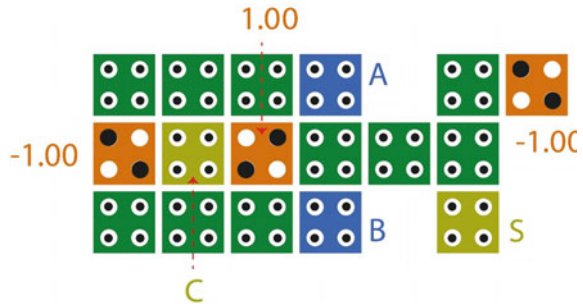
Mainly, two software's are used in modeling and analyzing QCA circuit layouts. Firstly, the QCADesigner tool is used to lay the basic QCA design layout and then simulate it for output waveforms [15]. Then, to calculate the energy dissipation, QCAPro software is used [16]. QCADesigner tool provides two simulation engines: bistable simulation engine and coherent vector engine. The bistable simulation engine uses the bistable property of QCA cells and calculated the polarization of the cells using the ground state computation method. The bistable simulation engine is fast but less accurate in calculating the polarization of cells. However, the coherent vector engine is slow but provides accurate simulation results. For the energy analysis, the coherent vector engine is used to verify the vector set and then calculate the average leakage and average switching energy dissipation of the circuits for three levels of kink energy. QCAPro tool provides the energy dissipation map of the circuit depicting the energy dissipation by each cell.

3 Proposed Work

Conventionally, all the QCA circuit designs are based on combinations of majority and inverter gates. The inputs of the majority gate are polarized to either -1 or 1 , and correspondingly, AND or OR logic functions are realized. In this work, the cell-cell interaction method is used in realizing the logic function of the half adder is presented. In the cell-cell interaction method, considering the effect of the radius of each polarized cell to be 65 nm , polarization of the surrounding cells is calculated. This approach aids in reducing the cell count and thus helps in reducing cell area.

Figure 4 presents the proposed design of the QCA half adder circuit. As depicted in Fig. 4, A and B represent the two input cells, while S and C represent the sum and carry output cells. The QCA layout of the half adder circuit uses 17 cells. The

Fig. 4 Proposed QCA layout of half adder circuit



presented design occupies a cell area of $18,618 \text{ nm}^2$ (or $0.018 \text{ }\mu\text{m}^2$). The design has a delay of 0.25 clock cycle.

The presented design is simulated for verification using the QCADesigner tool [15]. The simulation engine used for calculating the polarization is the bistable simulation engine. It is a fast but less precise simulation engine. The parameters configured for the simulation are (Table 1).

Figure 5 presents the simulated waveform. The output values are exactly as desired. The truth table and the simulated waveform of the half adder are in agreement.

QCAPro tool [16] is used to compute the power and energy requirements of the proposed circuit. The presented design is analyzed for average leakage energy dissipation and average switching energy dissipation. The energy maps for three kink energy levels are depicted in Fig. 6.

Table 2 presents energy dissipation estimates for three levels of kink energy. The presented design dissipates minimum energy of 26.92 meV (for $0.5 E_k$). The circuit is simulated at temperature, $T = 1\text{K}$.

Table 1 Values of design parameters used in the simulation

Parameter	Value
Type of cell	90°
Cell size	$18 \times 18 \text{ nm}^2$
Intercellular gap	2 nm
Simulation engine	Bistable simulation engine
Number of sample	12,800
Convergence tolerance	0.001
Radius of effect	65.000 nm^2
Relative permittivity	12.900
Maximum iterations per sample	100

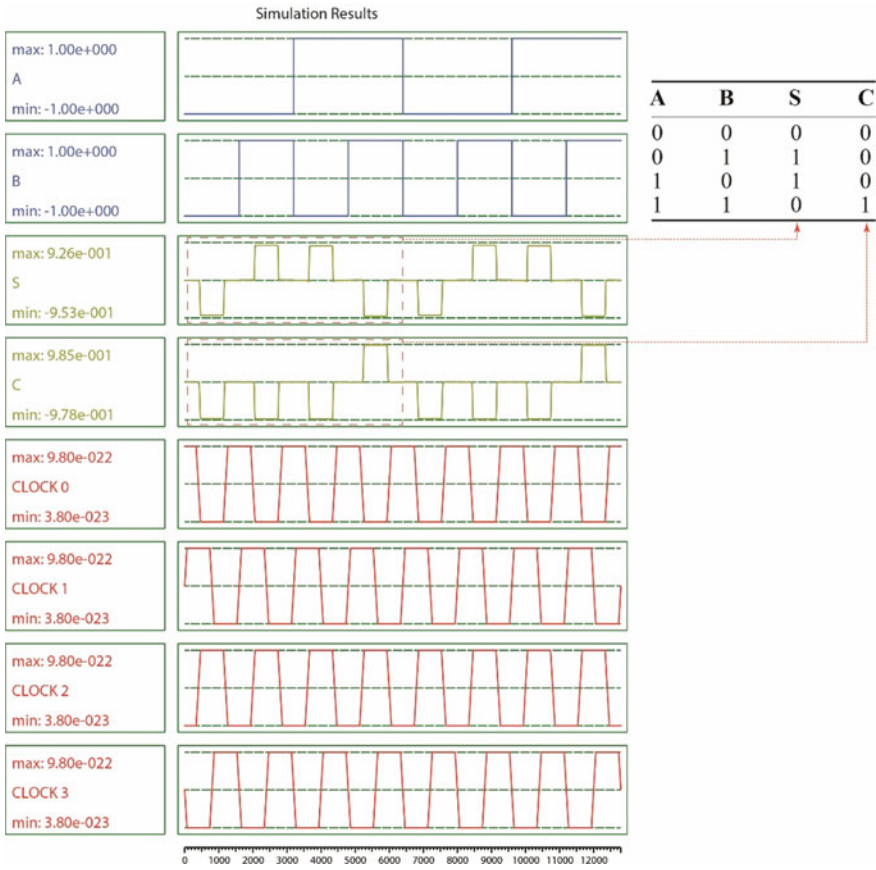


Fig. 5 Simulated waveform of half adder circuit

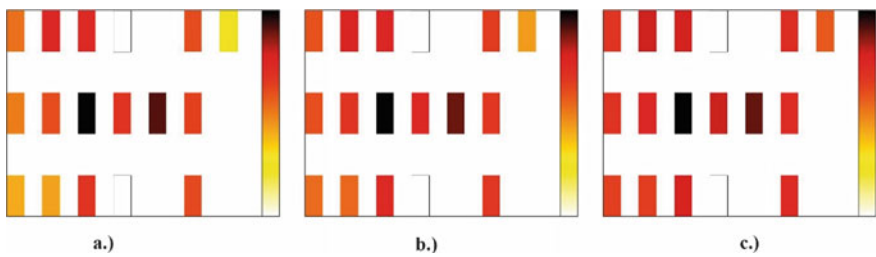


Fig. 6 Energy dissipation maps of proposed half adder at a) 0.5Ek, b) 1.0Ek, c) 1.5Ek

Table 2 Energy dissipation estimates of proposed QCA half adder

Average leakage energy dissipation (meV)			Average switching energy dissipation (meV)			Total energy dissipation (meV)		
0.5E _k	1.0E _k	1.5E _k	0.5E _k	1.0E _k	1.5E _k	0.5E _k	1.0E _k	1.5E _k
5.47	15.29	26.40	21.45	18.37	15.74	26.92	33.66	42.14

Table 3 Comparison of the proposed design with prior designs

Design	Cell count	Cell area μm ²	Latency (clock phases)	Area delay product	Crossover type
Ref. [8]	77	0.08	4	0.08	Multilayer
Ref. [10]	65	–	5	–	Coplanar
Ref. [11]	62	0.08	8	0.16	None
Ref. [9]	61	–	3	–	None
Ref. [13]	44	0.05	4	0.05	None
Ref. [12]	34	0.05	3	0.03	None
Ref. [7]	21	0.02	2	0.01	None
This Work	17	0.018	1	0.004	None

4 Discussions

The cell count of half adder designs in Refs. [7–13] ranges from 21 to 77. Also, the cell area consumption ranges from 0.08 to 0.02 μm². Thus, the proposed design shows an improvement of 20% on the cell count metric. A delay of only 1 clock phase or 0.25 clock cycle is induced thereby making presented design suitable for high-speed operation. The proposed design uses no crossovers thereby can be fabricated on a single layer. The energy dissipation analysis confirmed the ultra-low-power operation of the presented circuit. The design in Reference [12] provided a breakthrough in the half adder design by significantly decreasing the cell count. However still, the design in Reference [8] exploited the majority reduction technique and employed the cell–cell interaction technique to further decrease the cell count (Table 3).

5 Conclusion

In this paper, an area- and energy-efficient half adder is presented in QCA technology. The presented QCA circuit is designed using the intercellular interaction technique. The prior proposed designs are in the range of 21–77, with cell area in the range of 0.02–0.08 μm². The presented design uses 17 QCA cells, occupying only 0.018 μm² of area. A 20% area efficiency is reported. The energy and power dissipation analysis estimated that the presented design dissipates 26.92 meV of energy. It can now


be concluded that using the cell–cell interaction technique on a small circuit can optimize the circuit by 20%. Therefore, if this technique is employed on large and more complex circuits, it can increase the efficiency by many folds.

References

1. J.A.B. Fortes, in *Future Challenges in VLSI System Design*. IEEE Computer Society Annual Symposium on VLSI, 2003. Proceedings. IEEE (2003)
2. T. Skotnicki, et al., The end of CMOS scaling: toward the introduction of new materials and structural changes to improve MOSFET performance. *IEEE Circuits Devices Mag.* **21**(1), 16–26 (2005)
3. C. Lent et al., Quantum cellular automata. *Nanotechnology* **4**(1), 49 (1993)
4. IEEE: International symposium on circuit and systems. QCA: a promising research area for CAS society (2004)
5. A. Imre et al., Majority logic gate for magnetic quantum-dot cellular automata. *Science* **311**(5758), 205–208 (2006)
6. T.N. Sasamal, S. Ashutosh Kumar, M. Anand, *Quantum-Dot Cellular Automata Based Digital Logic Circuits: A Design Perspective* (Springer, 2020)
7. A.H. Majeed et al., Full adder circuit design with novel lower complexity xor gate in QCA technology. *Trans. Electr. Electron. Mater.* 1–10 (2020)
8. S.K. Lakshmi, G. Athisha. Design and analysis of adders using nanotechnology based quantum dot cellular automata (2011)
9. H.S. Jagarlamudi, S. Mousumi, J. Pavan Kumar, Quantum dot cellular automata based effective design of combinational and sequential logical structures. *World Acad. Sci. Eng. Technol.* **60**, 671–675 (2011)
10. S. Santra, U. Roy, Design and implementation of quantum cellular automata based novel adder circuits. *Int. J. Nucl. Quant. Eng.* **8**(1), 178–183 (2014)
11. F. Ahmad, M.B. Ghulam, A. Peer Zahoor, Novel adder circuits based on quantum-dot cellular automata (QCA). *Circuits Syst.* (2014)
12. D. Ajitha, K. Venkata Ramanaiah, V. Sumalatha, An efficient design of xor gate and its applications using qca. *i-Manager's J. Electron. Eng.* **5**(3), 22 (2015)
13. M. Poorhosseini, A.R. Hejazi, A fault-tolerant and efficient XOR structure for modular design of complex QCA circuits. *J. Circuits Syst. Comput.* **27**(07), 1850115 (2018)
14. C.S. Lent, P. Douglas Tougaw, A device architecture for computing with quantum dots. *Proc. IEEE* **85**(4), 541–557 (1997)
15. K. Walus et al., QCADesigner: a rapid design and simulation tool for quantum-dot cellular automata. *IEEE Trans. Nanotechnol.* **3**(1), 26–31 (2004)
16. S. Srivastava et al., in *QCAPro-an Error-Power Estimation Tool for QCA Circuit Design*. 2011 IEEE International Symposium of Circuits and Systems (ISCAS). IEEE (2011)

Observation of Proposed Triple Barrier δ -Doped Resonant Tunneling Diode



Man Mohan Singh , Ajay Kumar, and Ratneshwar Kr. Ratnesh

Abstract The authors have observed the influence of double quantum well and δ -doping for a triple barrier AlGaAs/GaAs resonant tunneling diode (RTD) on device performance, which has been investigated by means of numerical simulation using contact block reduction (CBR) technique. The introduction of Si δ -doping in triple barrier RTD tries to dominant the transport mechanism and increases the density of electrons between the tunneling energy levels. It can also increase the energy of electrons in resonant states at a lower voltage. Consequently, the peak current and peak-to-valley current difference of RTD have been increased. The optimized values of the quantum well and the triple barrier have been observed through mapped local density of states along with the electrical and structural performance of the device. In addition to this, the transmission coefficient of the device has been optimized with varying doping concentration, barrier length, and δ -doping. The most evident electrical parameters, a peak current density of 21.7×10^3 KA/m², a peak-to-valley current ratio of 11.12, could be achieved by designing RTD with the active region structure of δ -doped Al_{0.3}Ga_{0.7}As/GaAs/Al_{0.3}Ga_{0.7}As/GaAs/Al_{0.3}Ga_{0.7}As (2 nm-4 nm-2 nm-4 nm-2 nm). The results obtained in this paper may be useful to improve the device characteristics of resonant tunneling nanostructures.

Keywords Triple barriers · Quantum mechanics · Resonant tunneling diode · Delta-doping · NDR

1 Introduction

In the last few decades, continuous scaling of CMOS technology leads to a fundamental limitation with device geometry, eccentricity to develop new device that can extend the scaling of the device and reaches to new models and technology. Quantum mechanical tunneling-based resonant tunneling diodes (RTDs) act an essential part in scaling the geometry of the device for fabrication of new generation devices [1, 2].

M. M. Singh (✉) · A. Kumar · R. Kr. Ratnesh
Meerut Institute of Engineering and Technology, Meerut 250005, India
e-mail: manmohan.singh@miet.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022
P. Verma et al. (eds.), *Advances in Data Computing, Communication and Security*,
Lecture Notes on Data Engineering and Communications Technologies 106,
https://doi.org/10.1007/978-981-16-8403-6_63

We can play with device parameters of RTD and make the device more optimized for various applications with microwave frequencies. As this device has properties like negative differential region (NDR) which can augment the MOS-based technology which leads to logic with usage of ultra-low power and applications with compact embedded design [3].

Likewise, in designing of any quantum nanostructures, parameter like operating temperature is essential part in fabrication and modeling of these devices along with microelectronic circuits [4]. Considerably, we can alter the device performance at nano-level on changing the operating temperature up to a small extent [5]. Any heterostructure could prove its performance with its obtained characteristics in entire range of operating temperature. Applications at very high switching speed needs the reliability of the devices within broader range of temperature in which device works [6, 7].

Besides, multibarrier nanostructures are highly used with resonant states in quantum devices to investigate the process of tunneling and transmission [8]. Triple barrier structures are improved over double barrier in terms of multiple peaks and coverage of application range [9–11]. Therefore, multiple barriers are essential to generate multiple peaks in J-V characteristics of quantum devices like triple barrier RTD. In logic circuits, multiple NDR regions are highly useful to generate memories and other logic circuits which are achieved with these increments of barriers in the devices [12–14].

In this article, we observed and analyses the characteristics of proposed triple barrier RTDs that include the delta-doping at a broad range of operating temperature. Delta-doping in the well increases the number of electrons tunneled into the device, and a very high current will flow, and good PVCR is achieved with multiple peaks. The full article is further expressed below in appropriate fashion. Next section discusses the proposed device structure along with computational model of RTD parameters. Third section comprises of characteristics mapping and their detailed explanation. Last section summarizes the full observation received from the other section.

2 Proposed Device Structure and Simulation Model

We use the technique named contact block reduction (CBR) to calculate the current, transmission coefficients, local density of state (LDOS) in super lattice quantum nanostructures [13]. This technique is a resourceful method that uses few leads propagate with fixed number of eigenstates for quantum devices to estimate the retarded non-equilibrium Green's function with attached external contacts of nanostructured heterostructure. From this established Green's function estimation technique, the current density, LDOS and doping profile could be obtained from the formula of Landauer with set number of leads. Ballistic limits have also been taken to perform this action of quantum transport. Device simulation package nextnano3 has been incorporated with these techniques [14, 15].

The proposed structure of RTD with triple barrier could be seen in Fig. 1; the triple

Fig. 1 Proposed RTD device structure



barriers region of 1 nm/2 nm/3 nm/5 nm along with spacer layer of 3 nm is sandwiched between the heavily doped emitter and collector regions of 12 nm, respectively. The spacers aim to prevent the impurity doping diffusing into the barrier regions during the process of epitaxial growth. The heavily doped emitter and collector region could ensure the ohmic contact of 0.5 nm between the RTD and electrodes.

In almost all the cases, carrier transport within the device is accelerated first then coherent in nature. But, relaxation energy and momentum of carriers due to an externally applied voltage are also included. Fundamentally, an arrangement of both the technique is named as Landauer-Buttiker (LB) formalization. This is proficient and nearly equal to the formalization, i.e., non-equilibrium Green’s function. With the help of mentioned LB formalization, the abridged current in inverted form is derived as given below.

$$I_{xy} = \frac{-g_s q}{h} \int T_{xy}(E) [f(E, \mu_x) - f(E, \mu_y)] dE \tag{1}$$

where g_s is termed as spin degeneracy and q as charge. I_{xy} is termed as current of the device, Planck’s constant is denoted as h and $T_{xy}(E)$ the transmission function coefficient. Chemical potential in the contacts is known as μ_x and μ_y . For this device, the Fermi–Dirac function with energy E with the contact x is conveyed as

$$f(E, \mu_x) = \frac{1}{1 + \exp[E - \mu_x]/(k_B T)} \tag{2}$$

The nanostructured heterojunction in 3D form grown along x - y - z direction homogeneously; then, the explanation in the Schrodinger calculation can deal with

$$H_{K_p}^0 \psi(z, K_p) = E_n(k_p) \psi(z, k_p) \quad (3)$$

where H is Hamiltonian function, wave function is given by $\psi(z, K_p)$ which could be resolved into clarification within z -direction. Plane wave for this wave function in perpendicular planes is given as

$$\psi(z, K_p) = (z, k_p) \exp^{ik_p \cdot r} \quad (4)$$

In above mentioned equations, dependence of K_p (parallel momentum) can be ignored within the device on $\psi(z, K_p)$. Simultaneously, one dimensional solution of Schrodinger equation can be given as

$$H^0 \psi_n(z) = E_n \psi_n(z) \quad (5)$$

$$\left[-\frac{\hbar^2}{2} \frac{\partial}{\partial z} \left(\frac{1}{m^*(z)} \frac{\partial}{\partial z} \right) + V(z) \right] \psi_n(z) = E_n \psi_n(z) \quad (6)$$

Here, tensor component of effective mass in z -direction can be given as $m^*(z)$. \hbar is Planck's constant. $E_{n,0}$ is the edge profile of conduction band within material interference. $V(z) = E_C(z) = E_0(z) - \varphi(z)$ is potential energy of variable spatial. After solving Poisson's equation, we calculate the electrostatic potential which is termed as $\varphi(z)$.

3 Characterization and Discussion

Simulation of proposed 48 nm triple barrier with δ -doping has been done in this section along with achieved characteristics discussion. The details regarding the layers and models have given in above section. This section is dividing in three sub-sections. Firstly, local density of states within proposed nanostructure has been calculated. In other subsection, analyses the effect of tunneling with transmission coefficient, and peak-to-valley currents have been discussed. Finally, we explain the current voltage characteristics of the device at different barrier length along with delta-doping.

3.1 Simulate LDOS $\rho(z, E)$ for Nanostructured Triple Barrier RTD

The local density of states (LDOS) as a function of position in the device has been shown in Fig. 2. This LDOS is calculated for different barrier lengths, and its density

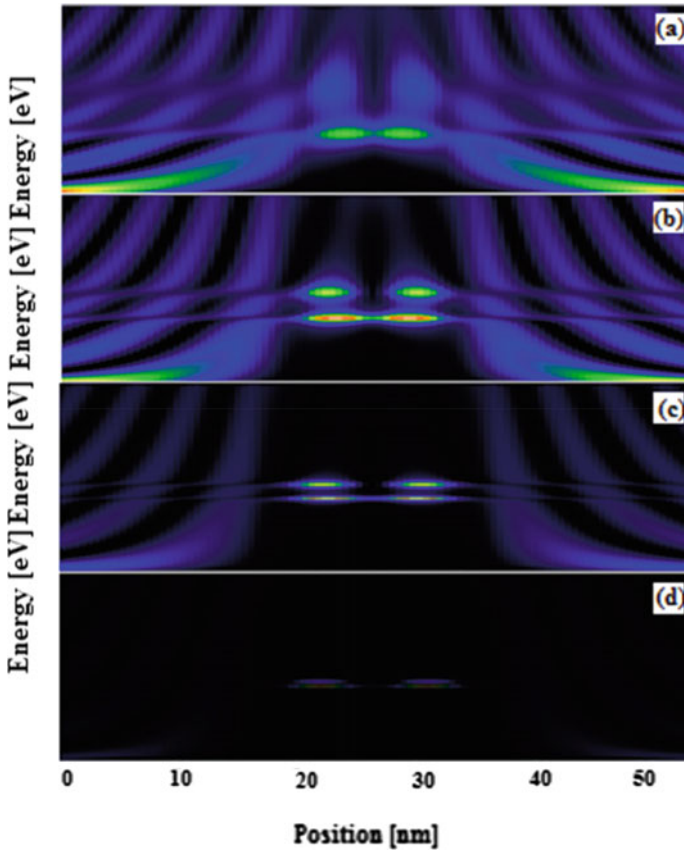
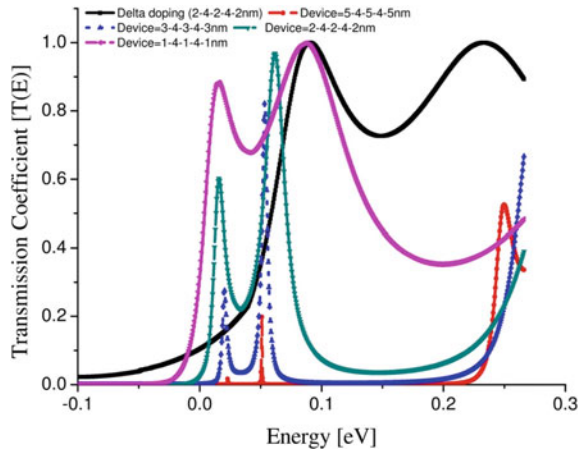


Fig. 2 Simulated LDOS $\rho(z, E)$ for nanostructured triple barrier $\text{Al}_x\text{Ga}_{1-x}\text{As}/\text{GaAs}$ resonant tunneling diode versus the energy function for different lengths of barriers: **a)** 1-nm barrier, **b)** 2-nm barrier, **c)** 3-nm barrier, and **d)** 5-nm barrier

pattern has shown in graph. Numerical calculation has used very fine energy grid with spacing of 0.5 meV to get almost all the values of sharp resonance. But still, some values are missing to achieve the transmission coefficient of unity. If resonance energy will match the grid point, then only we get the peaks as shown in graph. And this energy is the same energy at which tunneling exists within the RTD structure. With 2-nm barrier, we achieve highest value of calculated density of states (i.e., 432 eV^{-1}) as 5-nm barrier reaches to lowest energy density of 144 eV^{-1} . As we compared with existing work [10], there is no description about density of states, but here, we calculate this in the form of presented pictogram (Fig. 2).

Fig. 3 Calculated transmission coefficient versus energy function graph at different device structures



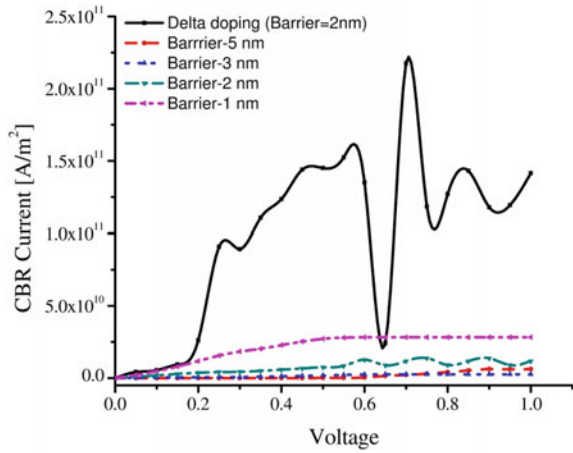
3.2 Transmission Coefficient as a Function of $Al_xGa_{1-x}As/GaAs$ Nanostructure

Figure 3 provides the graph of transmission coefficient versus energy function for the proposed nanostructure at varying the barrier lengths and its δ -doping structure. With increasing barrier length, effective masses within the device mismatches with it. Due to this mismatch, the tunneling probability has decreases as expected and shown in Fig. 3. 2-nm barrier length has the sharp peak (i.e., $T(E) = 1$) as compared to other device structures. But with δ -doping, value of tunneling carrier has been increased, and it provides the value of transmission coefficient nearest to unity as calculated through the extracted plot.

3.3 Current Voltage Characteristics at Altered Barrier Length and Nanostructure δ Doping

Figure 4 shows the graph of current density versus operating voltages at different values of barrier lengths and the δ -doping. Tunneling current should be increased as the energy difference decreases between the resonant levels which are done with the help of δ -doping. Multiple maxima and minima have been achieved with the usage of triple barrier as shown in graph. Here, we have achieved current density of maximum value, i.e., 21.7×10^3 KA/m² and a PVCR of 11.12. This is achieved at the optimized value of barrier length, i.e., 2 nm. On comparing with the result of [12], we have worked on different layered structured device along with delta-doping.

Fig. 4 J-V characteristics of different barrier lengths and Si-delta-doping



4 Conclusion

Our observation suggested that the introduction of δ -doping in proposed TBRTD has drastic changes in terms of peak current density to a maximum value of $21.7 \times 10^8 \text{ KA/m}^2$ at optimized values of 2-nm barrier length. In addition, the performance parameters like local density of states, transmission coefficient, and electrical characteristics could promote the all over performance of the triple barrier RTD device through the numerical simulation. The influence of δ -doping on these device parameters is also discussed successfully with the help of the simulated graphs. It is evident that by modifying the concentration of Si-delta-doping can expressively rise in the forward current and unity transmission coefficient. Moreover, optimizations of parameters length are also done with the help of varying their values within permissible limits. Additionally, studies with different material structures used to model the triple barrier RTD are requisite in future to make these devices more projecting.

Acknowledgements Special thanks to Dr. S. Birner, Nextnano GmbH, Munchen, Germany to provide the simulation package of Nextnano3 for this research.

References

1. A. Ramesh et al., Boron delta-doping dependence on Si/SiGe resonant interband tunneling diodes grown by chemical vapor deposition. *IEEE Trans. Electron Devices* **59**(3), 602–609 (2012)
2. L.K.S. Herval et al., in *Circular Polarization in n-Type Resonant Tunneling Diodes with Si Delta-Doping in the Quantum Well*. 29th Symposium on Microelectronics Technology and Devices (SBMicro) (Vol. 29, 2014), pp. 1–5
3. S.Y. Park et al., Si/SiGe resonant interband tunneling diodes incorporating delta-doping layers grown by chemical vapor deposition. *IEEE Electron. Device Lett.* **30**(11), 1173–1175 (2009)

4. A. Pfenning et al., Nano thermometer based on resonant tunneling diodes: from cryogenic to room temperatures. *ACS Nano* **9**, 6272–6277 (2015)
5. W. Lu, C.M. Lieber, Nanoelectronics from the bottom up. *Nat. Mater.* **6**, 841–850 (2007)
6. A. Taube et al., Temperature-dependent electrical characterization of high-voltage AlGaIn/GaN-on-Si HEMTs with Schottky and ohmic drain contacts. *Solid-State Electron.* **111**, 12–17 (2015)
7. M. Bhattacharya, J. Jogi, R.S. Gupta, M. Gupta, Impact of temperature and indium composition in the channel on the microwave performance of single-gate and double-gate InAlAs/InGaAs HEMT. *IEEE Tran. Nanotechnol.* **12**(6), 965–970 (2013)
8. C. Allford et al., Thermally activated resonant tunnelling in GaAs/AlGaAs triple barrier heterostructures. *Semi. Sci. and Tech.* **30**(10), 105035 (2015)
9. M. Asada, Y. Oguma, N. Sashinaka, Estimation of interwell terahertz gain by photon-assisted tunneling measurement in triple-barrier resonant tunneling diodes. *Appl. Phys. Lett.* **77**, 618–620 (2000)
10. M.M. Singh, M.J. Siddiqui, Electrical characterization of triple barrier GaAs/AlGaAs RTD with dependence of operating temperature and barrier lengths. *Mater. Sci. Semicond. Process.* **58**, 89–95 (2017)
11. S. Suzuki, M. Shiraishi, H. Shibayama, M. Asada, High-power operation of terahertz oscillators with resonant tunneling diodes using impedance-matched antennas and array configuration. *IEEE J. Quantum Elec.* **19**(1), 8500108 (2013)
12. M.M. Singh, M.J. Siddiqui, in Effect of Si-Delta Doping and Barrier Lengths on the Performance of Triple Barrier GaAs/AlGaAs Resonant Tunneling Diode. *IEEE Int. Conf. Elec. Devices and Solid-State Circuits.*, V0. 12 (2016) , pp. 30–34
13. S. Birner, C. Schindler, P. Greck, M. Sabathil, P. Vogl, Ballistic quantum transport using the contact block reduction (CBR) method. *J. Comp. Electron.* **8**, 267–286 (2009)
14. S. Birner, T. Zibold, T. Andlauer, T. Kubis, M. Sabathil, A. Trellakis, P. Vogl, Nextnano: general purpose 3-D simulations. *IEEE Trans. Electron Dev.* **54**(9), 2137–2142 (2007)
15. <http://www.wsi.tum.de/nextnano3>; <http://www.nextnano.de>

A Node-RED-Based MIPS-32 Processor Simulator



Ethan Anderson, S. M. Abrar Jahin, Niloy Talukder, Yul Chu,
and John J. Lee

Abstract Processor simulators are imperative tools that well facilitate the understanding of modern processors. Therefore, numerous attempts have been made to develop better simulators, and some of them have been very widely used. There exist several categories of such simulators in terms of simulation speed, cycle accuracy, functional validation, cache focus, multiprocessor target, behavioral visualization, and education purpose. Our recent study focuses on developing a simulator with the following objectives: (i) to help students understand the organization and operation of processor faster and (ii) to provide students with much easier ways to build their own simulators while learning. Along the study, this paper describes our recent development of a Node-RED-based MIPS-32 processor simulator. Its functionality includes 5-stage pipeline visualization, 2-phase clocking (i.e., mimicking master/slave behavior), various cache configuration, cache statistics visualization, operand forwarding for the resolution of data dependency, and branch prediction mechanisms. Our study demonstrates the feasibility of good simulator implementations using Node-RED.

E. Anderson (✉) · S. M. Abrar Jahin · N. Talukder · J. J. Lee
IUPUI, Indianapolis, IN, USA
e-mail: ethander@iu.edu

S. M. Abrar Jahin
e-mail: ajahin@iu.edu

N. Talukder
e-mail: ntalukde@iu.edu

J. J. Lee
e-mail: johnlee@iu.edu

Y. Chu
UTRGV, Rio Grande Vally, TX, USA
e-mail: mark.chu@utrgv.edu

1 Introduction

Students studying computer engineering, especially processor architecture, are often exposed to the MIPS ISA [1] as an educational medium for learning processor and system architectures. MIPS is a preferred choice while teaching CPU architecture due to its relative simplicity in comparison to other processor architectures such as ARM and RISC-V. As well-known practices, the best learning method is actually building something that one is learning. However, since it is not feasible to build a processor in a course, another way is to build its simulator or to utilize existing simulators at least. There have been several popular simulators. Gem5 [2] is one of the most popular simulators that supports various ISAs and cache hierarchy, and it even supports running operating systems. However, it requires steep learning curve, and thus, it is not a good fit for education purpose. MIPT-MIPS [3] is a cycle accurate pre-silicon simulator of MIPS and RISC-V CPU that has been developed recently, but as it is written in C++, thus, it requires a good knowledge of C++. One of the best educational processor simulators is the MIPS Assembler and Runtime Simulator (MARS [4]). However, MARS was written in Java language and old, and thus, it is hard to modify or integrate a new design. This may limit the students' knowledge of computer architecture. After considering aforementioned various aspects, we decided to create a MIPS processor simulator utilizing the well-known Node-RED programming environment [5] due to the following features provided.

- Event-driven, non-blocking model (good fit for processor simulation)
- Synchronized message passing
- GUI-based drag-and-drop framework (easier implementation)
- Dashboard feature (easier user interface and easier visualization)
- Module-based or node-based (good for modular design as it is analogous to hardware modules)
- Browser-based (easier usage)
- A large number of libraries and pre-built nodes (modules) openly available.

The motivation toward the development of this simulator was to demonstrate how difficult or easy to develop a reasonably good simulator, and to possibly assist future students in understanding the functionality of a MIPS pipeline with various enhancements. After considering several processor components and their importance of each, the following feature list was selected as the objectives for our current implementation, which are also our contributions along with Node-RED basis.

- ISA including branch, jump, multiply, and divide
- Data dependency detection and forwarding
- Branch prediction
- Memory and cache simulation
- Organized and intuitive user interface

The rest of this paper details the architecture and operations of our simulator with emphasis on various functionalities.

2 Pipeline Architecture

The simulator developed over the course of this project attempts to accurately simulate hardware behavior, within the abstract environment provided in Node-RED [5]. Node-RED is based on Node.js [6], which in turn utilizes JavaScript. Consequently, code snippets are written in JavaScript, using the abstractions and structures provided by Node-RED. The most notable feature that Node-RED provides is the ability for function blocks to synchronously pass ‘messages’ to each other. Messages can even be passed to multiple other functions in a single step, facilitating emulation of parallel execution. In addition to messages, data can also be passed between function blocks asynchronously using shared variables. These two methods of passing data form the basis for the architecture simulation and drive all pipeline events.

As shown in Fig. 1, the architecture on which our study is based utilizes the previously discussed ‘message’ mechanism in Node-RED to emulate clock cycle functionality. This is performed by creating a node with an infinite loop which sequentially generates ‘rising edge’ and ‘falling edge’ messages. These messages are passed to each of the pipeline stages simultaneously, allowing synchronous progression of data through the pipeline. Each stage of the pipeline is split roughly in half, with some operations being performed upon receiving a ‘rising edge’ message, and the rest being performed on the ‘falling edge’. Generally, the pipeline stage performs work during the rising edge and writes results for the next stage during the falling edge. This emulates the functionality of pipelining behavior of a real processor and prevents the introduction of race conditions depending on the exact order in which stages run/complete.

To facilitate communication between stages, shared ‘buffer’ variables are used which can be read and written by multiple stages. During the rising edge, the output

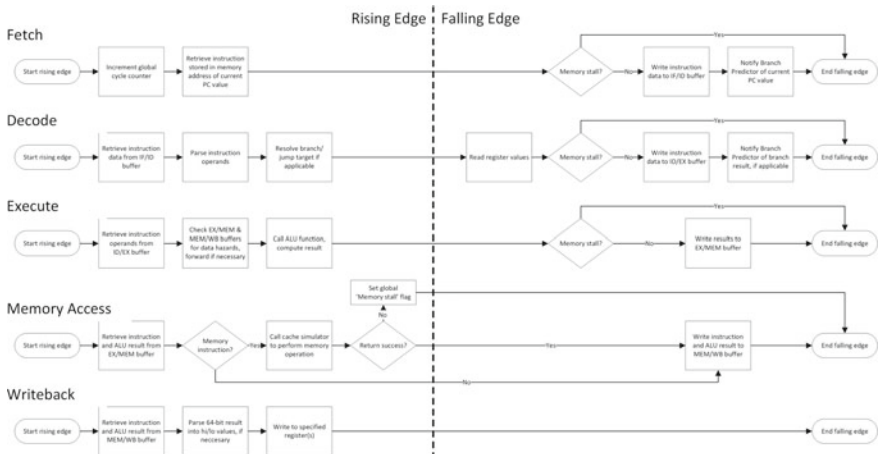


Fig. 1 Operation flowchart of our 5-stage pipelined CPU simulator

buffer of the previous stage is read and processed. On the falling edge, the calculated result is stored to the stage's output buffer, which provides input to the next stage. This mimics a real processor, which would use an array of latches to store results between stages. In our implementation, the data passed between stages is represented as a JavaScript object with attributes added by each stage depending on the work performed. For example, the decode stage adds attributes denoting the opcode and operands, while the execution stage adds a 'result' attribute. In this way, each stage always has access to previously calculated values while only using a single buffer to pass data. This data object also contains the address of the source instruction, and other text fields are displayed in the UI for appropriate visualization.

3 Branch Prediction

The branch prediction mechanism implemented functions more as a control unit, rather than a standalone branch predictor (BP) and, thus, plays an important role. The BP is not clocked to the main system; it instead waits for messages from other functions and then reacts accordingly. The two primary sources of information for the branch predictor are the Fetch and Decode stages. The messages serve to keep the BP up to date and allow it to make timely decisions. The first way the BP receives information is from the Fetch stage, which sends a message containing the current PC on the falling edge of each clock cycle. The message tells the BP that the clock cycle is nearly finished, and a new PC value must be selected. As shown in Fig. 2, upon receiving this type of message, the BP consults the Branch History Table (BHT) and either increments the PC to the next instruction or sets the PC to the cached branch target address. Because messages in Node-RED can trigger synchronous execution, the Fetch stage triggers the BP unit during the falling edge, and it will always finish selecting a new PC before the start of the rising edge.

The second type of messages which the BP relies on is branch resolution messages from the Decode stage. These messages indicate the address of a branch instruction and whether or not the branch was taken. This information is used by the BP to create or update the BHT and to correct mispredicted branches as needed. Each time the BP receives a message from the Decode stage, it updates the corresponding BHT entry for the specified address. Various branch prediction modes are implemented and concurrently maintained in the BHT in our simulator (details are shown later). After updating the BHT, the BP checks to see whether the branch is predicted correctly or not. If a misprediction is detected, the PC is corrected, and the outputs of the Fetch stage (which has just retrieved an errant instruction) are cleared. This effectively flushes the pipeline, as the incorrect branch decision only affects a single cycle.

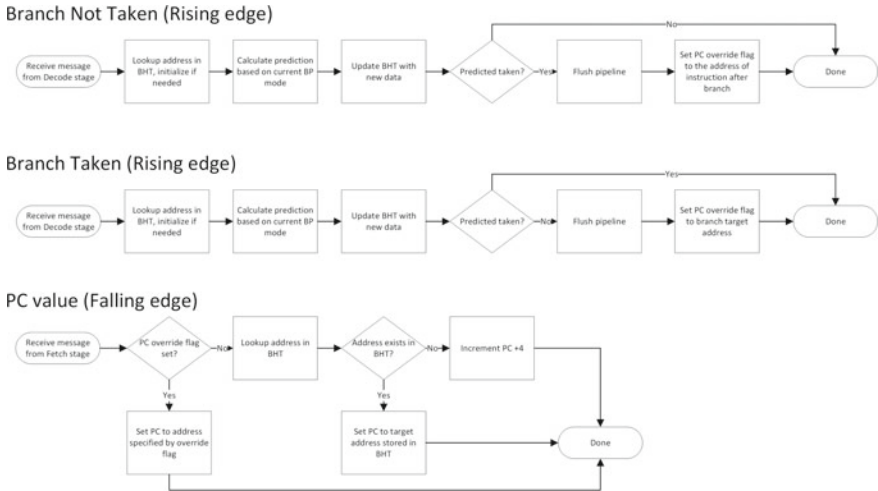


Fig. 2 Operation flowchart of branch predictor

4 ISA Implementation

4.1 Memory Model

For the implementation of memory model with cache, to build easiest and most efficient support of memory and cache accesses from ALU, we decided to use a table-structured unified memory. That is, the memory is implemented as a single variable in a node, with updating all functions within the node accordingly. This variable is structured as an object with named attributes, each corresponding to a memory address. This allowed us to use format `memvar[address]` and access arbitrary addresses rather than sequential addresses. In addition, an interpreter function was created which constructs the memory object from a text file containing key-value pairs of address and contents. The single text file includes both the program instructions as well as any data which need to be loaded prior to starting the program.

4.2 Support for Jump Instructions

Implementation of jump instructions was done by following the definition of jump instruction as provided by the MIPS specification [1]. The resulting address is formed by concatenating 4-bit MSB's from the PC, 26-bit absolute target address provided by the instruction, and two zeros for word-distance jumps. We implemented the 'j' and 'jal' instructions, where 'jal' instructions when executed store the current PC to register 31, which is used for return purpose.

4.3 Register File Implementation

Originally, the Decode stage of the pipeline would resolve register values on the rising clock edge. Although this aligns with the other stages, which perform all work on the rising edge, it does not accurately represent hardware. Furthermore, because the write-back stage writes to registers on the rising edge, inconsistent results could occur if a register is written and read within the same clock cycle. In order to resolve this issue, modifications were made to the Decode stage, which delays register resolution until the falling edge. This accurately simulates hardware and allows for a register to be written and read during the same clock cycle with consistent results. This change is reflected in Fig. 1.

5 Cache Simulation

When the pipeline needs to access memory, the cache simulation node ('cache simulator') will be activated and involved. Its implementation is illustrated in Figs. 3 and 4. Since it can process one memory access at a time, if another memory access is already being processed in the cache, a new access should not be allowed to use the cache. Next step is to check if the cache has been initialized. If cache is not initialized, it is initialized for the first-time use. For initialization, it collects cache configuration from the user input (UI portion of Node-RED, described in the UI section) and sets the configurations into cache simulator's configuration variables (if nothing is selected, default values are used). If the cache has been already initialized, the logic extracts necessary information including what type of memory operation (read or write), memory address, and data if the operation is a write. After obtaining those, it splits the address into following fields: tag, index, and offset.

Next step checks if requested word is in the cache (cache hit) or not (miss), and the operation depends on the current configuration of caching mode/types, which is illustrated in Fig. 4. Thus, the next step of action will differ based on the cache types. If it is a direct mapping cache, it will directly find the line with index. When line is found, it will be updated with new data. If miss occurs and the line has modified data, it is written into memory first and then filled with a new memory block. After that, the remaining process will resume.

If the cache is configured as a set-associative cache, the set can be accessed with the index field, as shown in Fig. 4. Then, we need to select and replace data in one of the lines for which we need a replacement policy. Our simulator also supports four kinds of replacement policies as follows: random, least recently used (LRU), last in first out (LIFO), and first in first out (FIFO).

For implementing the replacement policies, we are using two extra properties in our code for simplicity, which are `insertion_time` and `last_access_time`, thereby finding our replacement line easily. When we get the address to replace, then we will drop or perform write back the data and insert the new data in the location.

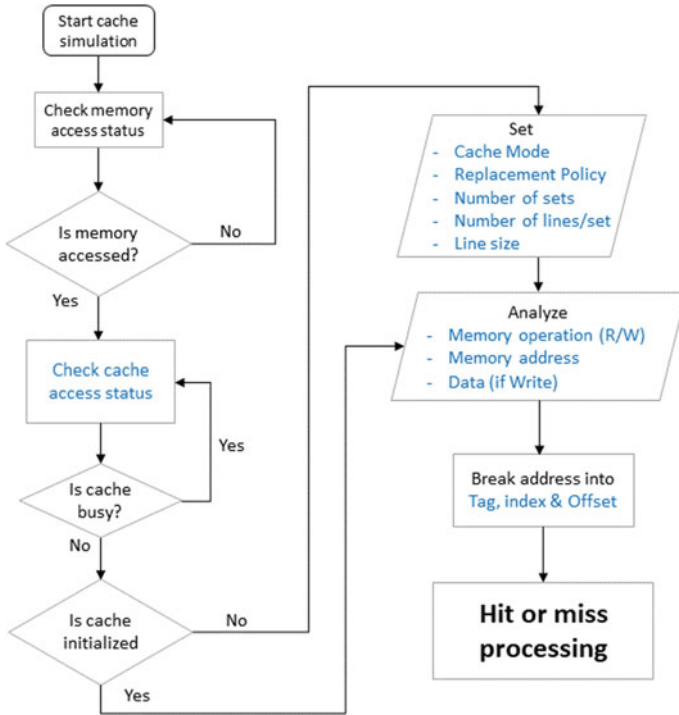


Fig. 3 Operation flowchart of cache simulation—part 1

If the cache is configured as a fully associative cache, it checks if any line is available. If available, it simply chooses the line and accesses the data. If all lines are fully occupied and cache miss occurs, it needs to replace a line and find the replace location. It follows one of the replacement policies from the UI selection. Once the replacement location is chosen, it will if necessary perform write back and insert new data into the location.

6 User Interface

The user interface (UI) of our simulator is composed of three tabs: pipeline tab, branch prediction tab, and cache tab. These are separated in different windows. We will discuss components of each tab in the following subsections.

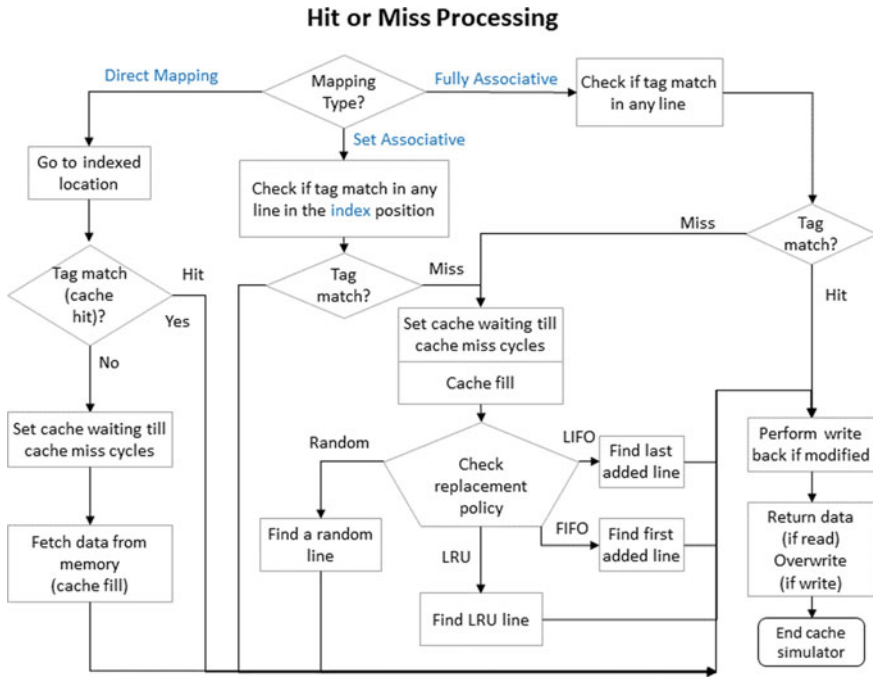


Fig. 4 Operation flowchart of cache simulation—part 2

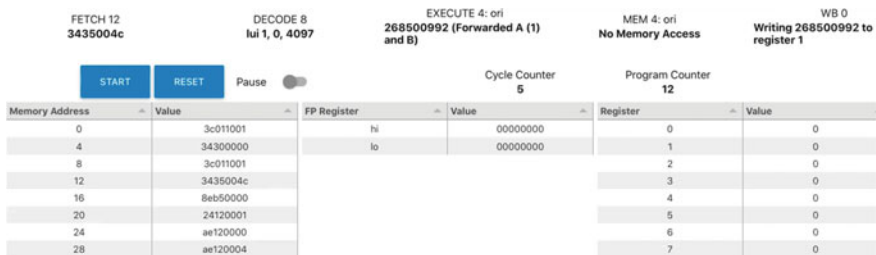


Fig. 5 Elements of pipeline visualization tab

6.1 Pipeline Tab

First tab is pipeline tab. Its elements are shown in Fig. 5. The top row has the view of five stages side by side and shows the names of the stages along with instruction address above with the machine code, disassembled instruction, operands, and computation result below.

Branch Prediction



Fig. 6 Elements of the branch prediction tab

The second row of Fig. 5 shows the RESET button that can be used to initialize the simulator. Also, the START button will allow to start the execution. The PAUSE radio button is used to pause the execution. Cycle counter increments by 1, and program counter increments by 4 if an instruction can be fetched. In the third row, three tables are shown. First one is memory table which shows unified view of instruction and data memory. During execution, this table is updated as results are added to the memory. Second one is floating point register table which shows register values used to store multiplication and division results. Third one is register table which stores the latest values of integer registers.

6.2 Branch Prediction Tab

Figure 6 depicts the branch prediction tab with left side being prediction strategy selector, right side showing Branch History Table (BHT), and bottom side predictor messages field. Left side has a dropdown list from which a user can select a prediction strategy from the following options: 1-bit, always taken, always not taken, 2-bit saturating, and 2-bit hysteresis. By default, ‘always not taken’ is initialized. The BHT stores the latest predictions for different strategies. The predictor messages field displays what actually happened for a recent branch, what was predicted, and if pipeline was flushed due to misprediction.

6.3 Cache Tab

There exist two cache tabs: One being used to select different cache parameters and the other for the visualization of cache performance. Users can use the first cache tab (not shown due to space limit) to select various cache configurations as follows.

- Cache Mode: Direct mapping, N-way set associative, and fully associative
- Number of Sets
- Number of Lines: Choice 1 through 1024 in power of two increment
- Line Size (# words): Choice 1 through 1024 in power of two increment
- Replacement Policy: random, LRU, FIFO, and LIFO

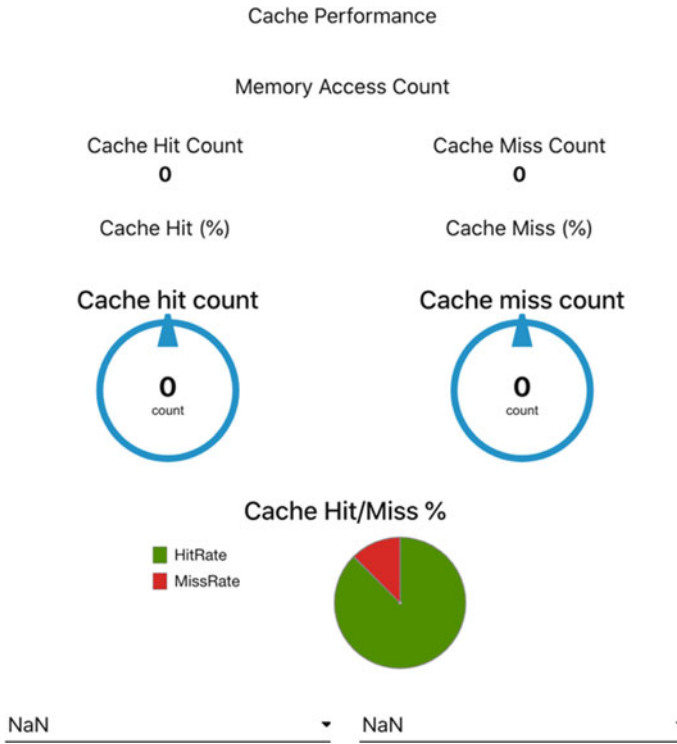


Fig. 7 Elements of the cache tab—statistics

The second cache tab (Fig. 7) delineates the performance statistics of cache behavior. The statistics include total memory access counts, hit and miss counts, their percentages, and hit and miss percentages in a pie chart.

7 Conclusion

In this paper, we presented a novel Node-RED-based MIPS-32 processor simulator with detailed explanation of key modules and a variety of configurations. The simulator is expected to be very helpful to students learning processor architecture. Additional enhancements can be implemented in the future, extending the simulator to include all materials covered in the computer architecture subjects. However, even without further improvement, the simulator is reasonably complete and allows users to experiment with many different programs and processor configurations.

References

1. mips.com, MIPS Architecture For Programmers Volume 1-A, Available via DIALOG. <https://s3-eu-west-1.amazonaws.com/downloads-mips/documents/MD00082-2B-MIPS32INT-AFP-06.01.pdf>. Accessed May 7, 2021
2. N. Binkert, B. Beckmann, G. Black, S.K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D.R. Hower, T. Krishna, S. Sadashti, R. Sen, K. Sewell, M. Shoaib, N. Vaish, M.D. Hill, D.A. Wood, The gem5 simulator. *ACM SIGARCH Comput. Archit. News* **39**(2), 1–7 (2011). <https://doi.org/10.1145/2024716.2024718>
3. MIPT-ILab, “MIPT-ILab/mipt-mips: Cycle-accurate pre-silicon simulator of RISC-V and MIPS CPUs” <https://github.com/MIPT-ILab/mipt-mips>. Accessed June 10, 2021
4. K. Vollmar, P. Sanderson, MARS: an education-oriented mips assembly language simulator. *SIGCSE* **6**, 239–243 (2006)
5. IBM, Node-RED. <https://nodered.org/>. Accessed May 7, 2021
6. NodeJS. <https://nodejs.org/>. Accessed June 10, 2021

Simulating Modern CPU Vulnerabilities on a 5-stage MIPS Pipeline Using Node-RED



Samuel Miles, Corey McDonough, Emmanuel Obichukwu Michael, Valli Sanghami Shankar Kumar, and John J. Lee

Abstract This paper proposes a simulation of the 5-stage pipelined MIPS processor using Node-RED and illustrates the basic effects of modern CPU vulnerabilities. Demonstrated in this study are Spectre vulnerability attack and load value injection (LVI) transient-execution attack. The storing of secret data within the cache is shown for Spectre, and through the use of an attacker's injected page number after a page fault has occurred, we demonstrate LVI's ability to access the host secrets via simulated memory hierarchy. The persistence of the secret data in the cache can also be observed in the case of both attacks. The characteristics of such security vulnerabilities are successfully simulated with the proposed Node-RED-based processor simulator.

1 Introduction

Transient-execution CPU vulnerabilities are defined as attacks in which a speculative execution optimization within microprocessors can be used by an attacker to access secret information. Due to the parallel nature of modern computers, if an operation is not able to be performed due to a previous operation that is taking more time and has not yet completed, the microprocessor may try to predict the result of the

S. Miles (✉) · C. McDonough · E. O. Michael · V. S. Shankar Kumar · J. J. Lee
Department of Electrical and Computer Engineering, IUPUI, Indianapolis, IN 46202, USA
e-mail: sammiles@iupui.edu

C. McDonough
e-mail: cojomcdo@iupui.edu

E. O. Michael
e-mail: emobmich@iupui.edu

V. S. Shankar Kumar
e-mail: vshanka@iupui.edu

J. J. Lee
e-mail: johnlee@iupui.edu

aforementioned previous operation. This is called speculative execution and is what enables many modern CPU vulnerabilities, such as Spectre [1] and Meltdown [2], which are capable of accessing secret information. Many transient-execution attacks abuse transient instructions to encode unauthorized data in a victim program. In this study, we demonstrate the effects of two modern CPU vulnerabilities by showing how speculative and injection-based attacks can load secret information into cache using a processor simulator built with Node-RED [3].

2 Prior Work

Recent research works on transient-execution attacks have predominantly focused on miss-prediction (diverting control flow) and data extraction-type attacks (exploiting illegal data flows). Several studies talk about using page faults or microcode assists to obtain data from various micro-architectural elements such as caches [4] and store buffers [5]. Others discuss the way that transient-execution attacks have evolved [6] and what defenses exist [7]. However, these studies primarily focus on the breadth of capabilities for the attacks by fully dictating the attack vector, rather than clearly illustrating and demonstrating some of these effects in a tangible and intuitive format. Specifically, the idea of simulating Spectre-type vulnerabilities is discussed by researches in [8] and is not new, but there is no current works using Node-RED to simulate CPU vulnerabilities. The goal of this study is to provide a visual, evidence-based framework for illustrating the effects of these exploits in a novel and intuitive way. Specifically, our contributions include:

- Simulation of a Node-RED-based 5-stage pipelined MIPS processor.
- Demonstration of the effects of modern CPU's transient-execution vulnerabilities.
- Simulation of the two-phase operation of a real CPU by using both rising and falling edges of a clock.

3 Simulator Implementation

The implementation is broken into sub-sections that discuss the details for each aspect of the pipeline as well as the additional pieces of the simulator that were developed in order to properly replicate the necessary functionalities for vulnerability demonstration. The diagram for the proposed simulator is shown in Fig. 1.

3.1.2 Fetch

During the rising edge, the Fetch stage of the pipeline reads the current PC register value from flow memory. The block verifies that code is loaded and the program has not finished execution. It then fetches the next instruction from the code relative to the PC value and sets the data into the pipeline register during the falling edge.

3.1.3 Decode

During the rising edge, the node reads the instruction register written by the previous stage and parses the information. This includes OP code, instruction type (i, j, r), rs, rt, rd , immediate, and address fields. During the falling edge, the parsed instruction is moved to the pipeline buffers. If the instruction is a branch, the node determines if the branch is predicted to be taken or not. The PC predictor evaluates miss-predictions to update the Branch History Table (BHT). If the Spectre is enabled, resulting prediction information is sent during the write-back stage instead.

3.1.4 Execute

During the rising edge, the data written by the Decode stage is read and executed. Forwarded operands from the execute/memory and memory/write-back stages replace values if a data hazard is detected. On the falling edge, the results are written to pipeline buffers for the next stage.

3.1.5 Memory and Cache

In the memory stage, for store and load word (SW/LW) operations, during the rising edge, the address to memory access passed from the execute stage is read. Then, the cache is checked first. A fully associative write-through cache with least recently used (LRU) replacement scheme has been implemented. If the address is found in the cache, the cache is accessed. For SW, the cache and memory are updated. For LW, if miss occurs, the cache checks if there is an empty block. If cache is full, blocks are replaced based on the LRU scheme. In case of cache miss, the system will check the simulated RAM storage for the specified virtual memory address. We implement a small amount of simulated memory that can be directly mapped into four equal-sized pages. In case of RAM miss, a page fault will occur and the desired page will be loaded from the simulated DISK. The memory hierarchy is simulated using RAM and DISK storage representations. The RAM is set to be 1KB, with a 4KB DISK. The DISK is initially loaded with data from an external text file where most entries are set to 0 except for some *TRUSTED* and *SECRET* values, which are represented as strings containing the representative text 'TRUSTED' and 'SECRET' for the purposes of simulating load value injection attacks, located at

specified physical memory addresses used in the simulation. Cache hit/miss statistics as well as declared memory are shown in the UI and updated every clock cycle.

3.1.6 Write-back

During the rising edge, the data from the memory stage is read. If the instruction writes back to memory, the result is written back. If the Spectre attack is enabled, the branch calculation determined during the Decode stage is sent to the PC predictor. No results are written to pipeline buffers during the falling edge because the pipeline is complete.

3.1.7 Hazard Detection and Stalls

The hazard detection block receives information from the Fetch and Decode stages. The node parses necessary information from the fetched instruction and determines if there is a true dependency caused by a LW instruction. If a dependency is detected, a stall signal is sent to the PC predictor. The PC predictor will not increase the PC for a stall cycle and will override the value in the current Fetch stage's register with a stall instruction. This will be decoded next cycle, and future stages become idle when they read the instruction. If the Spectre attack is enabled, the PC predictor will also flush the decode, execution, and memory pipeline buffers.

3.1.8 PC Predictor

The PC predictor receives messages from the Fetch stage informing the node to continue updating/predicting the PC. If the instruction address is in the BHT, the predictor will update the PC based on the prediction strategy selected. The PC predictor also receives branch messages from the Decode stage stating the results of branch calculations (calculated taken or not taken) and the address of the instruction. The node updates the BHT and flushes the Fetch stage's register if the branch was miss-predicted. Various branch prediction strategies can be chosen between. On hit/misses, the BHT is updated with the most current predictions for each strategy. On first encountering a branch instruction, the program will always predict not taken and create a new entry in the BHT using the branch address.

3.1.9 Simulation Settings

The simulator gets input instructions as a file of hexadecimal instructions separated by newlines. A program file can be created using a MIPS-32 assembler provided by MARS [10] and was used for this purpose. The simulator can start or stop with a button and can be resumed from a stop. A stepping function is present to step

through one instruction at a time to better visualize the pipeline. A reset button clears the memory, registers, cache, pipeline, BHT, and reloading the input files. The branch prediction strategy defaults to 1-bit prediction, and the attack settings default to off. Simulating either attacks can be done through the dropdown menu in the UI. Both the active and inactive states for each attack can be demonstrated. Enabling Spectre activates the delay branch prediction setting in the system. This forces branch resolution to occur during the write-back clock cycle. LW and SW operations can alter the cache if executed directly after a miss-predicted branch instruction. Any write to memory will be invalidated if the branch was miss-predicted. In LVI, the malicious page number is injected to load data onto RAM on a page fault. Data from outside the allowed access page for the process is loaded into RAM and the cache, which allows attackers to access the host *SECRET* information.

4 Simulation of Vulnerabilities

Spectre attack and load value injection (LVI) attack on MIPS processor are vulnerabilities that affect modern processors through techniques like branch prediction and out-of-order execution that were originally designed for performance purposes.

4.1 Spectre

Spectre attack is a security vulnerability that was discovered as part of Google's project *Zero* which was launched in 2018 to identify security issues with modern day processors [1]. Spectre exploits speculative execution on modern processors to steal secret data [1]. Speculative execution utilizes branch prediction to execute a set of instructions and avoid stalling while trying to decide if a branch is taken or not. This is used whenever there is a branch or when there is a control dependency between the branch instruction and an earlier instruction that has not written back to a register. The speculated instructions are flushed on a branch miss-prediction. However, footprints left in the processor allow for the Spectre exploits. A simple Spectre attack involves training a branch to access data from an array with a check on the array size. Once the attacker is confident that the branch is sufficiently trained to always access this array, it then uses this behavior to try to access secret data in an unreachable memory space using the array source address. Since the branch is always taken, the data is accessed and loaded in the cache before the system realizes that the branch was miss-predicted and flushes the pipeline leaving the cache untouched. A series of timed cache accesses is then used to grab this secret data from the cache. Spectre attacks can occur in virtually any computing environment.

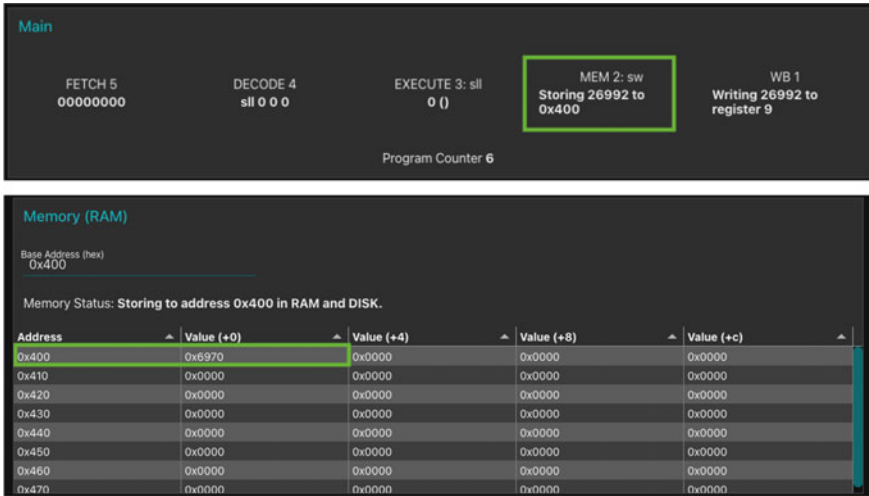


Fig. 2 Secret data is stored by the program in memory address 0×400

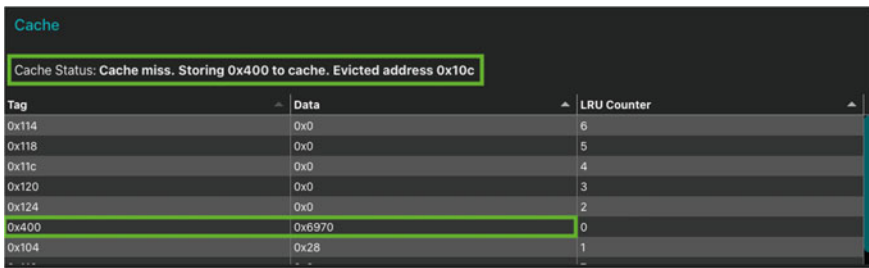


Fig. 3 Secret data is cached by the Spectre attack

As demonstrated in Fig. 2, secret data (0×6970) is initially stored at address 0×400 . An array of size 10 is initialized (array is set to all zeros for the demo but can be set to any value) starting from the base memory address 0×108 . A loop is used to train the array access branch that checks the array access index to be less than ten. The looping check uses branch prediction to train the BHT entry to 'always taken'. As the array size is always flushed from the cache, memory access for store operations will require additional clock cycles, creating a window for the Spectre attack on a miss-prediction. This loop is run four times, allowing array access and giving the process the false impression that array access is always going to be taken. After a miss-prediction, a full pipeline flush occurs, but the secret is already cached. Figure 3 shows the outcomes of the Spectre attack simulation.

4.2 Load Value Injection (LVI)

LVI attack reversely exploits Meltdown-type micro-architectural data leakage that was discovered by researchers in 2019. Meltdown is a vulnerability allowing a process to read all memory in a given system [2]. This attack was initially thought to target only Intel SGX enclave, but authors in [11] have shown that LVI-type attacks are applicable in other domains. LVI attack is a lethal attack since it mimics certain existing attacks, but none of those attack's defenses can actually mitigate it. LVI combines Spectre-style code gadgets with Meltdown-type illegal data flows to bypass existing defenses, but none of the existing Spectre/Meltdown defenses can mitigate LVI. LVI is one of the most advanced exploitation techniques presented to date since it combines page table manipulations and invalid transient executions. An attacker can use poisoned data loaded into various micro-architectural buffers in the legitimate victim program. It is shown that these buffers can be injected with a value in advance [11]. Recent research also suggests the existence of different LVI variants [11]. Some suggest an LVI-type attack to hijack the control flow through an injection of a poisoned address upon returning from a branch or jump instruction. This allows the attacker to execute malicious code on the host system, further illustrating the dangers of LVI attacks. By targeting load-type instructions, LVI drastically widens the spectrum of incorrect transient paths. In order to successfully exploit LVI, it needs to be possible to induce a page fault or microcode assists during execution on the host. LVI attacks are said to occur in three phases:

- Micro-architectural poisoning,
- Faulting loads,
- Secret data transmission.

Figure 4 shows the first step in LVI, where it can be observed that the cache is loaded with the *SECRET*, as defined in Sect. 3.1.5 previously. Figure 5 shows that even after another page fault occurs and resets the RAM, it remains in the cache.

The system is designed such that typically when a page fault occurs, the page that is being used to map the storage changes based on what physical address is desired. In the case of simulating LVI, the page number that normally is calculated based on the desired physical address is instead replaced with an attacker's page number that then allows for memory access outside of the currently accessible process to be breached. Through a page fault, the attacker can access *SECRET*, and even upon an additional page fault where RAM is cleared, *SECRET* is still accessible to the attacker from the data cache. Currently with LVI-type attacks, mitigation requires the serialization of the processor pipeline with 'lfence' (load fence) [12] instructions after every memory load. Research shows these defenses lead to performance penalties anywhere from 2x to 19x loss [13]. Intel has developed LLVM-based compiler pass which inserts lfence instructions in an optimal way to reduce overhead, but even this technique has proven to introduce more than a negligible penalty [11].

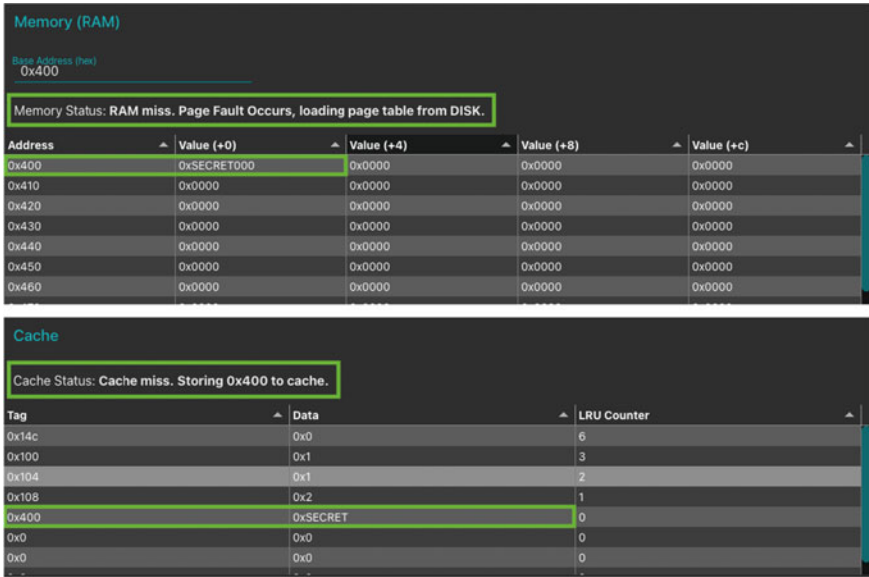


Fig. 4 LVI loads secret data into the cache

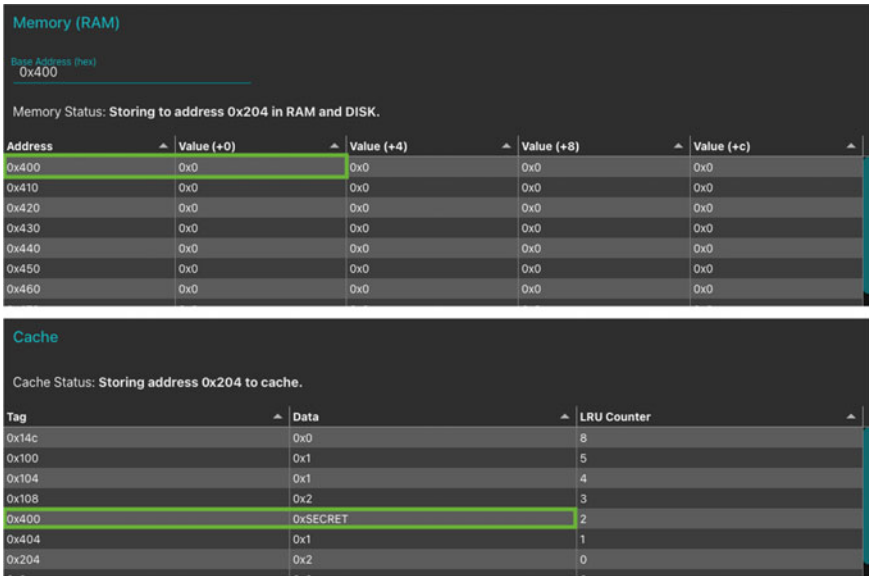


Fig. 5 Secret data remains in the cache, even after another page fault

5 Conclusions

This study achieves the goal of providing a tangible framework to show the effects of modern CPU exploits. We do this by showing how secret data can be accessed through side channels and loaded into the data cache to be accessed by the attacker. We contribute an educational simulator of some CPU exploits that shows the effects of these attacks such that understanding them is easier, as well as to encourage further research and development in combating these exploits. Future work would be to add support for J-type instructions, floating point values, and the ability to change any of the cache mapping or replacement policies.

References

1. P. Kocher, J. Horn, A. Fogh, D. Genkin, D. Gruss, W. Haas, M. Hamburg, M. Lipp, S. Mangard, T. Prescher, M. Schwarz, Y. Yarom, Spectre attacks: Exploiting speculative execution, in *2019 IEEE Symposium on Security and Privacy (SP)* (2019), pp. 1–19
2. M. Lipp, M. Schwarz, D. Gruss, T. Prescher, W. Haas, S. Mangard, P. Kocher, D. Genkin, Y. Yarom, M. Hamburg, Meltdown. [arXiv:1801.01207](https://arxiv.org/abs/1801.01207) (2018)
3. IBM, Node-RED. <https://nodered.org/>. Accessed May 7, 2021
4. J. Van Bulck, M. Minkin, O. Weisse, D. Genkin, B. Kasikxi, F. Piessens, M. Silberstein, T. Wenisch, Y. Yarom, R. Strackx, Foreshadow: extracting the keys to the Intel SGX kingdom with transient out-of-order execution. in *27th USENIX Security Symposium (USENIX Security 18)* (2018), pp. 991–1008
5. C. Canella, D. Genkin, L. Giner, D. Gruss, M. Lipp, M. Minkin, D. Moghimi, F. Piessens, M. Schwarz, B. Sunar, J. Van Bulck, Y. Yarom, Fallout: Leaking data on meltdown-resistant CPUs, in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (2019), pp. 769–784
6. C. Canella, K. Khasawneh, D. Gruss, The evolution of transient-execution attacks, in *Proceedings of the 2020 on Great Lakes Symposium on VLSI* (2020), pp. 163–168
7. C. Canella, S.M. Pudukotai Dinakarrao, D. Gruss, K. Khasawneh, Evolution of defenses against transient-execution attacks, in *Proceedings of the 2020 on Great Lakes Symposium on VLSI* (2020), pp. 169–174
8. O. Oleksenko, B. Trach, M. Silberstein, C. Fetzer, SpecFuzz: bringing spectre-type vulnerabilities to the surface, in *29th USENIX Security Symposium (USENIX Security 20)* (2020), pp. 1481–1498
9. mips.com, MIPS Architecture For Programmers Volume 1-A, Available via DIALOG. <https://s3-eu-west-1.amazonaws.com/downloads-mips/documents/MD00082-2B-MIPS32INT-AFP-06.01.pdf>. Accessed May 7, 2021
10. K. Vollmar, P. Sanderson, MARS: an education-oriented MIPS assembly language simulator. *SIGCSE* **6**, 239–243 (2006)
11. J. Van Bulck, D. Moghimi, M. Schwarz, M. Lipp, M. Minkin, D. Genkin, Y. Yarom, B. Sunar, D. Gruss, F. Piessens, LVI: Hijacking transient execution through microarchitectural load value injection in *2020 IEEE Symposium on Security and Privacy (SP)* (2020), pp. 54–72
12. Intel, An optimized mitigation approach for load value injection. <https://intel.ly/2CSsHwp>. Accessed May 10, 2021
13. M. Larabel, The brutal performance impact from mitigating the LVI vulnerability. <https://www.phoronix.com/scan.php?page=article&item=lvi-attack-perf>

Author Index

A

Abinaya, K., 509
Abrar Jahin, S. M., 695
Adithya, V., 3
Agarwal, Akshay, 443
Agarwal, Mitushi, 647
Aggarwal, Ritu, 99
Al-Azawi, Mohammad A. N., 15
Ananth kumar, T., 509
Anderson, Ethan, 695
Anwar, Khalid, 169
Areeb, Qazi Mohammad, 239
Arora, Jatin, 371

B

Balabantaray, Rakesh Chandra, 659
Bansal, R. K., 613
Bansal, Savina, 613
Basani, Vyomikaa, 181
Bhosale, Madhura M., 423
Boda, Rutvi, 379
Bora, Saranga, 581

C

Chaba, Mridul, 601
Chaba, Yogesh, 601
Chatrath, Sarvjeet Kaur, 601
Chatterjee, Rajdeep, 193
Chaurasiya, Rahul Kumar, 623
Chavan, Ashish, 569
Chawla, Paras, 477
Chik, Abdullah bin, 671
Chopra, Muskaan, 229

Choudhary, Rohan, 51
Chu, Yul, 695

D

Dalai, Asish Kumar, 133
Dave, Mayank, 397
Deepak, Gerard, 3, 273, 283, 293, 315, 325
Dembla, Deepak, 601
Deshpande, Aaditya, 569
Dhalaria, Meghna, 591
Dheenadhayalan, R., 315
Dhotre, Prashant, 569
Dogra, Aayush, 477

F

Fazili, Mohammad Mudakir, 679

G

Galphade, Manisha, 251, 261
Gandotra, Ekta, 591
Gill, Shabeg Singh, 229
Goel, Sonia, 63
Goyal, Bhawna, 477
Goyal, Gulshan, 89
Goyal, Hemlata, 661
Gupta, Amit, 251
Gupta, Anshul, 229
Gupta, Anupma, 477
Gupta, Surbhi, 207

H

Hatti, Vaishali, 433

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

P. Verma et al. (eds.), *Advances in Data Computing, Communication and Security*, Lecture Notes on Data Engineering and Communications Technologies 106, <https://doi.org/10.1007/978-981-16-8403-6>

I

Iqbal, Arshad, 169
 Ismail, Azlan, 219

J

Jain, Megha, 635
 Jain, P. C., 487
 Jain, Puneet Kumar, 407
 Jaiswal, Kavita, 347
 Jaiswal, Vaibhav, 39
 Jawaddi, Siti Nuraishah Agos, 219
 Jindal, Poonam, 533
 Johari, Muhammad Hamizan, 219

K

Karhade, Ashish, 251
 Kaur, Amanpreet, 543
 Kaur, Gagandeep, 543
 Kaushik, Baijnath, 77
 Kedas, Shweta, 407
 Kesavan, T., 467
 Khan, Yusera Farooq, 77
 Kochhar, Tanveer Singh, 89
 Kollengode, Chidambaran, 109
 Krishnan, N., 283
 Kumar, Ajay, 687
 Kumar, Amit, 533
 Kumar, Arun, 407
 Kumar, Santosh, 443

L

Lakshmi, K., 467
 Landge, Pallavi, 251
 Lee, John J., 695, 707

M

Mahajan, Pulkit, 443
 Mamatha, I., 497
 Manaswini, S., 293
 Mathur, Aradhya, 647
 Mazumdar, Saptarshi, 193
 McDonough, Corey, 707
 Meena, Om Prakash, 623
 Mendes, Isaac Kennedy Alexandre, 347
 Michael, Emmanuel Obichukwu, 707
 Miles, Samuel, 707
 Mittal, Ajay, 305
 More, Nilkamal, 261

N

Nadeem, Mohammad, 239
 Nair, Anjitha, 27
 Nemane, Rutuja, 27
 Nikam, V. B., 261
 Ningthoujam, Sanjit, 133
 Nithyusha, Nemani, 623

O

Ojha, Rituraj, 325

P

Padmavathy, N., 555
 Pandey, Dhiraj, 635
 Pangotra, Suksham, 679
 Panigrahy, Saroj Kumar, 133, 379
 Parashar, Priyanka, 647
 Patel, Falguni N., 121
 Pattanaik, Anmol, 659
 Pavana, C., 109
 Pavithra, M., 509
 Phani Raghavendra Sai, K., 497
 Pillai, Shreya, 27
 Pradhan, Astik Kumar, 337
 Prakash, Shruti, 181
 Pranav, M., 273
 Praveen Kumar, P., 509

R

Rajesh, R., 521
 Rajmohan, R., 509
 Ranjan, Rishwari, 361
 Rao, Shilpa, 581
 Ratnesh, Ratneshwar Kr., 687
 Ray, Niranjana Kumar, 337
 Rout, Jitendra Kumar, 337

S

Sadhvani, Sapna, 451
 Sahoo, Somya Ranjan, 133, 379
 Sahu, Satya Prakash, 157
 Sahu, Tirath Prasad, 157
 Saini, Kriti, 397
 Sai Vyshnavi, T., 181
 Santhanakrishnan, T., 521
 Santhanavijayan, A., 3, 273, 293
 Saxena, Ankit Sahai, 361
 Seshadri, Karthick, 109
 Shah, Hitesh B., 121
 Shah, Shishir, 121

Shankar Kumar, Valli Sanghami, [707](#)
 Sharma, Jatin, [391](#)
 Sharma, Krishna Kewal, [635](#)
 Sharma, Manju, [145](#)
 Sharma, Yashender, [613](#)
 Shiturkar, Nupur, [27](#)
 Shivam, [679](#)
 Simunic, Dina, [423](#)
 Sindhu, Korrapati, [109](#)
 Singh, Abhishek, [679](#)
 Singh, Animesh, [305](#)
 Singh, Jaiteg, [371](#)
 Singh, Man Mohan, [687](#)
 Singh, Saravjeet, [371](#)
 Singh, Sunil Kr., [229](#), [305](#)
 Sinha, Adwitiya, [647](#)
 Sinthuja, U., [433](#)
 Sobhanayak, Srichandan, [347](#)
 Sohail, Shahab Saquib, [169](#)
 Solanki, Arun, [39](#), [51](#)
 Srikanth, Kriti, [451](#)
 Srinivas, Nallapalem Neeraj, [487](#)
 Swain, Satyajit, [337](#)

T

Taherdoost, Hamed, [391](#), [417](#)
 Talukder, Niloy, [695](#)
 Tanuja Satish Dhope (Shendkar), [423](#)
 Thakare, Anuradha, [27](#)
 Thakral, Prateek, [99](#)
 Thakran, Ajay, [443](#)
 Thavamani, S., [433](#)
 Tripathi, Neeraj, [679](#)

Tushir, Meena, [63](#)
 Tyagi, Parul, [601](#)
 Tyagi, Sanjay, [145](#)

U

Uma Rao, K., [181](#)
 Urolagin, Siddhaling, [451](#)
 Utkarsh, Khushi, [647](#)

V

Velapure, Akshay P., [423](#)
 Vellisetty, Yasaswini, [487](#)
 Venkateswara Rao, Ch., [555](#)
 Verma, Pankaj, [533](#)
 Verma, Satya, [157](#)
 Vijayan Pillai, S., [521](#)

W

Wagh, Abhishek, [261](#)

X

Xin, Lam Zi, [671](#)

Y

Yogi, Abhishek, [251](#)

Z

Zafar, Aasim, [169](#)