Lecture Notes in Electrical Engineering 855

Pradeep Kumar Singh ·
Yashwant Singh ·
Jitender Kumar Chhabra · Zoltán Illés ·
Chaman Verma *Editors*

Recent Innovations in Computing

Proceedings of ICRIC 2021, Volume 2

Springer

Lecture Notes in Electrical Engineering

Volume 855

Series Editors

Leopoldo Angrisani, Department of Electrical and Information Technologies Engineering, University of Napoli Federico II, Naples, Italy
Marco Arteaga, Departament de Control y Robótica, Universidad Nacional Autónoma de México, Coyoacán, Mexico
Bijaya Ketan Panigrahi, Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, Delhi, India
Samarjit Chakraborty, Fakultät für Elektrotechnik und Informationstechnik, TU München, Munich, Germany
Jiming Chen, Zhejiang University, Hangzhou, Zhejiang, China
Shanben Chen, Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
Tan Kay Chen, Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore
Rüdiger Dillmann, Humanoids and Intelligent Systems Laboratory, Karlsruhe Institute for Technology, Karlsruhe, Germany
Haibin Duan, Beijing University of Aeronautics and Astronautics, Beijing, China
Gianluigi Ferrari, Università di Parma, Parma, Italy
Manuel Ferre, Centre for Automation and Robotics CAR (UPM-CSIC), Universidad Politécnica de Madrid, Madrid, Spain
Sandra Hirche, Department of Electrical Engineering and Information Science, Technische Universität München, Munich, Germany
Faryar Jabbari, Department of Mechanical and Aerospace Engineering, University of California, Irvine, CA, USA
Limin Jia, State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China
Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
Alaa Khamis, German University in Egypt El Tagamoa El Khames, New Cairo City, Egypt
Torsten Kroeger, Stanford University, Stanford, CA, USA
Yong Li, Hunan University, Changsha, Hunan, China
Qilian Liang, Department of Electrical Engineering, University of Texas at Arlington, Arlington, TX, USA
Ferran Martín, Departament d'Enginyeria Electrònica, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain
Tan Cher Ming, College of Engineering, Nanyang Technological University, Singapore, Singapore
Wolfgang Minker, Institute of Information Technology, University of Ulm, Ulm, Germany
Pradeep Misra, Department of Electrical Engineering, Wright State University, Dayton, OH, USA
Sebastian Möller, Quality and Usability Laboratory, TU Berlin, Berlin, Germany
Subhas Mukhopadhyay, School of Engineering & Advanced Technology, Massey University, Palmerston North, Manawatu-Wanganui, New Zealand
Cun-Zheng Ning, Electrical Engineering, Arizona State University, Tempe, AZ, USA
Toyoaki Nishida, Graduate School of Informatics, Kyoto University, Kyoto, Japan
Federica Pascucci, Dipartimento di Ingegneria, Università degli Studi "Roma Tre", Rome, Italy
Yong Qin, State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China
Gan Woon Seng, School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore, Singapore
Joachim Speidel, Institute of Telecommunications, Universität Stuttgart, Stuttgart, Germany
Germano Veiga, Campus da FEUP, INESC Porto, Porto, Portugal
Haitao Wu, Academy of Opto-electronics, Chinese Academy of Sciences, Beijing, China
Walter Zamboni, DIEM - Università degli studi di Salerno, Fisciano, Salerno, Italy
Junjie James Zhang, Charlotte, NC, USA

The book series *Lecture Notes in Electrical Engineering* (LNEE) publishes the latest developments in Electrical Engineering - quickly, informally and in high quality. While original research reported in proceedings and monographs has traditionally formed the core of LNEE, we also encourage authors to submit books devoted to supporting student education and professional training in the various fields and applications areas of electrical engineering. The series cover classical and emerging topics concerning:

- Communication Engineering, Information Theory and Networks
- Electronics Engineering and Microelectronics
- Signal, Image and Speech Processing
- Wireless and Mobile Communication
- Circuits and Systems
- Energy Systems, Power Electronics and Electrical Machines
- Electro-optical Engineering
- Instrumentation Engineering
- Avionics Engineering
- Control Systems
- Internet-of-Things and Cybersecurity
- Biomedical Devices, MEMS and NEMS

For general information about this book series, comments or suggestions, please contact leontina.dicecco@springer.com.

To submit a proposal or request further information, please contact the Publishing Editor in your country:

China

Jasmine Dou, Editor (jasmine.dou@springer.com)

India, Japan, Rest of Asia

Swati Meherishi, Editorial Director (Swati.Meherishi@springer.com)

Southeast Asia, Australia, New Zealand

Ramesh Nath Premnath, Editor (ramesh.premnath@springernature.com)

USA, Canada:

Michael Luby, Senior Editor (michael.luby@springer.com)

All other Countries:

Leontina Di Cecco, Senior Editor (leontina.dicecco@springer.com)

**** This series is indexed by EI Compendex and Scopus databases. ****

More information about this series at https://link.springer.com/bookseries/7818

Pradeep Kumar Singh · Yashwant Singh ·
Jitender Kumar Chhabra · Zoltán Illés ·
Chaman Verma
Editors

Recent Innovations in Computing

Proceedings of ICRIC 2021, Volume 2

Springer

Editors
Pradeep Kumar Singh
KIET Group of Institutions
Ghaziabad, India

Jitender Kumar Chhabra
Department of Computer Engineering
NIT Kurukshetra
Kurukshetra, India

Chaman Verma
Faculty of Informatics
Eötvös Loránd University (ELTE)
Budapest, Hungary

Yashwant Singh
Department of CSE
Central University of Jammu
Jammu and Kashmir, India

Zoltán Illés
Faculty of Informatics
Eötvös Loránd University (ELTE)
Budapest, Hungary

ISSN 1876-1100 ISSN 1876-1119 (electronic)
Lecture Notes in Electrical Engineering
ISBN 978-981-16-8891-1 ISBN 978-981-16-8892-8 (eBook)
https://doi.org/10.1007/978-981-16-8892-8

© 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

Preface

The fourth version of international conference was hosted on the theme Recent Innovations in Computing (ICRIC 2021), the conference was hosted by the Eötvös Loránd University (ELTE), Hungary in association with Knowledge University, Erbil, WSG University in Bydgoszcz, Poland and other academic associates, technical societies, IAC and other universities from India and abroad. The conference covers the tracks; advanced computing, intelligent networking, image processing and computer vision, e-learning, cloud and big data, security and privacy, and Digital India. The researchers were invited to present their research ideas at the Fourth International Conference on Recent Innovations in Computing (ICRIC 2021) on these tracks during two days of conference 8th to 9th June, 2021. We appreciate our valuable writers' contributions, as well as our Technical Program Committee's tremendous support and inspiration in making the 4th ICRIC 2021 a success. The conference was started with the opening remark of Dr. Zoltán Illés, Eötvös Loránd University (ELTE), Hungary. He welcomes all the participants and session chairs along with the keynotes. Knowledge University, Erbil was the academic partner for the conference and the inaugural speech with keynote was delivered by Dr. Kayhan Zrar Ghafoor, who is currently serving as a president of Knowledge University. Dr. Zdzislaw Polkowski delivers his talk on behalf of WSG University in Bydgoszcz, Poland.

We would like to express our sincere gratitude to our all session chairs—Prof. Jitendra Kumar Chhabra, NIT Kurukshetra, Dr. Arpan K.Kar, IIT Delhi; Dr. Maheshkumar H. Kolekar, IIT Patna; Prof. Manu Sood, Himachal Pradesh University; and Prof. Sudeep Tanwar, Nirma University, India, Dr. Ashutosh Sharma, Dr. Aruna Malik and Dr. Samayveer Singh from NIT Jalandhar, India. Dr. Ashima, Dr. Sumit Kumar, Dr. Nagesh Kumar, Dr. Vivek Sehgal, Dr. Yugal Kumar Chaired the session during the technical presentations. Dr. Veronika Stoffova, Trnava University in Trnava, Slovakia delivered a short speech followed by the session chair during the conference. Dr. Viktória Bakonyi, University of Eötvös Loránd, Hungary, and Dr. Chaman Verma from ELTE Hungary chaired the technical session and carried out a short panel discussion for the participants. Dr. Pljonkin Anton Pavlovich, Southern Federal University, Russia and Dr. Ashutosh Mishra, Yonsei University, South Korea also took part in panel discussion during the conference and chaired the technical

sessions as well. The last session during the conference was chaired by the Dr. Maria Simona Raboaca, Faculty of Electrical Engineering and Computer Science, University of Suceava, Romania and Dr. Praveen Kumar Singh, UIDAI, Lucknow, India.

We are also grateful to Eötvös Loránd University (ELTE), Hungary management board, rectors, vice rectors, deans, and professors for extending their help during the conference. Many other professors from different countries also deserve our gratitude for devoting their time to listen the paper presentations and for giving their valuable feedback to the authors. We extend our thanks to the Springer, LNEE Series, editorial board for believing in us.

Ghaziabad, India — Pradeep Kumar Singh
Jammu and Kashmir, India — Yashwant Singh
Kurukshetra, India — Jitender Kumar Chhabra
Budapest, Hungary — Zoltán Illés
Budapest, Hungary — Chaman Verma
June 2021

Contents

Advanced Computing

AWRPS-ROBO: Automated Weed Removal and Pesticides Spray 3
Sushopti Gawade, K. S. Charumathi, and Y. I. Jinesh Melvin

Designing of Cavity Filter with Slot Coupling Mechanism for k_u Band 13
Puneet Chandra Srivastava, Leena Sharma, Kiran Srivastava, and Praveen Kumar Malik

Use of Smart Mobile Applications with IoT in Diseases Prediction System for Apple Orchards 27
Karuna Sheel and Anil Sharma

Review on Miniaturized Flexible Wearable Antenna for Body Area Network 37
Utkarsh Pandey, Narbada Prasad Gupta, and Praveen Malik

DEERS: Design Energy-Efficient Routing Scheme for Harsh Environment Monitoring in Heterogeneous WSNs 53
Samayveer Singh, Aruna Malik, Pawan Singh Mehra, and Pradeep Kumar Singh

Context-Enriched Machine Learning-Based Approach for Sentiment Analysis 67
Hamza Abubakar Kheruwala, Mohammed S. Ahmad, Jai Prakash Verma, Sudeep Tanwar, and Pradeep Kumar Singh

Recommending Books Using RNN 85
Mala Saraswat, Rishi Saraswat, and Renu Bahuguna

A Survey on Applications of Unmanned Aerial Vehicles (UAVs) 95
Ritu Dewan and Khandakar Faridar Rahman

Early Detection of Influenza Using Machine Learning Techniques 111
Sajal Maheshwari, Anushka Sharma, Ranjan Kumar, and Pratyush

Fuzzy Time-Series Models Based on Intuitionistic Fuzzy, Rough Set Fuzzy, and Differential Evolution 125
Partha Pratim Deb, Diptendu Bhattacharya, and Indranath Chatterjee

Genetic Algorithm Application on 3D Pipe Routing: A Review 139
Vivechana Maan and Aruna Malik

Directed Undersampling Using Active Learning for Particle Identification 149
Zakarya Farou, Sofiane Ouaari, Balint Domian, and Tomáš Horváth

Smart Agriculture Using Internet of Things: An Empirical Study 163
Mohit Kumar Saini and Rakesh Kumar Saini

Intellegent Networking

A Study on the Implementation of Secure VANETs Using FPGA 179
Harsha Vardan Maddiboyina, V. A. Sankar Ponnapalli, and A. Naresh Kumar

Adoption of Microstrip Antenna to Multiple Input Multiple Output Microstrip Antenna for Wireless Applications: A Review 189
Nitasha Bisht and Praveen Kumar Malik

Massive MIMO System—Overview, Challenges, and Course of Future Research 207
Shailender, Shelej Khera, Sajjan Singh, and Jyoti

Millimeter-Wave Dual-Band (32/38 GHz) Microstrip Patch Antenna for 5G Communication 225
Jyoti Hatte, Shivleela Mudda, K. M. Gayathri, and Rupali B. Patil

Design and Analysis of Single Band and Wideband Wineglass-Shaped Patch Antenna for WLAN and Satellite Applications 239
Narbada Prasad Gupta, Parulpreet Singh, Sanjay Kumar Sahu, and Shelej Khera

ECICM: An Efficient Clustering and Information Collection Method in Heterogeneous Wireless Sensor Networks 249
Samayveer Singh, Aruna Malik, and Pradeep Kumar Singh

Exploring Trust in SDN Along with Network Monitoring 263
Gaurav Sharma and Sushopti Gawade

Improving LoRaWAN Networks Performance Through Optimized Radio Resource Management 277
Husam Rajab, Xi Tiansheng, and Tibor Cinkler

On Security and Performance Requirements of Decentralized Resource Discovery in IoT 289
Mohammed B. M. Kamel, Peter Ligeti, and Christoph Reich

EV Technology Trends & Placement of Electric Vehicle Charging Station: A Review 303
Tripti Kunj and Kirti Pal

Design of Multiband Pattern Reconfigurable Antenna Loaded with Circular Split Ring Resonators 315
Umhara Rasool Khan, Abdul Basit, Javaid A. Sheikh, G. M. Bhat, and Suhaib Ahmed

Optimal Thermal Coordination Dispatch for Demand Side Management 327
Nidhi and Kirti Pal

Optimal Routing in Wireless Sensor Networks: A Review 339
Jasleen Kaur, Punam Rattan, Brahm Prakash Dahiya, and Reenu

Perturbation by Sybil Attack in Clustering for Open IVC Networks (COIN) Protocol—A Protocol in Cluster-Based Routing Category for Infrastructure-Less VANETs 357
Nishtha, Devaashish Sharma, and Manu Sood

Image Processing and Computer Vision

Neuromorphic Computing: Review of Architecture, Issues, Applications and Research Opportunities 371
Hitesh Vora, Preeti Kathiria, Smita Agrawal, and Usha Patel

Computational Intelligence Approaches for Heart Disease Detection 385
Roseline Oluwaseun Ogundokun, Sanjay Misra, Peter Ogirima Sadiku, Himanshu Gupta, Robertas Damasevicius, and Rytis Maskeliunas

An Analysis of Different Machine Learning Algorithms for Image Classification 397
Ankur Chaturvedi, Vikram Rajpoot, Meghansh Bansal, and Hanuman Das Agrawal

Biotic Disease Recognition of Cassava Leaves Using Transfer Learning 411
Rahul Sharma and Amar Singh

A Sentiment Detection Tool for Multiple Domains 425
Priya Shrivastava and Dilip Sharma

Content-Based Image Retrieval (CBIR): A Review 439
Deepti Agrawal, Apurva Agarwal, and Dilip Kumar Sharma

Automatic Speech Emotion Recognition Using Cochleagram Features 453
Saumya Borwankar, Dhruv Shah, Jai Prakash Verma, and Sudeep Tanwar

An Analysis of Various Machine Learning Techniques Used for Diseases Prediction: A Review 467
Mudasir Hamid Sheikh, Sonu Mittal, and Rumaan Bashir

Credit Card Fraud Transaction Classification Using Improved Class Balancing and Support Vector Machines 477
Pradeep Verma and Poornima Tyagi

An Improved Lossless Algorithm for Text Compression 489
Kartik Bhatia, Anupam Singh, Anamol Verma, and Dipansh Mittal

Meta-Heuristic with Machine Learning-Based Smart e-Health System for Ambient Air Quality Monitoring 501
Pankaj Rahi, Sanjay P. Sood, and Rohit Bajaj

Smart eHealth System for Pervasive Healthcare 521
Pankaj Rahi, Sanjay P. Sood, and Sanjay K. Sharma

Identification of Missing Person Using Fusion of KNN and SVM Approach 537
Sandeep Rathor, Afreen Hasan, and Ankur Omar

Current Trends and Future Prospects: Detection of Breast Cancer Using Machine Learning Techniques 547
Ruqsar Zaitoon, Ashwani Kumar, and Syed Saba Raoof

E-Learning Cloud and Big Data

Analysis of Blockchain Secure Models and Approaches Based on Various Services in Multi-tenant Environment 563
Pooja Dhiman and Santosh Kumar Henge

Harmonic Minimization in Multilevel Inverters Using Ant Lion Optimization Algorithm 573
Tushar Tyagi, Amit Kumar Singh, Himanshu Sharma, and Rintu Khanna

Examine the Indian Tweets to Determine Society Emphasis on Novel Corona-Viruses (COVID-19) 587
Anil Kumar Dubey, Mala Saraswat, Raman Kapoor, and Rishu Gupta

Real-Time Rendering with OpenGL and Vulkan in C# 599
Dávid Szabó and Zoltán Illés

Prediction of Students' Performance with Artificial Neural Network Using Demographic Traits 613
Adeniyi Jide Kehinde, Abidemi Emmanuel Adeniyi, Roseline Oluwaseun Ogundokun, Himanshu Gupta, and Sanjay Misra

Real-Time Interaction Tools in Virtual Classroom Systems 625
Viktória Bakonyi, Zoltán Illés, and Tibor Szabó

Cost-Efficient BAT Algorithm for Task Scheduling in Cloud 637
Yagya Malik, Daanish Goyal, Abhiti Sachdeva, and Punit Gupta

Systemic Thinking in Programming Education 645
Szilárd Korom and Zoltán Illés

Technology Based University Identification Model for Real-Time 659
Chaman Verma, Zoltán Illés, and Veronika Stoffová

Comparison of Multi-Criteria Decision-Making Techniques for Cloud Services Selection 669
Neha Thakur, Avtar Singh, and A. L. Sangal

OULAD Learners' Withdrawal Prediction Framework 683
Moohanad Jawthari and Veronika Stoffa

Cloud Computing in Healthcare Industries: Opportunities and Challenges 695
Vinayak Rai, Karan Bagoria, Kapil Mehta, Vandana Mohindru Sood, Kartik Gupta, Lakshya Sharma, and Manav Chauhan

Security and Privacy

The Latest Trends in Collaborative Security System 711
Monika Arora and Sonia

Security Analysis and Deployment Measurement of Transport Layer Security Protocol 725
Marwah Yaseen, Mohammed B. M. Kamel, and Peter Ligeti

A Web Application Vulnerability Testing System 741
Roseline Oluwaseun Ogundokun, Sanjay Misra, Tobe Segun-Owolabi, Abhiram Anand Gulanikar, Akshat Agrawal, and Robertas Damasevicius

iReportNow: A Mobile-Based Lost and Stolen Reporting System 753
Bilkisu Larai Muhammad-Bello, Olatunde Petwilson Lewu, Sanjay Misra, Ajay Kumar Garg, Jonathan Oluranti, and Rytis Maskeliunas

Improving Security and Privacy in Attribute-Based Encryption with Anonymous Credential 767
Yuping Yan and Péter Ligeti

Next Generation Wireless Communication: Facilitated by Machine Learning 779
Praveen Kumar Singh

Simulation-Based Method for Analyzing Timing Attack Against Pass-Code Breaking System 795
Shaminder Kaur, Balwinder Singh, and Lipika Gupta

Quantum Dot Cellular Automata-Based Design of 4 × 4 TKG Gate and Multiplier with Energy Dissipation Analysis 809
Soha Maqbool Bhat, Suhaib Ahmed, and Vipan Kakkar

Artificial Intelligence with Enhanced Prospects by Blockchain in the Cyber Domain 827
Praveen Kumar Singh

Author Index 841

About the Editors

Dr. Pradeep Kumar Singh is currently working as Professor and Head in the department of CS at KIET Group of Institutions, Delhi-NCR, Ghaziabad, India. Dr. Singh is Life Membership of Computer Society of India (CSI), Life Member of IEI and promoted to Senior Member Grade from CSI and ACM. He is Associate Editor of the International Journal of Information System Modeling and Design (IJISMD), Indexed by Scopus and Web of Science. He is also Associate Editor of International Journal of Applied Evolutionary Computation (IJAEC), IGI Global USA, Security and Privacy, Wiley & International Journal of Information Security and Cybercrime (IJISC) a scientific peer-reviewed journal from Romania. He has published nearly 122 research papers in various International Journals and Conferences of repute. Some of his publications are in, e.g., IEEE Access, Applied Intelligence, Multimedia Tools and Applications, and Computer Communications. He has received three sponsored research projects grant from Govt. of India and Govt. of HP worth Rs 25 Lakhs. He has edited a total 12 books from Springer and Elsevier. He has Google scholar citations 1450, H-index 17, and i-10 Index 45.

Dr. Yashwant Singh is Head & Associate Professor in the Department of Computer Science & Information Technology at the Central University of Jammu.

Yashwant completed his Ph.D. from Himachal Pradesh University Shimla, his post Graduate study from Punjab Engineering College Chandigarh and undergraduate studies from SLIET Longowal. His research interests lie in the area of Internet of Things, Vulnerability Assessment of IoT and Embedded Devices, Wireless Sensor Networks, Secure and Energy Efficient Routing, ICS/SCADA Cyber Security, ranging from theory to design to implementation. He has collaborated actively with researchers in several other disciplines of computer science, particularly Machine Learning, Electrical Engineering.

Yashwant has served on Thirty International Conference and Workshop Program Committees and served as the General Chair for PDGC-2014, ICRIC-2018, ICRIC-2019, ICRIC-2020, and ICRIC-2021. He currently serves as coordinator of Kalam Centre for Science and Technology (KCST), Computational Systems Security Vertical at Central University of Jammu established by DRDO.

Yashwant has published more than 80 Research Papers in the International Journals, International Conferences and Book Chapters of repute that are indexed in SCI and SCOPUS. He has 548 Citations, i10-index 23 and h-index 14. He has Research Projects worth Rs.46.322 Lakhs in his credit from DRDO and Rs. 12.19 Lakhs from NCW. He has guided 4 Ph.Ds., 24 M.Tech. students and supervising 4 Ph.Ds. and 6 M.Tech.

Dr. Yashwant Has visited 8 countries for his academic visits e.g. U.K., Germany, Poland, Chez Republic, Hungary, Slovakia, Austria, Romania. He is Visiting Professor at Jan Wyzykowski University, Polkowice, Poland.

Prof (Dr.) Jitender Kumar Chhabra is Professor, Computer Engineering Department at National Institute of Technology, Kurukshetra India. He has published 120 papers in reputed International and National Journals and conferences including more than 40 publications from IEEE, ACM, Elsevier and Springer, most of which are SCI/Scopus indexed. His research interest includes Software Metrics, Data Mining, Soft Computing, Machine Learning, Algorithms & related areas. He is Reviewer for most reputed journals such as IEEE Transactions, ACM Transactions, Elsevier, Wiley, and Springer. He has total 1122 Google Scholar Citations, H-Index 16, and i-10 Index 24.

Dr. Zoltán Illés, Ph.D. Habil. has started higher education studies in subjects Mathematics and Physics at Eötvös Loránd University. He later took up the Computer Science supplementary course, which was started at that time. He got a Hungarian's Republic scholarship based on his outstanding academic achievements during his university studies. He graduated in 1985, after which he started working at the Department of Computer Science of Eötvös Loránd University. He completed his Ph.D. dissertation entitled "Implementation of Real-Time Measurements for High-Energy Ion Radiations" in 2001. In 2004, at the request of Jedlik Publisher, he also wrote a textbook on the C# programming language. This book has a second, expanded edition in 2008. In 2007, he was awarded a scholarship by the Slovak Academy of Sciences, where he spent six months researching and teaching at the Constantine the Philosopher University in Nitra. The NJSZT awarded the Rezső Tarján Prize in 2016 for the success of the joint work that has been going on ever since. He and his colleagues also researched the issue of mobile devices and applications in the framework of a TéT_SK tender won in 2014. Based on their research findings, he launched a pilot project to support real-time, innovative performance management. The first results of this research are an integral part of his habilitation dissertation. He has been Invited Speaker at several international conferences and Member of the Amity University Advisory Board since 2020.

Dr. Chaman Verma, Ph.D. is working as Faculty/Computer Research Scientist and Professional Assistant in the Department of Media and Educational Informatics, Faculty of Informatics, Eötvös Loránd University, Hungary. He has been awarded scholarship ÚNKP, MIT (Ministry of Innovation and Technology) and

National Research, Development and Innovation (NRDI) Fund, Hungarian Government. Additionally, He won the EFOP scholarship of the European Social Fund under the project: "Talent Management in Autonomous Vehicle Control Technologies" sponsored by the European Union and Hungarian Government. He was also nominated and awarded a scientific publication award in 2021 at the Faculty of Informatics, Eötvös Loránd University, Hungary. He got the best research paper awards at IEEE Conefernce, Uzbekistan (2021), and Springer Conefernce, CDAC, India (2020). His Ph.D. has been sponsored under the Stipendium Hungaricum Scholarship of Tempus Public Foundation, Government of Hungary. He has completed his M.Tech. in Computer Science and Engineering from Ch. Devi Lal University, Haryana, India. He has around 10 years of teaching and industry experience. He has more than 60 scientific publications in the IEEE, Elsevier, Springer, IOP Science, MDPI, and Walter de Gruyter. His research interests include statistical analysis, machine learning, deep learning, optimization and feature engineering, real-time systems, and educational informatics. He is a guest editor for Special Issue "Advancements in Machine Learning and Statistical Modeling, and Real-World Applications" in Mathematics Journal, MDPI, Basel, Switzerland. He is a Member of the Editorial Board and a Reviewer of various international journals and scientific conferences. He is a Life Member of ISTE, New Delhi, India. He is a reviewer of many scientific journals of IEEE, Springer, Elsevier, Wiley, MDPI. He has Scopus citations 405 with H-index 12. He has Web of Science citations 90 with H-index 6. He has google citations 613 with H-index 15, i10-index 23.

Advanced Computing

AWRPS-ROBO: Automated Weed Removal and Pesticides Spray

Sushopti Gawade, K. S. Charumathi, and Y. I. Jinesh Melvin

Abstract Weed Management is one of the important aspects in the agricultural area, for increasing yield in cotton crop production. Increasing weeds will damage the crop and it causes to lose the agricultural business also it increases the labor cost, time, more pesticide, spraying unwanted chemicals over the crop, and also its tedium and manual weeding are unfavorable. However, the incessant use of herbicide-resistant weedicides has led to a severe impact on crops and the surrounding environment. The method of mechanical weeding doesn't produce accurate results because it weeds all over the crops without any targets. To avoid such kinds of issues in the agricultural field, we propose a system, which detects and removes the weeds automatically, also it checks the weeds by inter rows or within rows without damaging any other crops. This system can separate the weeds and soil by using robots, also this system helps to avoid the usage of herbicides in the agriculture field and also reduces manpower. The initial stage of our system is to collect the data of different types of weeds, after finding out the weeds, it gets removed with the help of an automated tool (robot), finally, it checks the crops which are affected by any disease and spray the pesticides in the proper area.

Keywords Binarization · Color transform · Color moment · Cotton disease detection · Gabor filter · Image segmentation · Median filter · Support vector machine · Random forest · Weed detection

S. Gawade (✉) · K. S. Charumathi · Y. I. Jinesh Melvin
Pillai College of Engineering, Panvel, Navi Mumbai, Maharashtra, India
e-mail: sgawade@mes.ac.in

K. S. Charumathi
e-mail: kscharumathi@mes.ac.in

Y. I. Jinesh Melvin
e-mail: yijmelvin@mes.ac.inrinceton

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_1

1 Introduction

1.1 Fundamentals

Cotton Plant is considered as a commercial plant and it plays a vital role in Economics as many Textile Industries use cotton as raw material. Today's Cotton Production System produces healthy rural economics and provides significant benefits to the farmers. Weeds are unwanted plants that grow in between cotton plants and which results in the drastic reduction in the yield. Another important aspect that affects the reduction in the yield is the disease that occurs in the cotton plants. For a healthy and good harvest, it is essential for the removal of weed as well as detecting disease from cultivated fields. Manually removing the weeds needs more manpower and a time-consuming process. To overcome this, a robot has been designed for weed removal and pesticide spray on the affected plant.

1.2 Objectives

The main focus of this paper is to design a robot for an automatic weed removal system and pesticide spray. For weed detection, some images of the field have been taken with the help of a camera which is attached in front of the tractor or the Robot or manually. The images are then given to the image processing technique [1] to detect the weed [2] and disease [3]. Then after detecting the weeds that will be removed with the help of a robot arm. Similarly, after detecting the diseases on the leaves, proper pesticides [4] must be sprayed by the robot to increase the yield.

1.3 Scope

Most farmers face the difficulty in identifying the weed and disease of the plant. Moreover, it is very tedious work for the farmers to remove the weed and spray the pesticides manually. The proposed system reduces the burden on the farmers, it identifies the weed and sprays the pesticides on the disease-affected plant automatically.

2 Literature Review

Kesavan et al. [5] proposed a robot vehicle steered by DC Motor with two sensors, Color sensor and IR sensor to detect the weed and the obstacle, respectively. In this proposed system, sometimes there is a possibility of removing the original crop as

no proper classification algorithms are used to detect the weeds. Weeds are detected only on the basis of the color of the leaves.

Patel et al. [1] proposed an Image Processing Technique for identifying the weed and the obstacle with the help of Raspberry Pi. After detecting the weeds, the robot sprayed the weedicides to kill the weeds. Using weedicides gives good results in the starting period but then later it dominates the field which results in the reduction of the yield. This system is successful only when the weed size is smaller than the crop size. Moreover, the initial investment for implementing the solar plant is very high for small farmers and not able to afford the investment.

Sarangdhar et al. [6] proposed a mobile app using Support Vector Machine based on Regression technique for detecting diseases for cotton plants as well as soil monitoring. For soil monitoring sensors are used to find soil humidity, moisture, temperature, and water level sensors. This app provides ON/OFF relay to control the motor and the sprinkler according to their need. After detecting the disease by the mobile app, it would be notified to the farmer and spray the pesticides manually.

Ngo et al. [7] proposed Convolution Neural Network for image classification to detect the weed and the crop, and a robot is designed to spray the herbicides on the weed or near to it. A DC Motor rotation is used for the robot which is not able to stop at a certain time because of its instability. It causes not capturing the image or the capturing of wrong images which leads to the wrong image classification. More care should be taken by the robot not to spray herbicide on the good crop.

Pusphavalli et al. [8] proposed automatic weed removal using Machine Vision. The system detects the weeds by giving an image as input and differentiated from the crop. After detecting the weed it could be removed mechanically using the Hoe or Knife which is attached with the machine.

3 Methodology

The proposed methodology includes three modules such as weed detection module, disease detection module, and a robot module. The following Fig. 1 describes the Proposed System Architecture.

The input image of the plant and the damaged leaf is fed as input to the weed detection module and disease detection module, respectively. Then the output from both the modules instructs the robot to take the following two actions. One action is to remove the weed and separate the soil from the weed plant. Then the soil is sent to soil testing for further research. Second action is, based on the disease detected, the robot is instructed to spray pesticides. The following section discusses the modules.

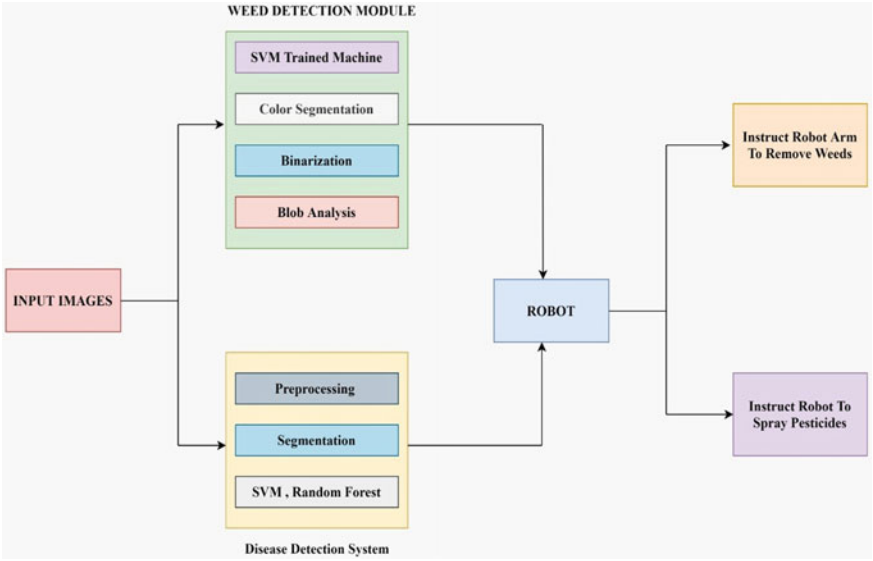


Fig. 1 Proposed system architecture

3.1 Weed Detection Module

Weed is detected using machine-learning with the implementation of Support Vector Machines (SVMs) and blob analysis for the effective classification of weeds. SVM is trained with two data samples. One data set is a crop data set and another data set is a weed data set. Analysis of consistent image regions is identified using the technique called Blob analysis [9]. To identify the weed spices the image processing techniques are used. Different data samples are collected and arranged in a multidimensional array. The image processing algorithm detects the weed and crop based on the feature extraction such as length, centroid, and RGB. After detecting the weed then the data is handed over to the robot for removing the weed. While removing the weed, soil is separated from the weed by using image segmentation method as shown in Fig. 2. Binarization method is implemented to differentiate the soil and the plant and then soil is separated from the weed with the use of the robot.

3.2 Disease Detection Module

Cotton Leaf Disease is detected using SVM and Supervised Learning Process based on Random Forest Algorithm. An input image (Damaged Leaf) is given to Image Acquisition, Pre-processing Image, Segmentation, feature extraction, Classification using SVM/RF Method to predict the disease. Image acquisition is the process of

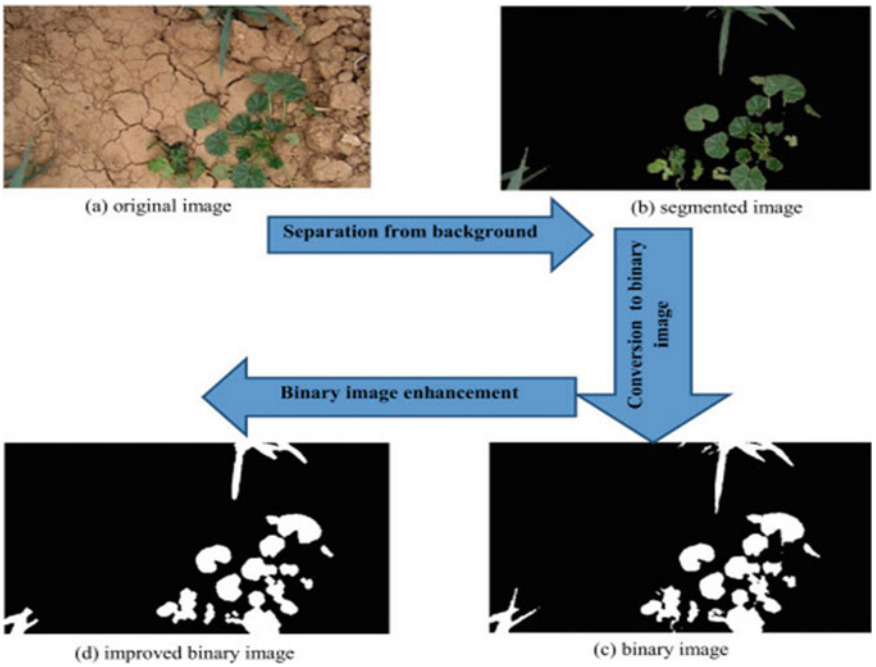


Fig. 2 Example of weed detection system image segmentation

taking the image by using a digital camera or mobile phone and the images are collected into the database. Then the image is given to the Pre-Processing Technique which removes the noise in the image and improves the quality of the image. Pre-processing includes enhancement of image (removing Noise), color conversion, resize, and filtering of the images. Median filter is used in this system to get accurate results with edge preservation while removing noise. Then segmentation is the next one which is used to separate the damaged portion of the infected leaf. This process separates the regions which are having similar properties by applying framing limits for these regions. The damaged area (Region of Interest) is detected using the color transformation method which transforms RGB color to YCbCr and then lastly to Gray Image. Then applying Feature Extraction, some of the features are extracted based on color and texture, and important extracted information is stored in a vector for further process such as to train the SVM/Random Forest for classification. This is the last step to detect the disease by comparing the data vector with the trained data. Accuracy can be analyzed for both SVM and Random Forest methods. Detailed process flow is shown in Fig. 3.

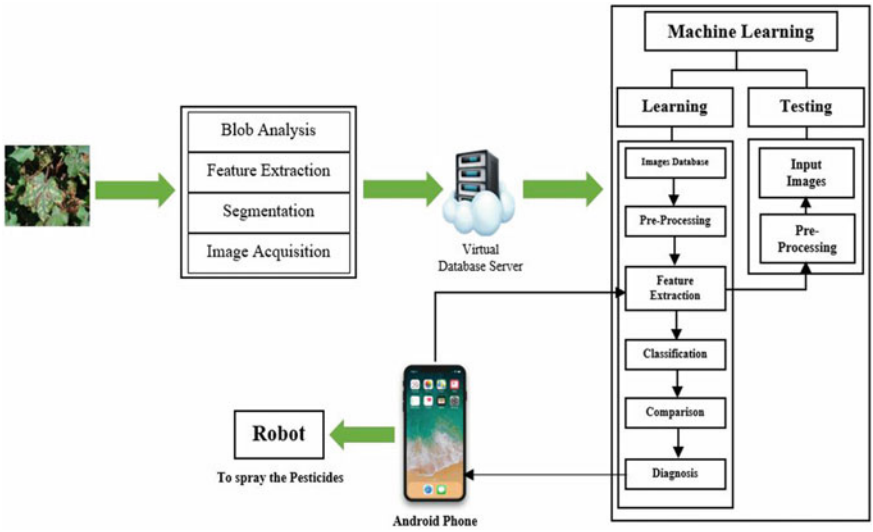


Fig. 3 Disease detection system

3.3 Robot Module

Robot is an electromechanical system embedded with software programming. The proposed system robot is designed with four wheels which are operated with two DC Motors. Each one is driven and directed individually using direct transmission of signals from a microcomputer Raspberry pi. In this system two robot arms are needed; one is used to remove the weed and another is to spray pesticides. For each arm, four servo motors are used. The arm has three directions of motion and a grip movement operated with four servo motors. Each servo motor is designed to do specific operation based on pulse width modulation. The programming language Python is used to control the servo motor rotation such as pick and drop operation. The arm movement is controlled by Raspberry pi 3 Model. Depending on the RPM of the DC Motor axle, it drives the vehicle at a particular distance. Once the DC motor starts the vehicle moves along column wise between the crop lines. Digital Camera is connected in front of the robot to take a snapshot of the leaf. The image is then given to the trained machine through raspberry pi. Sensors are also attached with raspberry pi to detect the temperature, humidity, and moisture of the soil. For detecting any obstacle in the path, IR Sensor is also attached with a raspberry pi module. Detailed robot system is shown in Figs. 4 and 5.

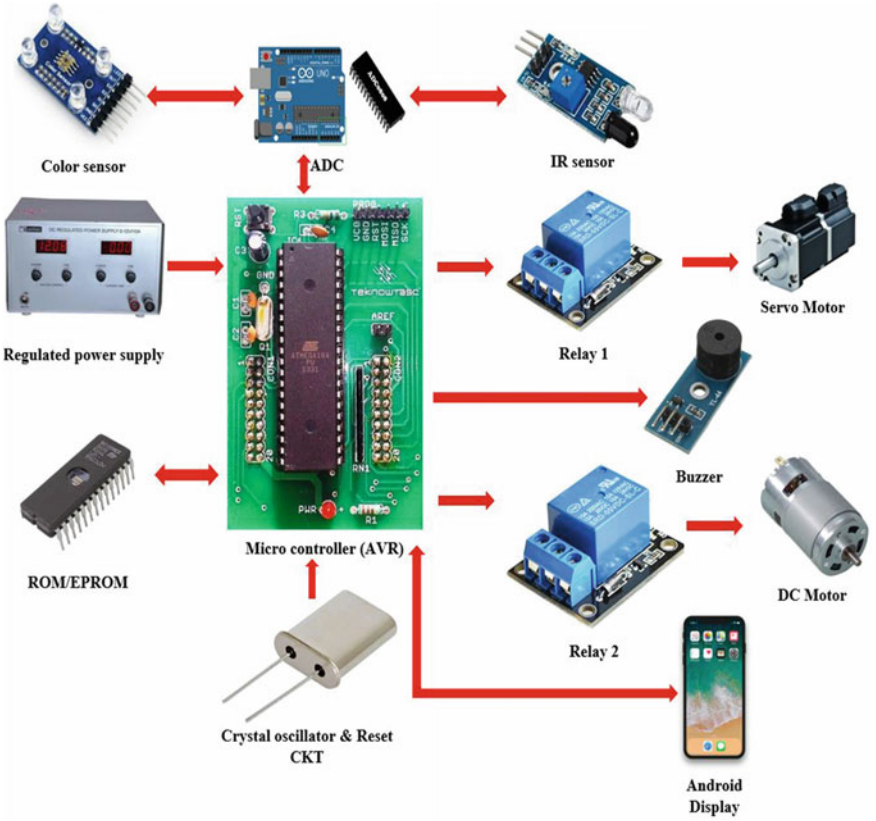


Fig. 4 Robot system

4 Conclusion

The main focus is on smarter, automatic, and more efficient models for crop growing in order to balance the demand and supply chain management for food in this increasing population of the country. This paper is considering all these aspects and discusses the role of various technologies such as SVM, Image Processing, and IoT as well. Image Processing, SVM/Random Forest method to identify the weed and detect the disease of the plant. It removes the weed as well as spray the pesticides automatically without burdening the farmers. The crop and weed are differentiated using three basic techniques: image conversion, SVM, and Blob analysis. These techniques are applicable for varying size of the leaf, weed, and lighting condition. The disease is also detected using Image segmentation, Feature extraction, and SVM/Random forest method. Accuracy of both SVM and Random Forest method is to be analyzed. A robot is designed to remove the weed and separate the soil from

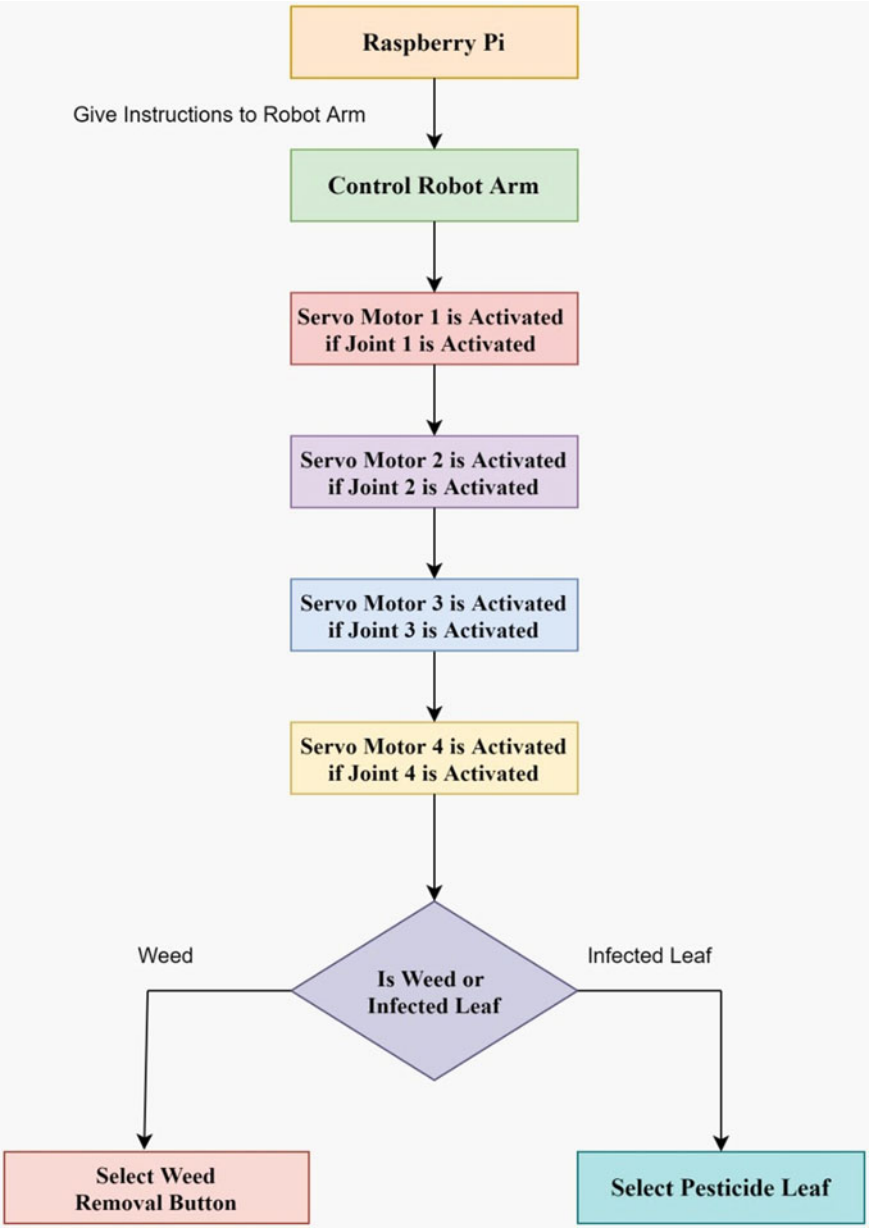


Fig. 5 Working mechanism of Robot Arm

the weed. In addition to this a sprayer is attached with the robot arm to spray pesticides on the defected plant. Robot is instructed and controlled through the Internet of Things. This paper provides an automatic system for weed and disease detection in order to increase the crop yield to the maximum extent.

References

1. H. Patel, A. Prajapati, R. Maheshwari, Design & implementation of solar weeding robot for cotton field. Int. Res. J. Eng. Technol. (IRJET) (2018)
2. https://data.world/datasets/weeds. Accessed Feb 2020
3. http://www.ikisan.com/mhcottondiseasemanagement.html. Accessed Feb 2020
4. https://data.world/datasets/pesticides. Accessed Jan 2020
5. K. Kesavan, R. Monish, S. Murali, Automatic weed removal system. Int. J. Res. Eng. Sci. Manag. (2018)
6. A.A. Sarangdhar, V.R. Pawar, Machine learning regression technique for cotton leaf disease detection and controlling using IoT. in *International Conference on Electronics, Communication and Aerospace Technology*, (2017)
7. H.C. Ngo, U.R. Hashim, Y.W. Sek, Y.J. Kumar, W.S. Ke, Weeds detection in agricultural fields using convolutional neural network. Int. J. Innov. Technol. Exploring Eng. (IJITEE) (2019)
8. M. Pusphavalli, R. Chandraleka, Automatic weed removal system using machine vision. Int. J. Adv. Res. Electron. Commun. Eng. (IJARECE) (2016)
9. S. Murawwat, A. Qureshi, S. Ahmad, Y. Shahid, Weed detection using SVMs. Eng. Technol. Appl. Sci. Res. (2018)
10. K. Raikar, S. Gawade, V. Turkar, Usability improvement with crop disease management as a service. in *2017 International Conference on Recent Innovations in Signal processing and Embedded Systems (RISE-IEEE)*
11. S. Gawade, V. Turkar, Analysis of digital media compatibility with framers in maharashtra and recommendation of service provider design framework E-Krishimitra. Int. J. Appl. Agric. Res. **12** (2017)
12. K. Raikar, S. Gawade, Review of usability and digital divide for ICT in Agriculture. Int. J. Adv. Res. (2017)
13. Plant village Cotton https://www.plantvillage.org/en/topics/cotton. Accessed Jan 2020
14. S. Sabzi, Y. Abbaspour-Gilandeh, J.I. Arriba An automatic visible-range video weed detection, segmentation and classification prototype in potato field. (Elsevier Ltd., 2020)

Designing of Cavity Filter with Slot Coupling Mechanism for k_u Band

Puneet Chandra Srivastava, Leena Sharma, Kiran Srivastava, and Praveen Kumar Malik

Abstract This paper presents investigations on variations in the physical dimension of cavity resonator and filter at microwave frequencies. In the previous papers, which have been studied so far, it has been observed that the characteristics of resonant cavity-based bandpass filter depend on many parameters such as cavity's length, width, depth, location and size of coupling slots, feed line parameters, and the gap between the two microstrip lines. In this paper, variations in cavity width with respect to guided wavelength have been done. Slot excitation with shorting vias has been applied and its performance has been evaluated by simulation. In this paper, the methodology to achieve the desired bandwidth, −3 dB lower edge frequency and higher edge frequency has been discussed. Three formulae have been proposed for estimating −3 dB lower edge frequency, −3 dB upper edge frequency and the bandwidth as a function of the cavity width. These formulae have been validated. A bandpass filter operating at k_u band has been used as an example. This filter exhibits an insertion loss of 1 dB at the center frequency of 14.95 GHz and a 3 dB bandwidth of 2.30 GHz with a roll-off of 8.3 dB/GHz and a return loss of 27.36 dB. The design and fabrication are fully compatible with modern microwave-integrated circuit (MIC) technology. These considerations have led to the choice of silicon as substrate material, micromachining for cavity formation and metal layers for ground and microstrip lines. By changing physical dimensions like cavity width of the filter with respect to the guided wavelength, the effects on filters parameters have been observed and through simulation, validated results have been presented. These results will be helpful in designing of cavity filter.

Keywords Microstrip · Cavity filter · Bandwidth · Cut-off frequencies · Roll-off

P. C. Srivastava (✉) · L. Sharma
Raj Kumar Goel Institute of Technology, Ghaziabad, U.P, India

K. Srivastava
Galgotias College of Engineering and Technology, Greater Noida, U.P, India

P. K. Malik
Lovely Professional University, Phagwara, Punjab, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_2

1 Introduction

In communication systems, for sorting and separation of signals filters are essential [1]. They are used in various communication systems that typically transmit and receive amplitude and/or phase-modulated signals across a communication channel [2]. Microstrip bandpass channels using conveyed segments are very well known in the present-day correspondence framework [3]. The plan approach with coupled-resonator and microstrip channel makes the channel recreation system straightforward [4]. By using different techniques, nowadays wideband filters have been developed, such as parallel coupled-resonator [5]. Because of simple and compact structure and easy fabrication of single-slot cavity structure [6], H slot microstrip filter was proposed [7] in which microstrip lines are used to feed H-shaped coupling slots on the grounded metal plane. H-shaped coupling slots feed top radiating patches, which can realize the radiation function of the antenna. While using a defect ground structure, a slot-coupled microstrip bandpass filter [8], a transversal filter with notch [9] etc. have been proposed. In 2017, a k_u band bandpass frequency selective surface (FSS) filter with high selectivity has been proposed [10]. In [11], for analyzing multiple slot couplings and simultaneous couplings on different walls of a rectangular cavity was proposed in which six-port junction is used to access a rectangular cavity from every direction. Moreover, the difficulty in deciding the parameters for design has not been discussed so far. In the present paper, cavity filter design using single slot having shorted via is proposed with three formulae which theoretically investigated, and its performance is validated through simulation by HFSS.

2 Rectangular Resonant Cavity

To form a closed box or cavity as shown in Fig. 1, a rectangular waveguide should be short-circuited at its two ends, $z = 0$ and $z = d$, hence a resonant structure constructs. Electric and magnetic energy is stored within the cavity, and power can be dissipated in the metallic wall of the cavity as well as in the dielectric material filling the cavity [13]. The energy stored by the resonator is associated with the electromagnetic field within the volume of the cavity. Coupling to the resonator can be done by a small aperture or a small probe or loop [14]. In the rectangular cavity, the incident and reflected passing waves superimpose each other to create standing waves [15]. In

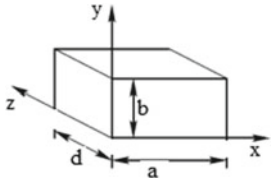


Fig. 1 A rectangular resonant cavity [1]

order to meet the boundary conditions of $E_x = E_y = 0$ on the end walls at $z = 0$ and $z = d$ where E_x and E_y are the electric fields in x and y directions [15], the length of the cavity must be a multiple of half a lead wavelength at the resonant frequency. The cavity is resonant at frequencies where d is a multiple of $\lambda_g/2$, λ_g being the guide wavelength for propagation along the z-axis. A signal entering at $z = 0$ produces a standing wave pattern. Because of the short, the transverse electric field (E_x and E_y) is zero at $z = d$ and at multiples of half-guided wavelength along the z-axis. At certain frequencies, an electric field null occurs at $z = 0$ [16]. Since the transverse electric field is zero at the $z = 0$ plane, the introduction of a shorting plate at that point does not disturb the standing wave pattern [17]. Thus at the selected frequencies, the electromagnetic field within the cavity can be sustained even in the absence of a signal source [18]. The resonant condition is therefore given by

$$d = p\left(\lambda_g/2\right)$$

where d is the length of the cavity and $p = 1, 2, 3 \ldots$, p being any positive integer. The cavity has an infinite number of resonances [18].

The resonant frequency of TE_{mnl} or TM_{mnl},the rectangular cavity mode shown in Fig. 1, is given by [1].

$$f_{mnl} = \frac{c_0}{2\pi\sqrt{\mu_r \varepsilon_r}}\sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2 + \left(\frac{l\pi}{d}\right)^2} \tag{1}$$

where μ_r and the relative permeability and permittivity of the material filling in the cavity volume are the relative permeability and permittivity of the material filling in the cavity volume; c_0 is the speed of light in free space. Here (*m; n* = 0, 1, 2….; *l* = 1, 2, 3…), the subscripts *m*, *n* and *l* represent the number of 1/2 sinusoidal periods in the standing wave pattern along *x*, *y* and *z* axes, respectively. a and d represent width and length, respectively, and b represents the height of the cavity. For $b < a < d$, the TE_{101} mode is the dominant resonant mode with the lowest resonant frequency [20–22].

2.1 Slot with a Shorting Via

Figure 2 shows the slot with shorting via. To excite the resonator, microstrip lines are used in such a way that the cavity's top metal layer is etched through coupling slots, and by using metallic via it gets short-circuited so that maximum magnetic current can be obtained [18]. Shorted vias provide the necessary impedance matching. The coupling slots are initially located at a quarter of the cavity length from the edge of the cavity to maximize the coupling. The slot width is varied while keeping the slot length constant ($\lambda_g = 4$ at the frequency of interest). The present investigations

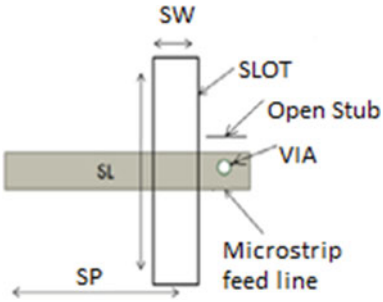


Fig. 2 Explanation of slot with a shorting pin [2]

indicate that the optimized slot width is $\lambda_g = 8$. Slot's position (cavity length/4) should be adjusted in order to achieve frequency at resonant, insertion loss and input impedance [19].

3 Filter Structure

Figure 3 shows the cavity filter in a silicon wafer. The cavity was given info and yield by two microstrip lines via slots [12]. This offers a simple way to combine integrated circuit microwave and integrated circuit monolithic microwave architectures, if required [3]. In microwave-integrated circuit (MIC) technology, the signal processing circuitry is also embedded in the MIC chip [5]. The circuit components should be compatible with this technology. Due to these considerations, silicon became the choice for the substrate material. Micromachining is a standard process in MIC technology. Therefore cavity resonator has been selected as the filtering element. The cavity filter's dimensions are determined by using formulae and concepts discussed above (Table 1).

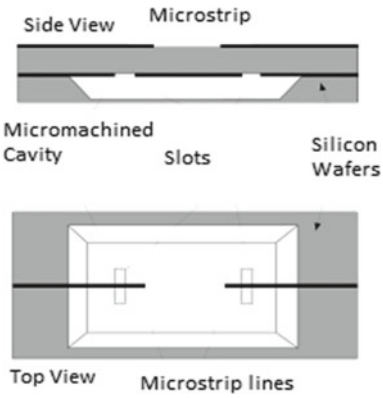


Fig. 3 Cavity filter in silicon wafer [3]

Table 1 Design parameters of cavity filter using shorting via

Design parameters	Dimensions in mm
Cavity length	5
Cavity width	1
Cavity height	0.65
Slot position (SP)	1.25
Slot length (SL)	0.8
Slot width (SW)	0.8

4 Results

We have designed and simulated silicon-based single cavity filter. The whole filter is modeled and simulated in HFSS with the slot duration and cavity size calculated, shown in Figs. 4, 5 and 6 as simulated curves. Parameters are obtained which help to calculate the following data:

$$\textbf{3 dB Bandwidth: } \Delta f = f_u - f_l = 2.30 \text{ GHz} \tag{2}$$

$$\textbf{Center Frequency: } f_r = \sqrt{(f_u * f_l)} = 14.59 \text{ GHz} \tag{3}$$

(f_u and f_l represent the upper edge and lower edge -3 dB frequency points of a bandpass filter).

Insertion Loss $= -1.15$ dB at 15.09 GHz.

Roll-Off: In a bandpass filter, the amplitude of the signal should be reduced sharply in the stopbands. Roll-off is a measure of the sharpness of this reduction. It is the slope of the amplitude–frequency curve. If the amplitude is in dB and the frequency range is 10 times the given frequency then the unit of roll-off is dB/decade. If the frequency

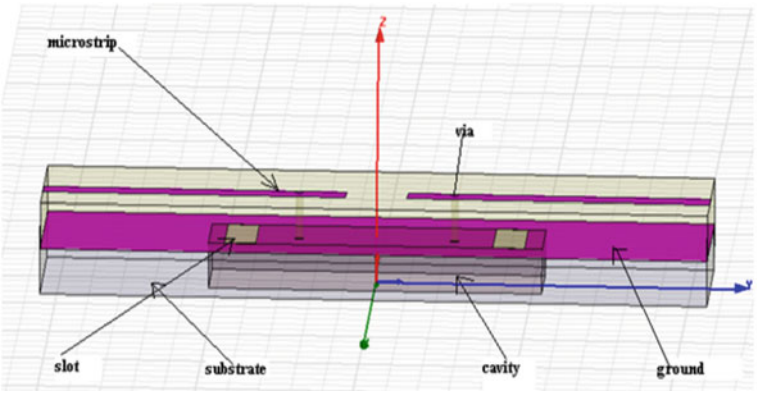


Fig. 4 HFSS modal of silicon-cavity filter with shorting via

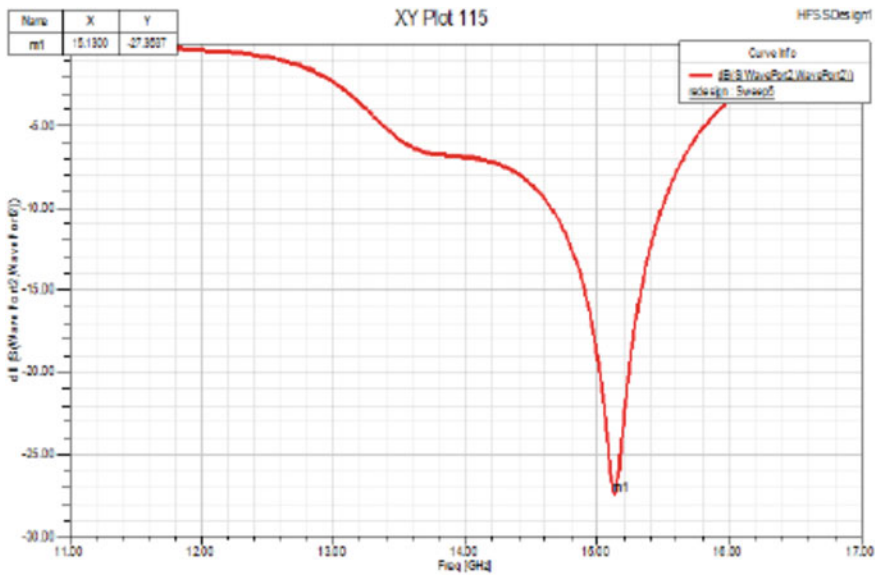


Fig. 5 Simulated S_{11} of single-cavity filter with shorting via

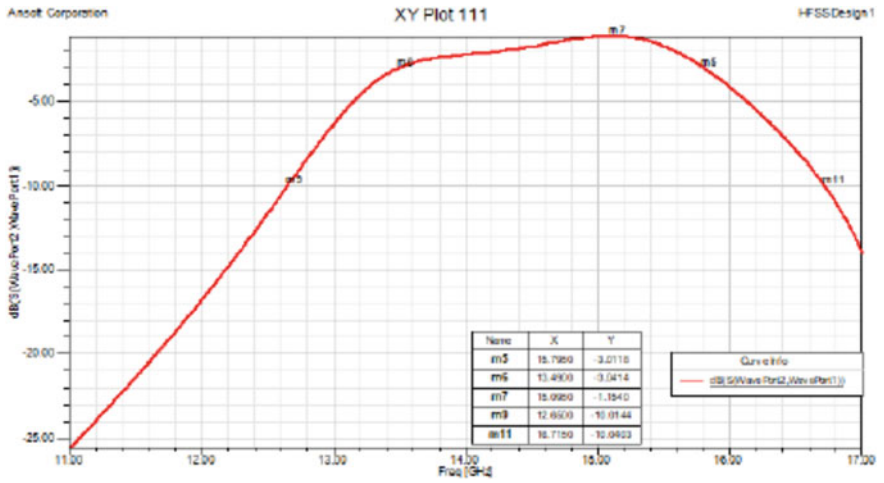


Fig. 6 Simulation curve S_{21} or insertion loss of single-cavity filter with shorting via

range is double the given frequency then the unit is dB/octave. Both these ranges are rather large. It is common to talk about –3 dB and –10 dB points. Therefore, the slope of the amplitude versus frequency curve has been determined between these two points. For the present investigations this gives (Table 2):

Table 2 Results obtained for the bandpass filter

Return loss (S_{11})	−27.36 dB
Resonant frequency	15.13 GHz
3 dB bandwidth	2.30 GHz
Insertion loss (S_{21})	−1.15 dB
Roll-off	8.3 dB/GHz and 7.6 dB/GHz at the lower end and higher end, respectively

Roll-off at lower frequency side = 8.3 dB/GHz, and roll-off at higher frequency side = 7.6 dB/GHz.

Effects of cavity dimensions on the filter properties were investigated. Cavity width was chosen as the physical dimension. Bandwidth, upper edge frequency and lower edge frequency were selected as filter parameters. Cavity width was systematically decreased up to 50% in steps of 10%. Analysis of results was done to optimize cavity width. The results are shown in Fig. 7.

Table 3 shows that bandwidth decreases as cavity width increases. The dependence is linear. This is graphically represented in Fig. 8.

The cavity width is a very important physical parameter of the filter. The cavity width was normalized with respect to guided wavelength as the latter includes wavelength and dielectric constant of the material in the cavity. This led us in a new direction. We plotted the curve between upper edge frequency, lower edge frequency and bandwidth (one by one) to the ratio of resonant frequency (Table 4).

The plot in Fig. 9 shows that the fractional upper edge frequency has a linear dependence on normalized cavity width. This can be mathematically expressed as

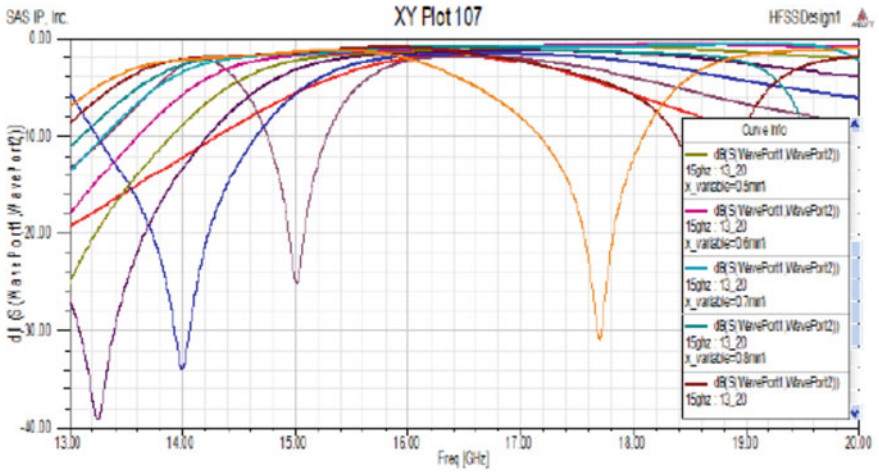


Fig. 7 HFSS simulated curve on varying cavity width

Table 3 Result obtained on varying cavity width

% decrement	w (mm)	f_u(GHz)	f_l(GHz)	Bandwidth (GHz)
0	1	15.79	13.28	2.51
10	0.9	17.10	13.60	3.5
20	0.8	18.14	13.75	4.39
30	0.7	19.00	13.95	5.05
40	0.6	19.57	14.45	5.12
50	0.5	21.90	14.80	7.1

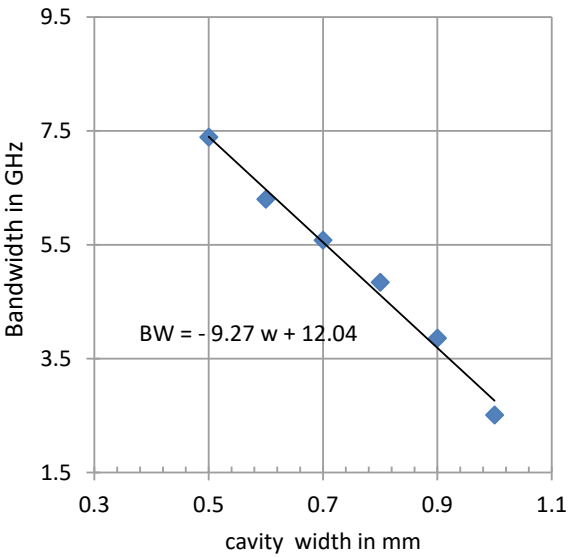


Fig. 8 Bandwidth with respect to variation in cavity width

Table 4 Result obtained on varying cavity width

Cavity width (w/λ_g)	f_u/f_r	f_l/f_r	Bandwidth in GHz
0.05	1.05	0.885	2.51
0.045	1.14	0.90	3.86
0.04	1.20	0.91	4.39
0.035	1.26	0.93	5.58
0.03	1.30	0.96	6.55
0.025	1.46	0.98	7.1

$$f_u/f_r = -14.51\frac{\omega}{\lambda_g} + 1.77 \tag{4}$$

This is an important result. While designing a cavity filter value of w can be chosen to get a particular value of f_u.

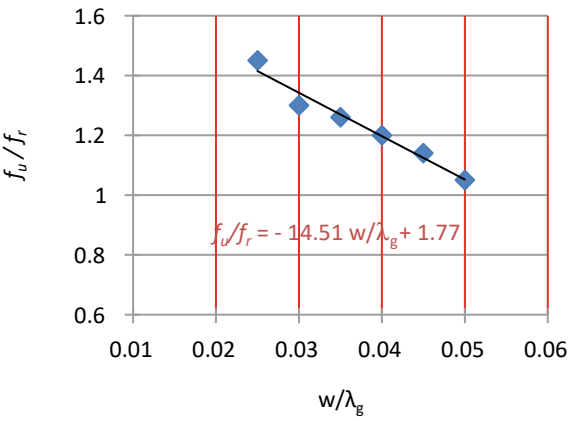


Fig. 9 Dependence of f_u/f_r on w/λ_g

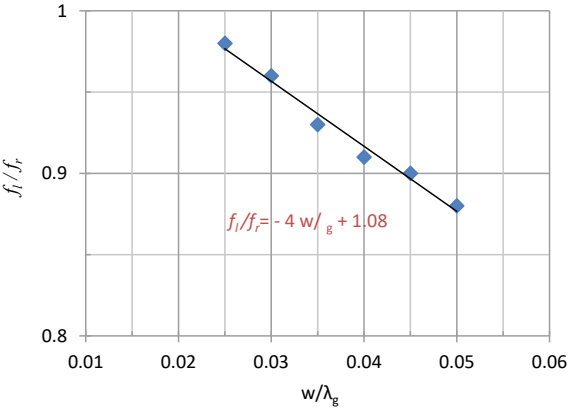


Fig. 10 Dependence of f_l/f_r on w/λ

The plot in Fig. 10 shows that the fractional lower edge frequency has a linear dependence on normalized cavity width. This can be mathematically expressed as

$$\frac{f_l}{f_r} = -4\frac{w}{\lambda_g} + 1.08 \tag{5}$$

This is an important result. While designing a cavity filter value of w can be chosen to get a particular value of f_l.

The plot of Fig. 11 shows that the bandwidth has a linear dependence on normalized cavity width. This can be mathematically expressed as

$$\frac{BW}{f_r} = -11\frac{w}{\lambda_g} + 0.75 \tag{6}$$

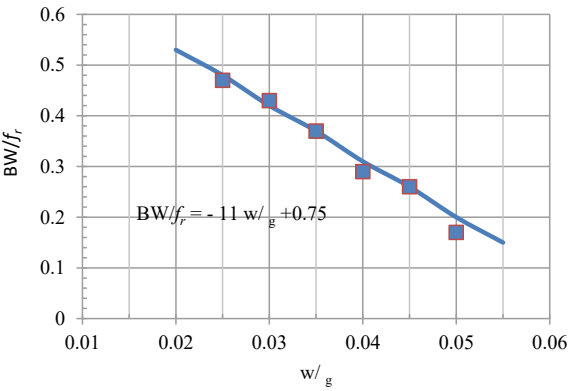


Fig. 11 Dependence of (BW/f_r) on w/λ_g

This is an important result. While designing a cavity filter value of w can be chosen to get a particular value of BW.

5 Validations of Mathematical Formulae

It may be noted from Figs. 6, 7 and 8 that f_u, f_l and BW all decrease as w is increased for a given f_r. For validating formulae (5), (6) and three different values of w/λ_g (0.048, 0.038 and 0.028) were selected. For each value of this parameter, the cavity filter was designed, modeled and simulated in HFSS (Table 5).

From the above figures the values of f_u, f_l and BW have been obtained in each case. These are shown in the table under the headings "Simulated". The same parameters have been estimated using new formulae (5), (6) and for three different values of w/λ_g → 0.048, 0.038 and 0.028. The agreement is quite good as only the first-order dependence of the frequencies on cavity width has been considered. The comparison of patch antenna performance parameters is given in Table 6.

Table 5 Validation of new formulae

w/λ_g	f_u (GHz)		f_l(GHz)		BW (GHz)	
	From new formula	Simulated	From new formula	Simulated	From new formula	Simulated
0.048	16.33	16.83	13.5	13.715	3.37	3.12
0.038	19.87	19.05	15.13	13.95	5.4	5.1
0.028	26.12	23.65	18.53	15.5	8.46	8.2

Table 6 Comparison of patch antenna performance parameter

References	Performance parameters (resonating bands, gain, VSWR and bandwidth)	Material used
Zhang et al. [12]	Operating band = 9.792 GHz, S11 = −18.44 dB, gain = 6.01 dB, bandwidth = 610 MHz	FR4 (lossy), Polyimide (lossy), Glass (lossy), εr = 4.3, 3.5 and 4.82, respectively
Malik et al. [22]	Operating bands = 10/60 GHz	RT/Duroid 5880 with εr = 2.2
[23]	Operating bands = 1.78/1.92 GHz, bandwidth = 138 MHz and 100 MHz	FR4, εr = 4.4, thickness = 1.6 mm
[24]	Operating band = 1.6 GHz, bandwidth = 53.3 MHz, VSWR < 2	FR-4 Epoxy, εr = 4.2, thickness = 1.6 mm
[25]	Operating bandwidth = 1.80–2.17 GHz, gain = 4.5 dBic	RO4003C, εr = 3.38, thickness = 6.2 mm
Our design	Return loss (S11), −27.36 dB Resonant frequency, 15.13 GHz 3 dB bandwidth, 2.30 GHz Insertion loss (S21), −1.15 dB Roll-off, 8.3 dB/GHz and 7.6 dB/GHz at lower end and higher end, respectively	Silicon

6 Discussions and Conclusions

This paper has reported work on a 3-D single, microstrip-fed, slot-coupled micromachined resonant cavity bandpass filter. Filter parameters have been assessed in terms of scattering parameters and bandwidth. The characteristics of resonant cavity-based bandpass filter depend on many parameters such as cavity's length, width and depth, location and size of coupling slots, feed line parameters, and the gap between the two microstrip lines. This paper has only considered the effects of varying cavity width. With these limitations in mind, the agreement between theory and simulations is very good. Further investigations need to be carried out to obtain more accurate predictions. In this paper, a bandpass cavity filter has been simulated for a resonant frequency of 15 GHz having 2.30 GHz bandwidth with 1.15 dB insertion loss over the entire band. The return loss is better than 10 dB, and this shows that a better return loss can be obtained, particularly at lower frequencies. Three new mathematical formulae have been proposed and validated. This paper will be helpful for the basic design of a single-cavity bandpass filter as well as to understand the effect of bandwidth and coupling in microwave devices.

References

1. S. Wong, R. Chen, J. Lin, L. Zhu, Q. Chu, Substrate integrated waveguide Quasi-Elliptic filter using slot-coupled and microstrip-line cross-coupled structures. IEEE Trans. Compon. Packag Manufact. Technol. **6**(12), 1881–1888 (2016). https://doi.org/10.1109/TCPMT.2016.2625744
2. Z. Ma, LIGA cavity resonators and filters for microwave and millimeter-wave applications. Int. J. Electric. Comput. Eng. **55**(8), 10–15 (2007)
3. K. Dhwaj, H. Tian, T. Itoh, Low-profile dual-band filtering antenna using common planar cavity. IEEE Antennas Wirel. Propag. Lett. **17**(6), 1081–1084 (2018). https://doi.org/10.1109/LAWP.2018.2832631
4. J.H. Lee, S. Pinel, J. Laskar, M.M. Tentzeris, Design and development of advanced cavity based dual mode filters using low temperature co-fired ceramic technology gigabit wireless systems. IEEE Trans. Microw. Theory Tech. (2007)
5. A.-R. Moznebi, K. Afrooz, M. Danaeian, P. Mousavi, Four-way filtering power divider using SIW and eighth-mode SIW cavities with ultrawide out-of-band rejection. IEEE Microwave Wirel. Compon. Lett. **29**(9), 586–588 (2019). https://doi.org/10.1109/LMWC.2019.2931115
6. D.R. Hendry, A.M. Abbosh, Analysis of compact triple mode ceramic cavity filters using parallel-coupled resonators approach. IEEE Trans. Microwave Theory Techn. **64**(6), 2529–2537 (2016)
7. B. Li, L. Zeng, Dynamically tunable switch and filter in single slot cavity structure. Sci. Rep. 9, Article no: 14583 (2019)
8. Z. Chen, X. Dai, G. Luo, A new H slot coupled microstrip filter antenna for modern wireless communication systems, in *2018 International Workshop on Antenna Technology (iWAT)* (2018), pp 1–3
9. S. Yang, Q. Chen, A miniaturized bandpass frequency selective surface with high selectivity based on slot coupling. Progr. Electromagn. Res. C **76**, 99–108 (2017)
10. A. Kumar, B. Chaturvedi, Novel CMOS dual—X current conveyor transconductance amplifier realization with current mode multifunction filter and quadrature oscillator. Circ. Syst. Signal Process. 1–28 (2017)
11. H. Ren, X. Li, Q. Zhang, M. Gu, On chip non interference angular momentum multiplexing of broadband light. Science **352**(6287), 805–809 (2016)
12. H. Zhang, W. Kang, W. Wu, Miniaturized dual-band SIW filters using E-shaped slotlines with controllable center frequencies. IEEE Microwave Wirel. Compon. Lett. **28**(4), 311–313 (2018). https://doi.org/10.1109/LMWC.2018.2811251
13. Z.H. Chen, Q.X. Chu, Dual-band reconfigurable bandpass filter with independently controlled passbands and constant absolute bandwidths. IEEE Microwave Wirel. Compon. Lett. 1–3 (2016)
14. X. Zheng, W. Liu, X. Zhang, T. Jiang, Design of dual band-notch UWB bandpass filter based on T-shaped resonator, in *Proceedings of the 2016 Progress in Electromagnetic Research Symposium (PIERS)*, 8–11 August 2016, pp. 4482–4486
15. X. Zheng, T. Jiang, Design of UWB bandpass filter with dual notched bands using E-shaped resonator, in *Proceedings of the IEEE/ACES International Conference on Wireless Information Technology and Systems*, 13–18 March 2016, pp. 1–2
16. D.W. Wuyan, Design of double mode microstrip pass filter for square loop patch. Electron. Technol. Appl. **12**, 102–105 (2017)
17. C. Zhong, C. Liuji et al., Design electronic components and materials for ceramic cross-reference microstrip pass filters. **1**, 48–50 (2017)
18. M.C. Tang, T. Shi, S. Chen, Dual-band bandpass filter based on a single triple mode ring resonator. Electron. Lett. **52**(9), 722–724 (2016)
19. Y. Peng, L. Zhang, J. Fu et al., Compact dual-band bandpass filter using coupled lines multimode resonator. Microw. Wirel. Compon. Lett. **25**(4), 235–237 (2016)
20. P. K. Malik, M. Singh, Multiple bandwidth design of micro strip antenna for future wireless communication. Int. J. Recent Technol. Eng. **8**(2), 5135–5138 (2019). ISSN: 2277–3878. https://doi.org/10.35940/ijrte.B2871.078219

21. P.K. Malik, H. Parthasarthy, M.P. Tripathi, Alternative mathematical design of vector potential and radiated fields for parabolic reflector surface, in *Advances in Computing, Communication, and Control*, ed. by S. Unnikrishnan, S. Surve, D. Bhoir, vol. 361, ICAC3 2013. Communications in Computer and Information Science (Springer, Heidelberg, 2013)
22. P.K. Malik, H. Parthasarthy, M.P. Tripathi, Analysis and design of Pocklingotn's equation for any arbitrary surface for radiation. Int. J. Sci. Eng. Res. **7**(9), 208–213 (2016). ISSN 2229–5518
23. N. Shaik, P.K. Malik, A Retrospection of Channel Estimation Techniques for 5G Wireless Communications: Opportunities and Challenges. Int. J. Advan. Sci. Technol, **29**(05), 8469–8479 (2020). ISSN: 2005-4238, June 2020
24. A. Rahim, P.K. Mallik, V.A. Sankar Ponnapalli, Fractal Antenna Design for Overtaking on Highways in 5G Vehicular Communication Ad-hoc Networks Environment. Int. J. Eng. Advan. Technol (IJEAT). **9**(1S6), 157–160 December (2019). ISSN: 2249–8958, https://doi.org/10.35940/ijeat.A1031.1291S619
25. A.S. Duggal, P.K. Malik, A. Gehlot, R. Singh, G.S. Gaba, M. Masud, J.F. Al-Amri, A sequential roadmap to industry 6.0: Exploring future manufacturing trends. IET Communication 1–11 (2021). (ISSN: 1751-8636), https://doi.org/10.1049/cmu2.12284

Use of Smart Mobile Applications with IoT in Diseases Prediction System for Apple Orchards

Karuna Sheel and Anil Sharma

Abstract Internet of Things or IoT refers to the connection of physical devices or things with network. This technology has a huge contribution now a days in the diverse application areas. In rapidly changing scenario, mobile applications are considered as the best way for marketing any work or business digitally. The present study focuses on the use of smart mobile applications with IoT. Mobile applications running on smart devices are important Supporter for internet of things. The study focuses on the integration of physical objects and external services like google firebase, for data and application management. The system is proposed to integrate the IoT services with smart mobile application which has the ability to detect and predict the conditions in which diseases can appear in apple.

Keywords Internet of Things [IoT] · Mobile application · Google firebase · Apple · Apple diseases

1 Introduction

Introduction to smart mobile application development has driven data gathering, information presentation and user interaction in various ways. When it comes to develop Smart mobile application with internet of things it becomes effective in the mean of interface, connection with the resources of IoT devices [1].

After the internet, Internet of things is viewed as a technology and economic trend in the industry of global information. It is a network which is enough intelligent and connects all physical things to the internet. The purpose of IoT is for communicating through information sensing devices and exchanging information in accordance with agreed protocols. It accomplishes the goal of locating, tracking, managing, intelligent identifying and monitoring the things [2].

K. Sheel · A. Sharma (✉)
School of Computer Science and Engineering, Lovely Professional University, Phagwara 144411, Punjab, India
e-mail: anil.19656@lpu.co.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_3

Google Firebase is an online database provider, that allows developers to connect sensor data with mobile application. Cloud firestore and Realtime database are some essential services which is provided by Firebase to store and manage the app data online.

IoT enables the end-to-end development and deployment of environment awareness by providing tools for collection of information on smart devices and tools for communication with external resources [3]. It offers facilities like backend and a real-time database.

Firebase provides an application programming interface to the application developers. The API allows synchronization of application data across clients and stored data on its cloud. Google providing the client libraries which enables integration with IOS, Android and JavaScript applications [4]. It simplifies easy and secure file transfer regardless of network quality for the Firebase apps. It is supported by Google Cloud Storage which is cost-effective object storage service. The developer can use firebase to store audio, video, images, and other user-generated content [5].

Apple farmers cannot exactly identify the level of humidity and temperature in the apple orchard. They can only feel it by themselves under temperature and humidity variation conditions. So, experience is the only thing which can play a bigger part on their daily farming methods. If the humidity level is too high, they left the orchard in sun light and if too low then they give water to the soil [6].

There are various types of mobile applications available now a days which support IoT components, for e.g.—BLYNK app. BLYNK or any other application are designed for some specific purpose. So instead of using these types of applications, developers can create their own mobile applications for more specific purpose or according to user's requirement [7]. For ex. Variation in environmental factor like temperature and humidity can cause to various diseases in plants and crops. So, there could be a possible solution to monitor the field's temperature and humidity level to control the diseases in crop. For this IoT components like DHT11 or DHT22 [Temperature and Humidity sensor], ESP8266 wi-fi transreceiver can be used with Arduino UNO board [8]. Google firebase provide the platform to connect and manage the mobile application data online with IoT sensors [9].

2 General Design of the System

An IoT Based Mobile Application System is proposed in this study which is based on Internet of Things Architecture. The architecture of IoT is becoming increasingly standardized with the development of IoT. Many researchers/ authors has written about IoT and its architecture [10].

With reference to other IoT based environment monitoring system, this system is characterized by its ease of use and mobility. So, it is a need to combine the monitoring of the system with IoT technology. The term monitoring means to get the notification based on prediction about environment variation that can affect the apple crop [11].

In the proposed study the system adopts an architecture containing three layers:

1. Input Layer
2. Transmission Layer
3. Application Layer

This embedded system for monitoring environment variables and predicting the diseases appearing conditions in apple orchard is mainly based on following IoT components [12].

1. Sensors

   Use of sensors is done at input layer of the proposed system. Sensors take the input from apple orchard. Temperature and humidity used as input for sensors in analog form and then will convert in digital form (Fig. 1).

2. Wi-Fi transreceiver

   Wi-Fi transreceiver is used at transmission layer which is a highly integrated chip used for wireless connection with android application. Here ESP8266 is used to transmit the sensor's data to Google firebase real time database (Fig. 2).

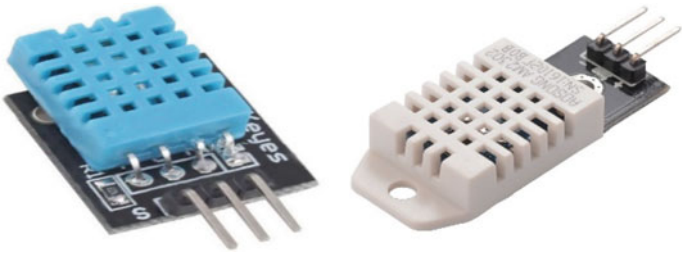
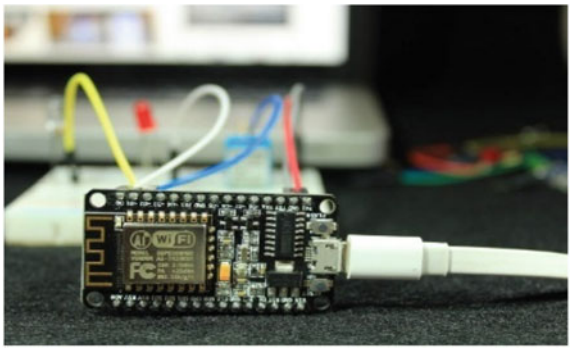


Fig. 1 DHT-11 and DHT-22 Sensors

Fig. 2 ESP8266 Wi-Fi Transreceiver



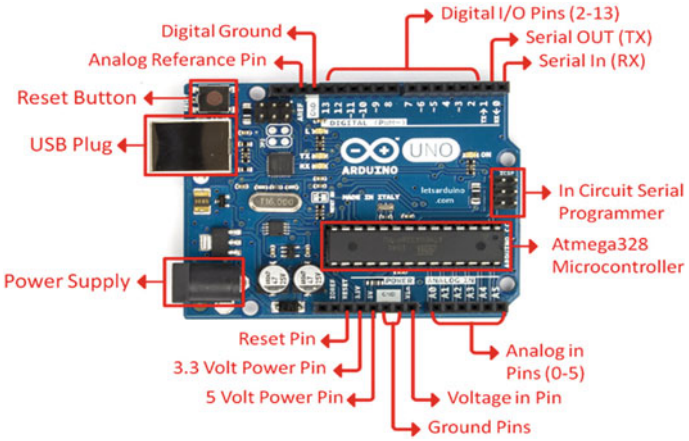


Fig. 3 Arduino UNO microcontroller

3. Arduino

Arduino is a general purpose IoT board used as microcontroller for this embedded system and also used at input layer. The purpose for this is to take input from sensors and transmit to real time database using transreceiver (Fig. 3).

The prediction is done through Android smartphone at application layer [13].

A. Hardware Description

There are four parameters in an apple orchard which may be responsible for diseases appearance—variation in temperature and humidity, height of the orchard and location of the orchard (either shady or sunny).

Based on these four parameters we can predict the diseases appearing condition by monitoring temperature, humidity level of the orchard.

Monitoring of orchard be made up of sensors for temperature and humidity, Arduino Uno Microcontroller and ESP8266 Wi-Fi transreceiver.

The out from sensors works as input to microcontroller and sent through Wi-Fi transreceiver to Google firebase real time database [14]. Below figure shows the schematic diagram of temperature and humidity monitoring using IoT with mobile application (Fig. 4).

B. Software Description

The program is written in Arduino IDE. This part is designed to process the temperature and humidity values, monitoring the orchard. The software comprises the various values of sensors, send temperature and humidity level of the orchard to Arduino UNO microcontroller. Then continue to send the values to Google firebase using ESP8266 transreceiver. Android application receives the values send by microcontroller to firebase [Google Realtime database] [5].

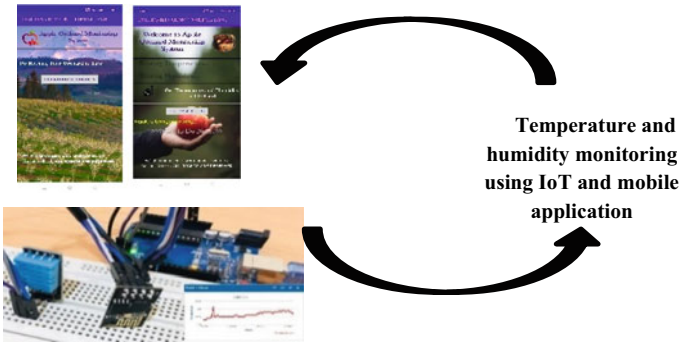


Fig. 4 Field's monitoring using IoT components and mobile application

2.1 Hardware Design

The Hardware part consist of 3 modules. Arduino UNO microcontroller, DHT-11/ DHT-22 temperature and humidity sensor, ESP8266 Wi-Fi transreceiver. The microcontroller is used for embedding sensors and wireless transmission. It is used to read the values from sensors, and transmit these values to real time database of google firebase [15]. Then these values are ruled by android application.

First, microcontroller get the analog voltage signal from the sensor and convert it to digital signal. After microcontroller receive the digital signal, microcontroller send the value from sensor to the Android via ESP8266 transreceiver and wireless connection [16].

2.2 Mobile Application Design

2.2.1 Android Studio

Android studio is a capable platform to develop mobile applications. Here it is used with java to develop the android application [17]. Using android studio there are four files used to demonstrate the orchard monitoring method. First is MainActivity.java with activity_main.xml file which is used for getting temperature and humidity from Google Realtime database. Use of Fuzzy rules are done on button event to get the prediction on the basis of above mentioned four parameters [18]. Second is nextPage.java with activity_next_page.xml which is used to predict the orchard condition either safe or need some precaution [19].

2.2.2 Google Firebase

Andrew Lee and James Tamplin back in 2011 established Firebase but formally it was launched in April 2012. For building portable and web applications for business which need real-time database, firebase framework is useful. It is used when one user updates any single record in the database then that update must be transferred to every user immediately. It provides a simple and integrated environment for many applications at a time. When it comes to the application development, Firebase handles most of the server-side work. With the help of google firebase services Realtime database can be created and managed. This database can be further linked with android application to utilize the stored data [20].

2.2.3 Fuzzy Logic

Fuzzy logic is a method of reasoning which imitates the way of decision making and involves all possible intermediate digital values between YES and NO. FL provides the way to identify the possible values of temperature and humidity of the apple orchard that can cause the diseases [21] (Figs. 5 and 6).

An algorithm has been designed to get the sensor values from orchard and sends the input values to android based prediction system.

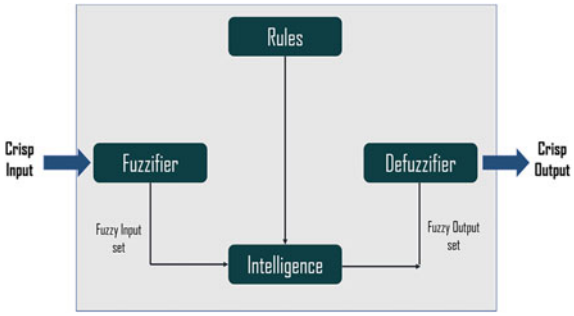


Fig. 5 Fuzzy logic architecture

LP	X is Large Positive
MP	X is Medium Positive
S	Small
MN	X is Medium Negative
LN	X is Large Negative

Fig. 6 Fuzy rule set

```
while True do
    collect temperature and humidity values from Sensors;
    purify received data;
    aggregate and analyze the data;
    if connection done then
    formulate the data for disease prediction;
    send formulated data to smartphone;
    if prediction done then
      switch to monitoring field again;
    else
      try to send the orchard's temperature and humidity level;
    endif
else
    try to match with smart device;
endif
endloop
```

Algorithm: Steps to collect data from sensors and processing by mobile application [22].

Here Fuzzy rules are applied to temperature and humidity values. Fuzzification and defuzzification are applied to predict the intermediate digital values between the diseases appearing parameters [23].

2.2.4 Real Time Database

Cloud hosted database is called a Real-Time Database. Data is stored as JSON and for each associated client it synchronized continuously. Real-Time Databases are used when user demands cross platform application, developed with iOS or Android, to update each new data instance. Real-Time database feature provide feature to developers to skip the step of developing a database. Most of the application's backend could be handled by Firebase [24].

3 Conclusion

As it is known now a days, every person wants to control the "things" by just a click or touch, this study reveals the use of mobile application with IoT to predict the forthcoming diseases appearance. The things are sensed and monitored by IoT components and related input data will be managed and controlled by supported

mobile application. Between IoT components and mobile application a real time database by google is introduced which is used to store the real time values inputted from sensors and then used by the mobile application to predict the possible conditions in which a disease may appear. In a nutshell the present study describes the role of mobile application to control the things with the help of IoT Components.

References

1. S. Albishi, B. Soh, A. Ullah, F. Algarni, Challenges and solutions for applications and technologies in the Internet of Things. Proc. Comput. Sci. (2017). https://doi.org/10.1016/j.procs.2017.12.196
2. S. Chen, H. Xu, D. Liu, B. Hu, H. Wang, A vision of IoT: applications, challenges, and opportunities with China perspective. IEEE Internet Things J. **1**(4), 349–359 (2014). https://doi.org/10.1109/JIOT.2014.2337336
3. S. K. Datta, C. Bonnet, Connect and control things: integrating lightweight IoT framework into a mobile application, in *Proceedings—NGMAST 2015 9th International Conference Next Generation Mobile Applications, Services* and *Technologies,* pp. 66–71 (2016). https://doi.org/10.1109/NGMAST.2015.23
4. L. Goswami, Power line transmission through GOOGLE firebase database, no. Icoei (2020), pp. 415–420
5. C. Khawas, P. Shah, Application of firebase in android app development—a study. Int. J. Comput. Appl. **179**(46), 49–53 (2018). https://doi.org/10.5120/ijca2018917200
6. B. Basannagari, C.P. Kala, Climate change and apple farming in Indian Himalayas: a study of local perceptions and responses. PLoS ONE **8**(10), 1–6 (2013). https://doi.org/10.1371/journal.pone.0077976
7. I. Mohanraj, K. Ashokumar, J. Naren, Field monitoring and automation using IOT in agriculture domain. Proc. Comput. Sci. **93**(September), 931–939 (2016). https://doi.org/10.1016/j.procs.2016.07.275
8. F.A. Khan, A.A. Ibrahim, A.M. Zeki, Environmental monitoring and disease detection of plants in smart greenhouse using internet of things. J. Phys. Commun. **4**(5) (2020). https://doi.org/10.1088/2399-6528/ab90c1.
9. S.A. Jaishetty, R. Patil, Iot sensor network based approach for agricultural field monitoring and control. Int. J. Res. Eng. Technol **05**(06), 45–48 (2016). https://doi.org/10.15623/ijret.2016.0506009.
10. S. Madakam, R. Ramaswamy, S. Tripathi, Internet of Things (IoT): a literature review. J. Comput. Commun. **03**(05), 164–173 (2015). https://doi.org/10.4236/jcc.2015.35021
11. C. Feng, H. Wu, H. Zhu, X. Sun, The design and realization of apple orchard intelligent monitoring system based on internet of things technology. Adv. Mater. Res. **546–547**, 898–902 (2012). https://doi.org/10.4028/www.scientific.net/AMR.546-547.898
12. S. Khummanee, S. Wiangsamut, P. Sorntepa, C. Jaiboon, Automated smart farming for orchids with the Internet of Things and fuzzy logic, in *Proceeding 2018 3rd International Conference on Information Technology, InCIT 2018* (2018), pp. 1–6. https://doi.org/10.23919/INCIT.2018.8584881
13. N. Magaia, P. Gomes, L. Silva, B. Sousa, Development of mobile IoT solutions: approaches, architectures, and methodologies. IEEE Internet Things J. (2020)
14. N. Chatterjee, S. Chakraborty, A. Decosta, A. Nath, Real-time communication application based on Android using Google Firebase. Ddfsdfdsfsd **6**(4), 74–79 (2018) www.ijarcsms.com
15. J. Gubbi, R. Buyya, S. Marusic, M. Palaniswami, Internet of Things (IoT): a vision, architectural elements, and future directions. Futur. Gener. Comput. Syst. **29**(7), 1645–1660 (2013). https://doi.org/10.1016/j.future.2013.01.010

16. M.J. Aitkenhead, D. Donnelly, M.C. Coull, E. Hastings, Innovations in environmental monitoring using mobile phone technology—a review. Int. J. Interact. Mob. Technol. **8**(2), 42–50 (2014). https://doi.org/10.3991/ijim.v8i2.3645
17. F. Afandi, R. Sarno, Android application for advanced security system based on voice recognition, biometric authentication, and Internet of Things, in *Proceeding—ICoSTA 2020 2020 International Conference on Smart Technology and Applications: Empowering Industrial IoT by Implementing Green Technology for Sustainable Development* (2020). https://doi.org/10.1109/ICoSTA48221.2020.1570615292.
18. K. Anand, C. Jayakumar, M. Muthu, S. Amirneni, Automatic drip irrigation system using fuzzy logic and mobile technology, in *Proceedings—2015 IEEE International Conference on Technological Innovations ICT Agriculture and Rural Development TIAR 2015*, no. Tiar (2015), pp. 54–58. https://doi.org/10.1109/TIAR.2015.7358531.
19. M. Poongothai, P.M. Subramanian, A. Rajeswari, Design and implementation of IoT based smart laboratory, in *2018 5th International Conference* on *Industrial Engineering* and *Applications ICIEA 2018* (2018), pp. 169–173. https://doi.org/10.1109/IEA.2018.8387090
20. W.J. Li, C. Yen, Y.S. Lin, S.C. Tung, S.M. Huang, Just IoT Internet of Things based on the firebase real-time database, in *Proceedings—2018 IEEE International Conference Manufacturing*, *Industrial & Logistics Engineering SMILE 2018*, vol. 2018 (2018), pp. 43–47. https://doi.org/10.1109/SMILE.2018.8353979
21. Sayantini, What is Fuzzy Logic in AI and What are its Applications? https://www.edureka.co/blog/fuzzy-logic-ai/. Accessed 02 Apr 2020
22. A.F. Santamaria, P. Raimondo, F. De Rango, A. Serianni, A two stages fuzzy logic approach for Internet of Things (IoT) wearable devices, in *IEEE International Symposium on Personal*, *Indoor and Mobile Radio Communications*, pp. 1–6 (2016). https://doi.org/10.1109/PIMRC.2016.7794563
23. S. Dubey, R.K. Pandey, S.S. Gautam, Literature review on fuzzy expert system in agriculture. Int. J. Soft Comput. Eng. **2**(6), 289–291 (2013)
24. M.A. Mokar, S.O. Fageeri, S.E. Fattoh, Using firebase cloud messaging to control mobile applications, in *Proceedings of International Conference on Control, Computing and Electronics Engineering 2019, ICCCEEE 2019*, pp. 2–6 (2019). https://doi.org/10.1109/ICCCEEE46830.2019.9071008.

Review on Miniaturized Flexible Wearable Antenna for Body Area Network

Utkarsh Pandey, Narbada Prasad Gupta, and Praveen Malik

Abstract Body area network (BAN) is a small-range wireless network consisting of devices placed in, on and near about human body. It consists of wearable or implanted electronic devices that transmit ID or sensor data to a gateway device. It uses an electric field, electric current or electromagnetic communication technology. So there can be three different types of data transmission techniques in a BAN. These are really useful in various medical applications, such as monitoring the functionality of implants, tracking the health record of elders and various testing like endoscopic exams. But if we want this technology to be a thriving industry, we require to define its scope and potential in a better way, and a clear vision for its future is also very much needed. BAN is capable of providing highly useful functions. It is also a handy, unobtrusive, low-power system that consists of high data security. Seeing its importance and use in upcoming scenarios we have presented a review paper on the various challenges encountered while working through this body area network and the different ways to get the least errors by introducing various configurations of the microstrip patch antenna. Performance analysis of wearable antenna under BAN in various possible situations is discussed through this review study which will be very helpful in designing a miniaturized flexible wearable antenna working under BAN.

Keywords WBAN (Wireless Body Area Network) · Wearable device · Compact · Flexible · Performance analysis

1 Introduction

The present era is of wearable antenna because of its application in almost every field, for example, industrial application, study, biomedical application, military application and many more. It works under body area network. This wearable antenna is used for transmitting and receiving signals mostly in the ISM band (2.4 GHz). It is also used in transmitting critical readings of the patient to the doctor's device so that immediate diagnosis can be provided. The best feature of this wearable antenna

U. Pandey (✉) · N. P. Gupta · P. Malik
School of Electronics and Electrical Lovely Professional University Punjab, Jalandhar, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_4

is its simple design and its integration on the patient's clothes or sometimes inside the body as well. Our work is strictly focused on deep analysis of various antenna design methodologies to construct a flexible wearable antenna that can be mounted on textile material for body-centric wireless communication. The antennas designed earlier were using non-flexible substrate material such as RT/Duroid, FR-4, foam etc., which were having medium to high dielectric constant. Because of these high values of dielectric constant the surface wave losses were also high. Now with the use of these flexible textile materials substrate dielectric value is reduced, so the surface wave losses are also reduced and the efficiency of antenna parameters increases. These textile material-based antennas are closer to the body and flexible in nature, so contact of these antennas remains much longer to the body, and some of the antennas reviewed in this paper are even planted inside the body as well. These antennas can easily provide a more accurate value for body parameters such as body temperature or heart rate and even other parameters by the use of some sensors or transducers. The role of the material used for substrate matters is extremely vital, so we have to take this selection of substrate material as a key parameter in our antenna designing. Miniaturization can be understood as the process of manufacturing smaller and smaller size of mechanical, optical or electronic devices. The silicon chip can be the best example to understand the benefits of miniaturization. Miniaturization, low weight and low volume are important features in any kind of antenna design. In medical science we are already using pacemakers and defibrillators which require this miniaturization. In the present era, the requirement of designing the low-profile multiband or wideband miniaturized antenna has significantly increased [1] (Fig. 1).

The above figure represents the schematic diagram for wearable antenna and its various uses. Through this diagram various sensors have been linked to a mobile antenna. Here we have different antennas on the human body, which is shown in the diagram as various sensors that send body control unit, and from there it can be sent to desired server or receivers. We have so many advanced optimization techniques existing now for antenna size miniaturization but it is still a difficult task to shrink the physical dimensions without considerably affecting the antenna performance parameters and researchers are working for finding more and more ways to accomplish this task for making antenna design as little as possible [2]. The present-day antenna requires either a wideband of operation or multiple numbers of bands of operation. For this very purpose meta-materials are used in designing these miniaturized antennas which provide the capability of working under multiband frequency ranges. A modern antenna can work over Bluetooth, Wi-Fi, IR wireless communication, microwave radio, broadcast radio etc. As we decrease the size of the antenna the gain of the antenna also decreases. To enhance the gain of these multiband miniaturized antennas without affecting other parameters, we use the technique of integration of frequency selective surface [3]. This is also an area where a lot of opportunities exist and researchers are working upon it. We have tried to discuss here some of the miniaturized antennas in our study as well [4, 5]. Flexible electronics can be understood as the selection of those materials whose mechanical properties include bent, wrinkled collapsed or stressed, and these are suitable for the application of modern electronics to various non-flat and real scenarios such

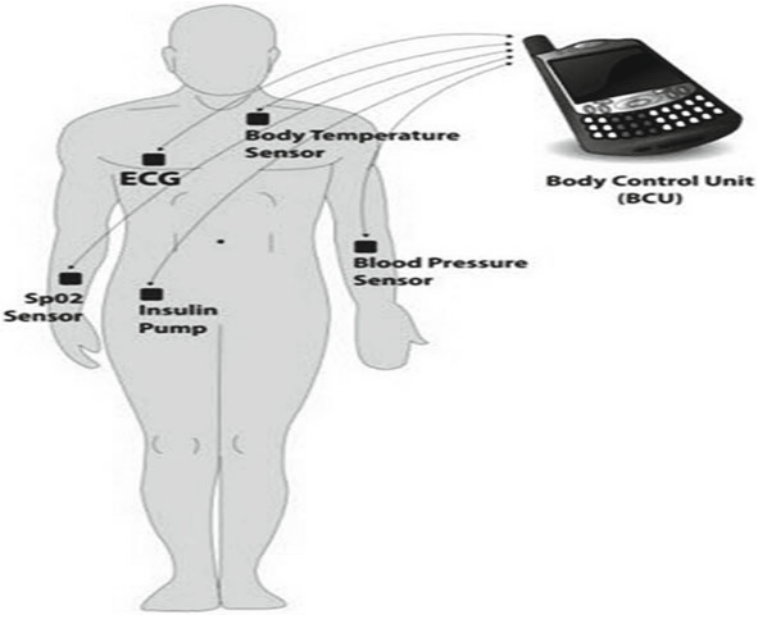


Fig. 1 Body area network application

as the shape of the human body. They offer various advantages such as low-cost manufacturing, lightweight, ease of operation and inexpensive flexible substrates [6]. In recent times researchers have focused a lot on this technology because of its simple configuration, passive operation, low cost and multimode sensing. These antennas are able to perform dual operations of sensing and can be employed easily by minimizing the number of components. These textile wearable antennas are becoming more and more popular for on-body applications as they have a better ability to sense moisture deformations, human body moments and to supervise as well as monitor human health [7]. These antennas are having better integration capability with cloths and offer so many features such as washing ability lightweight and comfort to the users. Medical studies and healthcare has taken a giant bounce ahead this decade with the creation of wearable scientific gadgets inside the market. Awareness has been provided to the masses about the presence of such gadgets for use. These devices are widely used by health-conscious people around the world. Slowly but surely, this knowledge of the presence of such device, its functionality and various advantages related to this are going to make them more and more popular and things to have essentially with everyone [2]. Early detection of critical diseases is now very much possible, all thanks to the researchers for their endeavors in the field of wireless transmission and remote sensing technology who combined these two things and have come up with such a nice solution.

2 Contribution

The motive of the presented study is to present different tactics proposed to design miniaturized flexible body-worn antennas, selection of material for implanting body-worn devices and different challenges present in designing the wearable antennas working under body area network. Our contribution can be mentioned through the following points.

- Vital literature has been reviewed to present the recent research progress in miniaturizing the body-worn antenna design.
- Through the presented study we wanted to bring your focus to the research gap still present in the field of designing wearable antennas.
- Through the presented study we also wanted to present the future scope in the field of designing flexible, miniaturized wearable antenna design as multiband or ultra-wideband antenna can be produced with the use of different technologies.

3 Literature Review

Selection of substrate material and design methodologies for wearable antennas is the biggest challenge in designing a novel antenna, so we are presenting some of the pioneering work done in this area by researchers through our literature review and we also see some of the modern techniques through which miniaturization is achieved. We would also take a look at how to get an ultra-wideband antenna so that our antenna could be used for various purposes. We studied the antenna designed by keeping flexible, ultra-wideband miniaturized characteristics in our focus. Let's see how we started with this flexible textile wearable antenna, then the design methodology for the conventional wearable antenna and at last, we will discuss some of the recent antennas and that will done with the tabular analysis of studied papers.

3.1 Flexible Textile Antenna Designs

The general textile material is often consumed in the procedure of selecting the substrate material for antenna design as it is both flexible and economical as well. As this antenna has the nature of flexibility, so these antennas are more suitable in body-worn applications. In the case of wearable mobile antenna application circular polarization is required. The first paper having circular polarization for wearable antenna is reported in [8]. The radiation pattern of circular polarized antenna is beautifully explained as to how energy is radiating in all directions [9]. The signal reception of a circular polarized antenna is independent of orientation.

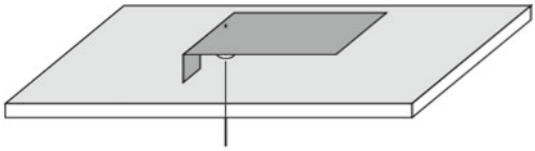


Fig. 2 A simple configuration of PIFA used in mounting on shirt's sleeve [8]

3.2 Conventional Body-Worn Antenna Designs

Basically, wearable antenna in its earlier stages was simply focused on monopole antenna, coplanar dipole antenna, microstrip patch antenna and planar inverted-F antenna designs. In its earlier stages it was very common that antenna was designed on PCB board using FR-4 substrates. This type of antenna was accepted at that time as it was economical and simple in manufacturing. An antenna that is easily wearable on the sleeve of a shirt has been reported through one of the classic works done [8]. Conceptually, the PIFA structure is generally a quarter-wave ($\lambda/4$) monopole antenna whose folded part is parallel to the ground plane. It can be easily understood through the configuration given in Fig. 2.

3.3 Related Works (in Recent Times)

Now the work done in the field of wearable antenna designing having the same focus as per our studies is summarized here with the help of a table given below, where we selected some papers from recent years in which miniaturization, flexibility and various other important parameters related to wearable devices are focused (Table 1).

4 Discussion

- Ritesh et al. in their paper presented a miniaturized antenna with dual coating and quad bands for wireless standards with increased gain. The size of the antenna is $0.239 \times 0.351 \times 0.0127\lambda$ mm^3, at a lower frequency of 2.39 GHz. This structure is having a slotted hexagonal form radiating component with metamaterial-inspired SRR cells for getting quad-band characteristics. The suggested designed antenna is having consistent and stable radiation patterns along with less cross-polarization. Its radiation efficiency and gain are also in the acceptable range. To get better impedance matching in selective bands, FSS (frequency selective surface) has been designed and implemented for enhanced gain value [3].
- Merih et al. in the presented paper proposed a miniaturized antenna with new innovative resonator geometry working under MICS and ISM bands for biometry applications. In this study they have taken three layers of human phantom model.

Table 1 Comparative analysis of reported microstrip patch antenna (MPA) studies

References	Antenna patch design, size (mm^2)/(mm^3) And number of operating bands	Remark
Ritesh et al. [3]	1. Slotted hexagonal form radiating component with metamaterial inspired SRR cells with FSS 2. $-30 \times 44 \times 1.6$ 3. Quad-band	1. For better impedance matching at selective bands FSS is designed and implemented for enhancing the gain value 2. Slotted hexagonal form radiating component with metamaterial inspired SRR cells for getting quad-band characteristics 3. This FSS is working in the band of 3.6–10.1 GHz and provides the gain of about 4–5 dB 4. FR4 thickness = 1.6 mm, $\varepsilon r = 4.4$, $\tan \delta = 0.02$ 5. Analyzed parameters: Band of operations—WLAN/WiMAX/ WAVE/C/X Band, (2.4/3.5/5.8/7.9) Gain—2.63/2.58/2.82/2.99 dB, Radiation efficiency (%)—33.5, 38.8, 84.4, 72.8 Bandwidth (in%)—8.80/38.78/15.54/35.45
Merih et al. [10]	1. L-shaped transmission line fed antispiral resonator structure 2. $15 \times 15 \times 1.92$ *3*. Dual-band	1. L-shaped transmission line fed anti spiral resonator structure that is loaded with spiral resonator to the end so that effective electrical length can be increased 2. SAR values at resonant frequencies are found in the appropriate range. The effectiveness of the three-layered human tissue model is pointed out with experimental and simulated results in free space 3. Thickness = 0.64 mm, Rogers 3210 $\varepsilon r = 10.2$ and loss tangent of 0.003 4. Analyzed parameters: Resonate at 403 MHz with 100 MHz B.W. and at 2.4 GHz with 360 MHz B.W., Gain = −12.25 dBi, −12.4 dBi, SAR = 1.6 mW

(continued)

Table 1 (continued)

References	Antenna patch design, size (mm^2)/(mm^3) And number of operating bands	Remark
Amiya et al. [11]	1. Proximity-coupled fed microstrip circular patch antenna 2. Substrate of 30 mm × 45 mm radius of the radiating patch = 13.005 mm 3. Single-band	1. This proposed antenna is having allure features regarding 5G applications such as machine-to-machine communication, Internet of Things etc. 2. The proposed antenna has an impedance of 50.99−j 0.15 Ω. This impedance matching is excellent which suits the maximum power transfer 3. Rogers RT/Duroid 5880 $\varepsilon_r = 2.2$ thickness = 1.5 mm and FR4 epoxy $\varepsilon_r = 4.4$ thickness = 1.6 mm 4. Resonate at 3.5 GHz, Gain = 5.8 dB, Return Loss = −40.28 dB, VSWR = 1.02 B.W. = 200 MHz, Impedance = 50.99 – j0.15
Siddat et al. [12]	1. T-shape msp placed over 6 six-layer phantom model 2. 29.99 mm × 29.99 mm × 0.59 mm 3. Single-band	1. Antenna is designed for the detection of brain tumors. It is placed above a whole head phantom model consisting of six layers. Numerous performance parameters have been checked for the usual state and tumor-affected state 2. By varying positions of tumor regards to the antenna. The position of the tumor can be detected by analyzing data, a tumor of radius of 5 mm with permittivity and conductivity of 7 S/m 55 and 7 S/m has been taken into concern for the purpose of simulation process 3. FR4 εr = 4.35 thickness = 0.55 mm 4. Resonate at 0.91 GHz, Directivity = 3.12 dB Return loss = −35.58 dB, SAR = 0.332 W/kg

(continued)

Table 1 (continued)

References	Antenna patch design, size (mm^2)/(mm^3) And number of operating bands	Remark
Chowdhury et al. [2]	1. Pentagonal-shaped patch with hexagonal slot 2. 33 × 23 × 1 3. Single-band	1. The size was taken very minute deliberately because the purpose of this antenna design was to detect this tumor in its early stages only. The possibility of getting more prominent results is there with an antenna having better gain and efficiency value 2. This antenna is placed on a human head phantom for getting the values of different parameters of the antenna. The value return loss we got for the two cases was different for normal head phantom and tumor-affected head phantom. The existence of brain tumor can be understood through the shift of the resonant frequency to get maximum return loss value. The measured difference is 18 MHz between the actual head phantom and tumor-affected head phantom 3. FR4 lossy εr = 4.4 thickness = 0.8 mm 4. Operating in ISM band (2.4–2.4835 GHz). Return loss = −31.42 dB, VSWR = 1.058 SAR = 0.3877 W/kg
Saraereh [13]	1. F-shaped msp with rectangular feed 2. 27.7 × 19.4 3. Dual-band	1. In this the author basically exposes a blend of F-shaped and rectangular configurations. Various scaling down techniques were used in miniaturizing this antenna. A human phantom model has also been used for verifying its SAR (specific absorption rate) value in practical conditions 2. Antenna parameters were checked under various twisting conditions when mounted on the human body with the phantom model presented. The designed antenna provides all the parameters in an acceptable range as return loss is less than −10 dB. Gain value is greater than 3 dB, VSWR value is so close to 1 and bandwidth for operation is more than 300 MHz. 3. FR4 εr = 4.4 thickness = 1.6 mm 4. Resonate at 2.33–3.06 GHz(730 MHz), 5.2–7.4 GHz (2.2 GHz), Gain = 6.13 dB and 7.19 dB, return loss = −19.5 dB, −34.31 dB

(continued)

Table 1 (continued)

References	Antenna patch design, size (mm^2)/(mm^3) And number of operating bands	Remark
Jiayuan et al. [4]	1. Miniaturized antenna using corrugated microstrip line 2. 5.3 × 5.3 × 1.3 3. Single-band	1. It exhibits good slow wave–wave characteristics in a specific frequency range. The radiating part of the antenna is employed as CM (corrugated microstrip) for achieving low profile as well as good radiating characteristics. A detailed discussion of the impact of CM propagation constant on the antenna is done in the presented paper 2. It has been found on measurement that the proposed antenna achieves beamwidth of 75° in H-plane and 70° in E-plane with a gain tolerance value of 3 dB. Peak gain value at central frequency has been found as 5.15 dBi which suits the anticipated antenna design 3. Rogers RO4003C thickness = 1.5 mm εr = 3.55 4. Resonate at 9.0 GHz, gain = 5.15 dBi, return loss = −18.27 dB Beamwidth of 75° in H-plane and 70° in E-plane
Li et al. [5]	1. A palisade-shaped metasurface (PSMS) structure 2. 0.38 λ02 × 0.08 λ0 (λ0 = 5 GHz) 3. 0.38 λ02 × 0.08 λ0 (λ0 = 5 GHz) 4. Single-band	1. The antenna preference has been flexibly attuned by tuning to the width of slot gap s1 and the substrate height hs 2. The miniaturized PSMS antenna is not only having minimum size but also increased value of gain as well. This designed antenna is having great potential applications in many wireless communication systems 3. Rogers RO3003 (εr1 = 3, tan δ = 0.0013) and Glass Epoxy (εr2 = 4.4) 4. Bandwidth 19.6, Gain = 6.76 dBi resonate at 5.0 GHz

(continued)

Table 1 (continued)

References	Antenna patch design, size (mm^2)/(mm^3) And number of operating bands	Remark
Ain et al. [14]	1. Polydimethylsiloxane (PDMS) and Copper (Cu) patch with SMA connector × 60 × 3 3. Multi-band	1. The flexible antenna is made of copper patch and polydimethylsiloxane (PDMS) substrate. The presented antenna is having an SMA connector as the coaxial feed. Here the patch radius is 21.5 mm 2. The inclusion of PDMS + glass reduces the relative permittivity to 1.9 and loss tangent to 0.014 which is really helpful in enhancing antenna performance 3. The presented antenna was encapsulated under another PDMS/PDMS + glass substrate having 0.6 mm thickness so that it can overcome the adhesive issues between PDMS substrate and copper patch that also maintain the constant space from the ground as well 4. PDMS/PDMS + glass substrate. Thickness = 0.6 mm (εr = 1.9) 5. Resonate 1.92/2.34/2.46/2.25 GHz, Return loss = −16.22/−9.19/−11.3/−38.84 dB Bandwidth 18% 5. Resonate 1.92/2.34/2.46/2.25 GHz, Return loss = −16.22/−9.19/−11.3/−38.84 dB Bandwidth 18%
Soumen et al. [15]	1. Half-mode substrate-integrated waveguide structure 2. Single-band	1. The designed antenna supports a textile on body SIW antenna that can work under IOT application well. As in wearable technology, fabric material can easily be used as a substrate that is also sporting this SIW-based wearable antenna design requirement 2. The IOT devices are set apart from other communication network as here almost no bandwidth necessities are required 3. Jeans as substrate thickness = 1 mm, εr = 1.6 and tan δ = 0.021 4. Resonate at 5.87 GHz, Gain = 6.02 dBi, Efficiency = 99%, Return loss = −18 dB 5. Resonate at 5.87 GHz, Gain = 6.02 dBi, Efficiency = 99%, Return loss = −18 dB

(continued)

Table 1 (continued)

References	Antenna patch design, size (mm^2)/(mm^3) And number of operating bands	Remark
Dr. Gupta et al. [16]	1. UWB antenna in a thin substrate through slot insertion technique 2. 35 × 31 3. UWB	1. Various parameters like gain radiation pattern and reflection coefficient have been computed. The particular design has been designed so that it can operate over the range of 3.1–10.6 GHz 2. The main part of the design is slot insertion which helped in optimizing the radiation parameter. The radiation pattern of presented antenna was optimized with the introduction of slots in patch 3. Substrate used is RT-Duroid Relative permittivity (εr) = 2.2, Height of substrate = 0.127 mm 4. Operating frequency band = 3.1 GHz–10.23 GHz, Gain = 2.5 dBi

This resonator model has an L-shaped feed line which is effective in providing increased electrical length. The antenna is kept inside the human tissues which are conductive in nature, so the upper and lower substrate sides are in danger of getting short circuit. This situation is avoided with the use of biocompatible superstrates. Rogers 3210 with a thickness of 0.64 mm has been selected as superstrates and substrate materials. The loss tangent value for selected material is 0.003 and the relative permittivity is 10.2. The size of the antenna is 15 × 15 × 1.92 mm^3. The resonant frequency for the presented design is 403 MHz (range comes under MICS) and 2.45 GHz (widely used range in wireless communication) [10].

- Amiya et al. in the paper "*Design of a Miniaturized Circular Microstrip Patch Antenna for 5G Applications*", presented a new proximity-coupled fed circular microstrip patch antenna working under 5G applications. The resonant frequency of this antenna is 3.5 GHz. The length and width of the substrate of an antenna are 30 m and 45 mm, respectively. The radiating patch is having a radius of 13.005 mm. The proposed antenna has an excellent return loss S_{11} as −40.2827 dB and a substantial gain value of 5.8 dB. The VSWR value for this design is 1.02 (nearly ideal) and efficiency is 88.04% having the 200 MHz bandwidth. This proposed antenna is having allure features regarding %g applications such as machine-to-machine communication Internet of Things etc. The impedance of 50.99 – j0.15 Ω is achieved through the proposed antenna which is desired for achieving maximum power transfer too [11].
- Md. Siddat et al. in the paper *"Design of a Miniaturized Slotted T-Shaped Microstrip Patch Antenna to Detect and Localize Brain Tumor",* presented a miniaturized T-shaped wearable antenna designed for detection of brain tumor that operates in the frequency range of 902–928 MHz in ISM band. The specialty of the presented antenna is its small size, huge bandwidth and parametric results, which identify the unaffected and affected human brain tissue. The presented antenna is of the size 29.99 mm × 29.99 mm × 0.59 mm^3, and this antenna is placed on a head phantom model having six layers. Performance parameters have been investigated for tumor-affected as well as non-affected conditions. SAR value for the presented antenna is 0.332 W/kg which is under the acceptable range. The location of the tumor can be detected correctly by analyzing the data of the antenna [12].
- Tulsi Chowdhury et al. in the paper "*Design of a Patch Antenna Operating at ISM Band for Brain Tumor Detection",* presented a wearable antenna for detection of brain tumor operating in ISM band (2.4–2.4835 GHz). The shape of the antenna is pentagonal which is having a dimension of 33 × 23 × 1 mm^3. This antenna is positioned over a phantom of the human head for getting the values of different parameters of the antenna. The value of return loss we got for the two cases was different for normal head phantom and tumor-affected head phantom. The value of return loss in benign tumor-affected head phantom is −31.42 dB at 2.43 GHz as the resonant frequency. The existence of brain tumor can be understood through the shift of the resonant frequency to get maximum return loss value. The measured difference is 18 MHz for the tumor-affected head phantom and actual head phantom [2].

- Saraereh in the paper *Microstrip Wearable Dual-Band Antenna Design for ON Body Wireless Communications* presented a dual-band wearable antenna for ON body communications. These dual bands belong to the 2.45 GHz and 5.4 GHz bands. The prominent feature of the presented antenna is its size. In this the author basically exposes a blend of F-shaped and rectangular configurations. Various scaling down techniques were used in miniaturizing this antenna. A human phantom model has also been used for verifying its SAR (specific absorption rate) value in practical conditions. Antenna parameters were checked under various twisting conditions when mounted on the human body with the phantom model presented. The designed antenna provides all the parameters in an acceptable range as the return loss is less than −10 dB, gain value is greater than 3 dB, VSWR value is so close to 1 and bandwidth for operation is more than 300 MHz [13].
- Jiayuan Lu et al. in the paper *Design of Miniaturized Antenna Using Corrugated Microstrip* presented a novel way to design a miniaturized antenna with a corrugated microstrip line that exhibits good slow wave–wave characteristics in the specific frequency range. The radiating part of the antenna is employed as CM (corrugated microstrip) for achieving low profile as well as good radiating characteristics. A detailed discussion of the impact of CM propagation constant on the antenna is done in the presented paper. The verification of simulated results and feasibility of the designed antenna has been checked. It has been found on measurement that proposed antenna achieves beamwidth of 75° in H-plane and 70° in E-plane with gain tolerance value of 3 dB, peak gain value at central frequency has been found as 5.15 dBi which suits the anticipated antenna design. On resonant mode antenna also exhibits excellent radiation characteristics. The size of miniaturized antenna is $0.16\lambda_0 \times 0.16\lambda_0 \times 0.04\lambda_0$ at central frequency of 9 GHz. The presented antenna design has shown a new way of miniaturizing the antenna dimensions which is a promising sign for development in future wireless communication systems [4].
- Ximing Li et al. in the paper *Design and Characterization of a Miniaturized Antenna Based on Palisade Shaped Metasurface* presented a miniaturized single-fed monopole microstrip patch antenna with a palisade-shaped metasurface. For revealing the working principle parametric analysis along with dispersion analysis has been carried out. The dual resonant modes are generated here with this antenna design. The proposed antenna is having a compact structure with a maximum size of $0.38\lambda_0{}^2$; a low profile of $0.08\lambda_0$ (here λ_0 denotes the free-space wavelength at 5.0 GHz) which shows an impedance bandwidth of 19.6% (where $S_{11} < -10$ dB) and the 6.76 dBi is an average value of gain across the bandwidth. The antenna preference has been flexibly attuned by tuning to the width of slot gap s1 and the substrate height h_s. The miniaturized PSMS antenna not only has minimum size but also increased the value of gain as well. This designed antenna is having great potential applications in many wireless communication systems [5].
- Ain et al. presented the flexible antenna made of copper patch and polydimethylsiloxane (PDMS) substrate for medical applications. The presented antenna is having an SMA connector as the coaxial feed. Here the patch radius is 21.5 mm,

substrate volume is $60 \times 60 \times 3$ mm^3 and ground plane area is 60×60 mm^2. In the presented work the authors have also produced a PDMS + glass microsphere composite for substituting PDMS substrate. The inclusion of PDMS + glass reduces the relative permittivity to 1.9 and loss tangent to 0.014 which is really helpful in enhancing antenna performance. The presented antenna was encapsulated under another PDMS/PDMS + glass substrate having 0.6 mm thickness so that it can overcome the adhesive issues between PDMS substrate and copper patch that also maintain the constant space from the ground as well [14].

- Soumen et al. *in the paper SIW based body wearable antenna for IOT applications* presented a half-mode substrate-integrated waveguide working at IOT frequency band. The proposed antenna is having a central frequency of 5.87 GHz with a gain value of 6.02 dBi. The designed antenna supports a textile on body SIW antenna that can work under IOT application as well. As in wearable technology, fabric material can be easily used as a substrate that is also sporting this SIW-based wearable antenna design requirement. The IOT devices are set apart from other communication networks as here almost no bandwidth necessities are required. The measured results clearly exhibit the return loss value as −18 dB at a frequency of 5.8 GHz. The radiation pattern and other parameters for the antenna have been analyzed thoroughly as well [15].
- Dr. Gupta et al. presented the chronological development of UWB technology used in wearable antenna design from scratch to modern sophisticated application of the technology mentioned and also discussed the shortcomings of efforts or failure of the previous work and in their work they highlighted the mistakes which were happened due to which technology was not able to boom up at that time. It's really important to understand these facts before we go to work for the technology advancement and that's wonderfully explained in the paper represented by the author here. In brief, we can understand this like a broader range of operation is covered in this system in comparison to other wireless protocols, such as Wi-MAX, Bluetooth, Wi-Fi etc. High gain and larger bandwidth for impedance matching is the major requirement for such a system. The information of allocation of UWB from 3.1 to 10.6 GHz in 2002 has been clearly mentioned here in the provided work [16].

So with the above-discussed content, you can clearly understand. By presenting this review paper we would like you to go through the different technology and design methodology by which we could achieve miniaturization and flexibility [3, 10]. Through the paper presented by Chowdhury et al. and Ain et al., we wanted to let you know how these wearable antennas can be made useful in the medical field as well [2, 14]. In order to make this antenna more useful for different frequency bands, that is why through the paper presented by N. Gupta et al. where techniques have been discussed for making it useful in the whole range from 3.3 to 10.2 GHz (ultra-wideband) [16]. Through this whole review work, you could understand the different ways for making your wearable antenna handier as it will be compact, flexible, multiband or ultra-wideband and appropriate for use in the medical field.

5 Conclusion

In this review paper we have discussed the various design techniques, applications and the selection of material used for antenna substrate for the wearable antenna. We have also seen the techniques to make our antenna work for the required frequency band. We have also focused on using a textile antenna which is increasing the flexibility of wearable antenna. We discussed many challenges which come while designing the wearable antenna on the human body, like the size and shape of the antenna. We have seen that the flexibility achieved by the antenna with the use of flexible substrate helps in improving parameters like radiation pattern with more transmission and reception area and impedance matching and reduces the losses that occurred due to crumpling and bending. For overcoming these challenges we should select the appropriate textile material as substrate so that flexibility can be increased and the SAR value can be decreased, and at the same time, we also took care of other parameters like return loss antenna gain, radiation pattern etc.

References

1. C.K. Nanda, S. Ballav, A. Chatterjee, S.K. Parui, A body wearable antenna based on jeans substrate with wide-band response, in *5th International Conference on Signal Processing and Integrated Networks (SPIN)* (2018). 978-1-5386-3045-7/18
2. T. Chowdhury, R. Farhin, R.R. Hassan, M.S.A. Bhuiyan, R. Raihan, Design of a patch antenna operating at ISM band for brain tumor detection, in *Proceedings of the 2017 4th International Conference on Advances in Electrical Engineering* (IEEE, 2017). 978-1-5386-0869-2/17
3. R.K. Saraswat, M. Kumar, A quad band metamaterial miniaturized antenna for wireless applications with gain enhancement. Wirel. Personal Commun. (Springer, 2020). https://doi.org/10.1007/s11277-020-07548-z
4. J. Lu, H.C. Zhang, P.H. He, T.J. Cui, Design of miniaturized antenna using corrugated microstrip. IEEE Trans. Antennas Propag. (IEEE, 2019). https://doi.org/10.1109/TAP.2019.2963209
5. X. Li, J. Yang, Z. Chen, P. Ren, M. Huang, Design and characterization of a miniaturized antenna based on palisade-shaped metasurface, Hindawi Int. J. Antennas Propag. Article ID 7838563, 9 pages (2018). https://doi.org/10.1155/2018/7838563
6. C. Mao, P.L. Werner, D.H. Werner, D. Vital, S. Bhardwaj, Dual-polarized armband embroidered textile antenna for on-/off-body wearable applications (IEEE, 2019). 978-1-7281-0692-2/19
7. G.-P. Gao, *Member, IEEE*, Chen Yang, *Student Member, IEEE*, A wide bandwidth wearable all-textile PIFA with dual resonance modes for 5-GHz WLAN applications. IEEE Trans. Antennas Propag. (2018). https://doi.org/10.1109/TAP.2019.2905976
8. M. Klemm, I. Locher, G. Troster, A novel circularly polarized textile antenna for wearable applications, in *The Wireless Technology* (2004)
9. C.A. Balanis, Antenna theory: analysis and design, 3rd edn (Wiley, New York, 1997), p. 859
10. M. Palandoken, Compact bioimplantable MICS and ISM band antenna design for wireless biotelemetry applications. Radioengineering **26**(4) (2017)
11. A.B. Sahoo, N. Patnaik, A. Ravi, S. Behera, B.B. Mangaraj, Design of a miniaturized circular microstrip patch antenna for 5G applications, in *International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)* (IEEE, 2020)

12. M.S.B. Nesar, N. Chakma, M.A. Muktadir, A. Biswas, Design of a miniaturized slotted T-shaped microstrip patch antenna to detect and localize brain tumor, in *2nd International Conference on Innovations in Science, Engineering and Technology (ICISET),* 27–28 October 2018 (IEEE, 2018). 978-1-5386-8524-2/18/
13. O.A. Saraere, Microstrip wearable dual-band antenna design for ON body wireless communications, in *IEEE 4th International Conference on Computer and Communication Systems* (2019). 978-1-7281-1322-7/19
14. A.S. Za'aba, S.N. Ibrahim, N.F.A. Malek, A.M. Ramly, Development of wearable patch antenna for medical application, in *Regional Symposium on Micro and Nanoelectronics (RSM)* (IEEE, 2017)
15. S. Banerjee, A. Singh, S. Dey, S. Chattopadhyay, S. Mukherjee, S. Saha, SIW based body wearable antenna for IoT applications (IEEE 2019). 978-1-7281-0070-8/19
16. N.P. Gupta, M. Kumar, Radiation performance improvement in wearable UWB antenna through slot insertion technique, in *Fifth International Conference on Communication Systems and Network Technologies 2015*. IEEE. https://doi.org/10.1109/CSNT.2015.41
17. P. Salonen, L. Sydanheimo, M. Keskilammi, M. Kivikoski, A small planar inverted-F antenna for wearable applications, in *The Wearable Computers* (1999)

DEERS: Design Energy-Efficient Routing Scheme for Harsh Environment Monitoring in Heterogeneous WSNs

Samayveer Singh, Aruna Malik, Pawan Singh Mehra, and Pradeep Kumar Singh

Abstract The wireless sensor networks (WSNs) are hardcore used for continuous monitoring in harsh environments, such as floods detection, forest fire detection, volcano detection, etc. It can be reviewed from the existing schemes that the network instability and delay in data collection and transmission are leading. Thus, there is a requirement of increasing the network stability and decrease the delay in communication. In this paper, an energy-efficient routing scheme for heterogeneous WSNs is discussed. The proposed work considers three clustering parameters such as cluster heads (CH) distance to the sink, density of the cluster heads, and network residual energy for electing the cluster heads dynamically. A threshold-based probability formula that correlates all three parameters is also proposed. The heterogeneous nodes are considered for deploying in harsh environment monitoring. The performance of the proposed method is compared with the existing work by considering alive and dead nodes and throughput per round as metrics. In the simulation results, the proposed method performs better than the existing protocol on all the matrices.

Keywords Clustering · Energy efficiency · Harsh environment monitoring · Wireless sensor networks · Network lifespan

1 Introduction

Nowadays, wireless sensor networks (WSNs) are capable to handle multifold applications because of the advancement in the sensing technology of the sensor nodes. These networks play a significant role in gathering data for a long duration from hostile

S. Singh · A. Malik
Department of Computer Science and Engineering, Dr. B. R. Ambedkar National Institute of Technology Jalandhar, Jalandhar, Punjab, India

P. S. Mehra
Department of Computer Science and Engineering, Delhi Technological University, New Delhi, India

P. K. Singh (✉)
Department of Computer Science, KIET Group of Institutions, Delhi-NCR, Ghaziabad, UP, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_5

environments where human intervention is not possible. There are many applications in the current scenarios where we can deploy such types of technologies as real-time monitoring of vaccines, real-time fertilizer monitoring, border surveillance, real-time building structure health monitoring, precision agriculture, crops threshing, underwater monitoring system, etc. These wireless technologies are deployed for solving the above-mentioned problems and provide very efficient and effective solutions [1].

In WSNs, sensors nodes are deployed in the monitoring field for collecting the data or information according to the defined applications. The sensor nodes gathered the information from the field and forward the information to the sink. After collecting the data by the sink, it forwards the data to the server with the help of the internet connection or the end-user server. The deployment of nodes can be dynamic or static in WSNs. In the static deployment of nodes, sensor nodes are deployed by the human being manually. This type of deployment is not useful in the harsh environment where human intervention is not possible, whereas dynamic deployment of nodes can be done by the aircraft as per the requirement of the applications. This type of deployment is also called random deployment of sensor nodes [2–4].

Generally, sensor nodes have various capabilities like energy, link, memory, microprocessor, and other various computing capabilities. If all the nodes have the same capabilities they are deployed in the monitoring field and are called homogeneous nodes. If the nodes have different capabilities such networks are called heterogeneous networks. As discussed in many research papers [5, 6], heterogeneous networks are more capable than homogeneous networks. The cost of the heterogeneous networks is higher than the homogeneous networks because these heterogeneous networks consist of higher networks capabilities as compared to the homogeneous networks. The proportion of the cost of heterogeneous networks is not much higher than the performance proportion of the network. These networks perform many times better than the addition of the cost in capabilities of the networks [5, 6].

There are many issues in the WSNs, including routing, deployment, energy efficiency, fault tolerance, calibration, stability, flexibility, etc. These problems can be easily solved by incorporating the clustering process with the nodes in the monitoring area. In this process, the total number of deployed sensor nodes are divided into groups and for each group a master node is selected which collects the data from the other nodes. Initially, some of the nodes are chosen as the master node or cluster head (CH) [7]. These CHs cover most of the sensors by initializing the hello message and making communication with them. If some of the nodes do not have any CH, then some random nodes are chosen as CH. After that, each node in the field has CH with the condition that only one sensor node can communicate with only one CH. After selecting the CH for the nodes data collection will take place and after a certain amount of time rotation of the CH will be performed which helps in making the networks more efficient and balancing the load among the networks efficiently [8].

In this paper, an energy-efficient routing scheme for heterogeneous WSNs is discussed. The proposed work considers three clustering parameters such as cluster heads (CH) distance to the sink, density of the cluster heads, and network residual energy for electing the cluster heads dynamically. A threshold-based probability formula that correlates all three parameters is proposed. The heterogeneous nodes

are considered for deploying in harsh environment monitoring. The performance of the proposed method is compared with the existing work by considering alive and dead nodes, throughput, residual energy, and cluster heads per round as metrics.

The organization of the paper is given as follows: Sect. 2 discusses the literature review of the existing techniques. The systems model discusses in the Sect. 3 of the paper. Section 4 discusses the proposed method and Sect. 5 discusses the simulation results of the paper. Finally, the paper is concluded in Sect. 6 with its future scope.

2 Literature Review

A lot of research is going on the routing algorithms in WSNs for preserving the energy of the nodes. In this paper, we especially consider the advancement in the hostile environment where longevity of the network increases and also prolongs the network stability. There are many methods discussed for increasing the lifetime and stability period of the networks and are discussed as follows. One of the very first protocols which are discussed for WSNs is called the low-energy adaptive clustering hierarchy (LEACH) protocol [1]. In the LEACH, clusters are selected based on the probability formulation. If the probability value of the node is higher than the defined value of the threshold, the node is considered the CH; otherwise, the node may be CH in the other node. Lindsey et al. discuss a protocol that collects data in the chaining form [2]. This approach starts to collect data from the farthest nodes and forward that data collection to the nearer nodes and the same process is followed till the data reached the sink. However, this data collection process is not suitable for large networks. In [3], Chand et al. discuss a heterogeneous HEED protocol for WSNs. This approach considers three types of nodes for collecting the data and the fuzzy-based clustering method. However, this approach is suitable for large networks. In [4], Singh et al. discuss a multilevel heterogeneous network model for WSNs. This method considers an approach that consists of multiple types of heterogeneity of nodes. It can define a general model for heterogeneity. In [5], Singh et al. discuss a performance investigation of energy-efficient HetSEP for prolonging lifetime in WSNs. This method shows the comparative study of the stable election protocol with three levels of heterogeneity.

In [6], Singh et al. discuss an energy-aware data gathering and clustering technique based on nature-inspired optimization in WSNs. This paper discusses a clustering method by considering four different parameters, such as residual energy, density, average energy, and distance. In [7], Ramteke et al. discuss a particle swarm optimization (PSO) and genetic mutation-based (GM) routing technique for IoT-based homogeneous software-defined WSNs. This approach discusses PSO and GM for cluster head election. However, this method does not consider the heterogeneity in the networks. In [8], the authors discuss a distributed energy-efficient clustering algorithm for HWSNs. In this work, a two-level and multilevel heterogeneous model is discussed. It uses residual energy as a clustering parameter for electing the cluster heads. However, this method does not consider the density and distance for cluster

head election. The papers [9, 10] are the extension of the [8]. In [9, 10] the threshold formulation which helps in clustering is modified. The paper [11] discusses the multi-level stable and energy-efficient clustering protocol in HWSNs. A new heterogeneous model is discussed by considering three types of nodes in this work. However, this paper increases the significant stability period and lifetime of the networks.

In [12], Singh et al. discuss a method called OSEP, an optimized stable election protocol in heterogeneous WSNs. This method is the extension of the [5] by considering the three levels of heterogeneity. In [13], Malik et al. discuss a method called DACHE, a data aggregation-based effective and optimized cluster head election routing protocol for HWSNs. This method uses data aggregation with the clustering process and heterogeneity. However, this method performs very well as compared to the existing methods. In paper [14], Singh et al. discuss a clustering-based optimized stable election protocol in WSNs. This clustering process considers the residual energy, distance, and density parameters for cluster head election. However, it does not consider the average energy of the nodes for cluster head election. In paper [15], Singh et al. discuss an effective analysis and performance investigation of energy heterogeneity in WSNs. This paper uses a comparative study of the stable election protocol, deterministic energy-efficient clustering, and hybrid energy-efficient distribution.

In [16], Singh et al. discuss a method called optimized cluster head election protocol for heterogeneous WSNs (OCHEP). This method uses the optimized cluster method based on probability and its threshold value. In [17], Singh et al. discuss energy-efficient clustering protocol using fuzzy logic for heterogeneous WSNs. This method uses a fuzzy logic system by considering residual energy, distance, and density parameters for cluster head election. However, it did not define the general model for heterogeneity. In [18, 19], a stable period enhancement for zonal-based (SPEZ) clustering in heterogeneous WSN is discussed. This paper uses fuzzy-based clustering for electing the cluster heads efficiently. However, this paper does not consider the average energy of the nodes as clustering parameters. In [20], the authors discuss an energy-efficient cross-layer-based adaptive threshold routing protocol for WSN. This paper discusses an adaptive method for clustering and data collection. However, the results are not satisfactory in this paper and lifetime may be improved by adding more concepts. In [21], a method is discussed for selective α-coverage-based heuristic in wireless sensor networks. This paper discusses a work for target coverage problem. It does not consider the data collection process among the nodes and sinks.

In this paper, an energy-efficient routing scheme for heterogeneous WSNs is discussed. The proposed work considers three clustering parameters, such as cluster heads (CH) distance to the sink, density of the cluster heads, and network residual energy for electing the cluster heads dynamically. A threshold-based probability formula that correlates all three parameters is proposed. The heterogeneous nodes are considered for deploying in harsh environment monitoring. In this section, we will discuss the network energy model and radio energy dissipation model.

3 Assumptions of the Network Energy Model and Radio Dissipation Model

Following are the assumptions of the network energy model and radio dissipation model.

- Nodes are not movable and an ID is fixed for node identification.
- Initial energy of nodes is defined.
- Location of the sink is fixed and symmetric connections are used between the sensor nodes and sink for communication.
- Heterogeneous nodes are deployed in the network.

The proposed network consists of three types of nodes called nrm, int, and adv nodes. The energies and number of nrm, int, and adv nodes are denoted as e_1, e_2, ande_3 and $\mathrm{n_{nnrm}}$, $\mathrm{n_{nint}}$, $\mathrm{n_{nadv}}$, with the condition $\mathrm{n_{nnrm}} > \mathrm{n_{nint}} > \mathrm{n_{nadv}}$, respectively. The total energy of the network is calculated as follows:

$$e_{\mathrm{Tot}} = \oint \times n_n \times e_1 + \oint^2 \times n_n \times e_2 + \left(1 - \oint - \oint^2\right) \times n_n \times e_3 \tag{1}$$

where $\oint$ is the model parameter. The number of nrm, int, and adv nodes are $\oint \times n_n$, $\oint^2 \times n_n$, and $(n_n - (\oint * n_n + \oint^2 * n_n))$ with the e_1, e_2, and e_3 energy, respectively. When we are putting $\oint = 0$ in (1), it gives one type of nodes with the following energy of the networks, i.e., $\mathrm{E_{tot}} = n_n * \mathrm{E_3}$ in one-level heterogeneity. This is called the adv nodes instead of nrm nodes where a condition is imposed for changing the adv node energy into nrm nodes energy as

$$\oint = \frac{\mathrm{e_3} - \mathrm{e_1}}{\beta * f(\mathrm{e_2},\ \mathrm{e_3})} \tag{2}$$

When we are putting $1 - \oint - \oint^2 = 0$ in Eq. (1), it gives two types of sensors, namely nrm and int nodes. This relation $1 - \oint - \oint^2 = 0$ have two solutions, i.e., $(\left(\sqrt{5}\right) - 1)/2$ and $(\left(\sqrt{5}\right) + 1)/2$. The value $(\left(\sqrt{5}\right) - 1)/2$ lies in the range between 0 and 1. Thus, it may be considered the true value for solving the expressions in two levels of heterogeneity.

In three levels of heterogeneity, the upper bound is $(\left(\sqrt{5}\right) - 1)/2$ denoted by $\oint_{\mathrm{ub}}$ and let the lower bound is $\oint_{\mathrm{lb}}$. The range of $\oint$ is $\oint_{\mathrm{lb}} < \oint < \oint_{\mathrm{ub}}$ and function value f is considered as $(\mathrm{e_3} - \mathrm{e_2})$ from (2). Thus, $\oint_{\mathrm{lb}}$ is considered as $\oint_{\mathrm{lb}} < \oint < \oint_{\mathrm{ub}}$. By using the values of $\oint$ and$\oint_{\mathrm{ub}}$, calculate $\oint_{\mathrm{lb}}$ as below:

$$\oint_{\mathrm{L}} < \frac{\mathrm{e_3} - \mathrm{e_1}}{\beta * (\mathrm{e_3} - \mathrm{e_2})} < \left(\left(\sqrt{5}\right) - 1\right)/2 \tag{3}$$

Let $e_2 = \alpha_1 + e_1$ and $e_3 = \alpha_2 + e_2$ and we have the following relation: $\oint_{lb} < \frac{\alpha_2+\alpha_1}{\beta*\alpha_2}$

$$\frac{\alpha_2}{\alpha_1} < \frac{1}{\beta*\oint_{lb} - 1} = -\frac{\alpha_2}{\alpha_1} \geq \frac{1}{1 - \beta*\oint_{lb}} \tag{4}$$

By putting $-\frac{\alpha_2}{\alpha_1} = 0$ in (4) because LHS is negative we get the following relation:

$$1 - \beta*\oint_{lb} < 0 = \frac{1}{\beta} < \oint_{lb} \tag{5}$$

Putting the value in (3) as follows:

$$(e_3 - e_1) \leq \frac{\beta*\left(\left(\sqrt{5}\right) - 1\right)}{2}*(e_3 - e_2) \tag{6}$$

$$\beta*\left(\left(\sqrt{5}\right) - 1\right)*e_2 - 2*e_1 \leq \left(\beta*\left(\left(\sqrt{5}\right) - 1\right) - 2\right)*e_3 \tag{7}$$

The nrm, int, and adv nodes energy are given as e_1, $e_2 = e_1 * (1 + \omega)$, and $e_3 = e_1 * (1 + \eta)$, respectively. And the values of coefficients ω and η are 0.06 and 0.11, respectively.

The energy consumption in transmitting, receiving, and sensing L-bit data between the sensor nodes and BS is given as follows [1]:

$$E_{TXS} = L * \in_{elec} + L* \in_{fs} * d^2 \quad \text{if } d \leq d_0 \tag{8}$$

$$E_{TXL} = L * \in_{elec} + L* \in_{mp} * d^4 \quad \text{if } d > d_0 \tag{9}$$

$$E_{Rx} = L* \in_{elec} \tag{10}$$

$$E_{Sx} = L* \in_{elec} \tag{11}$$

where $\in_{elec}$, $\in_{fs}$, and $\in_{mp}$ are the energy consumption in different spaces and d_0 is threshold distance given as follows:

$$d_0 = \sqrt{\frac{\in_{fs}}{\in_{mp}}} \tag{12}$$

4 DEERS: Design Energy-Efficient Routing Scheme for Harsh Environment Monitoring in HWSNs

The proposed methodology of the energy-efficient routing scheme for harsh environment monitoring in HWSNs is divided into two phases, namely the setup phase and the steady phase. In the setup phase cluster heads are selected using the clustering process, whereas in the steady phase data collection will take place.

Setup Phase: In the setup phase, the whole monitoring area is divided into small clusters and cluster heads are selected using the defined process. In this work three clustering parameters such as cluster heads (CH) distance to sink, the density of the cluster heads, and network residual energy are considered for electing the cluster heads dynamically. This CH election method consumes very less energy as compared to the latest papers in data collection. The process of efficient clustering with data collection is divided into rounds. Initially, a predefined percentage of nodes is used as cluster heads (e.g., 5% nodes) by considering a constant as two CH cannot be equipped in the range of each. The selection of CH depends on the threshold value between 0 and 1 initially, but in the later rounds it is calculated as follows:

$$T(n_i) = \frac{p_i}{1 - p_i\left[rmod\left(\frac{1}{p_i}\right)\right]} \times \frac{\mathrm{Dist_{avg}}}{\mathrm{Dist_{BS_to_n}}} \times \mathrm{Dest_n} \times \frac{\mathrm{E_{resi}}}{\mathrm{E_{tot}}} \tag{13}$$

where $T(n_i)$ indicates the threshold value for a particular type of node. p_i and r indicate the probability of the particular node and the round. $\mathrm{Dist_{avg}}$, $\mathrm{Dist_{BS_to_n}}$, $\mathrm{E_{resi}}$, $\mathrm{Dest_n}$, and $\mathrm{E_{tot}}$ indicate the average distance from the CH, distance between the CH and nodes, residual energy of the network, node density, and total energy of the networks, respectively.

The total energy of the networks (e_{tot}) is given as follows:

$$\begin{aligned} e_{\mathrm{tot}} &= \oint \times n_n \times e_1 + \oint^2 \times n_n \times e_2 + \left(1 - \oint - \oint^2\right) \times n_n \times e_3 \\ &= n_n \times \left(\oint \times e_1 + \oint^2 \times e_2 + \left(1 - \oint - \oint^2\right) \times e_3\right) \end{aligned} \tag{14}$$

$$e_{\mathrm{tot}} = e_1 \times n_n \times \left(\oint + \oint^2 \times e_2/e_1 + \left(1 - \oint - \oint^2\right) \times e_3/e_1\right) \tag{15}$$

e_1 is defined as the initial energy of a node and $e_1 \times n_n$ is the total network energy. $(\oint + \oint^2 \times e_2/e_1 + (1 - \oint - \oint^2) \times e_3/e_1)$ indicates the incremental factor due to their energy heterogeneity, which means heterogeneous network having $(\oint + \oint^2 \times e_2/e_1 + (1 - \oint - \oint^2) \times e_3/e_1)$ times more energy than that of the homogeneous nodes. The nodes become CH after $1/\mathrm{p}_i \times (\oint + \oint^2 \times e_2/e_1 + (1 - \oint - \oint^2) \times e_3/e_1)$ rounds and the average CH in the heterogeneous networks are as follows:

$$n_n \times 1/\mathrm{p}_i \times \left(\oint + \oint^2 \times e_2/e_1 + \left(1 - \oint - \oint^2\right) \times e_3/e_1 \right).$$

The weighted probabilities of the proposed protocol DEERS nrm, int, and adv nodes are p_{nrm}, p_{int}, and p_{adv} given as follows:

$$\begin{aligned} p_{nrm} &= \frac{\mathrm{p_i} \times \mathrm{E_1}}{\left(\oint \times e_1 + \oint^2 \times e_2 + \left(1 - \oint - \oint^2\right) \times e_3 \right)} \\ &= \frac{\mathrm{p_i}}{\left(\oint + \oint^2 \times e_2/e_1 + \left(1 - \oint - \oint^2\right) \times e_3/e_1 \right)} \end{aligned} \tag{16}$$

$$\begin{aligned} p_{int} &= \frac{\mathrm{p_i} \times \mathrm{E_1}}{\left(\oint \times e_1 + \oint^2 \times e_2 + \left(1 - \oint - \oint^2\right) \times e_3 \right)} \\ &= \frac{\mathrm{p_i}}{\left(\oint + \oint^2 \times e_2/e_1 + \left(1 - \oint - \oint^2\right) \times e_3/e_1 \right)} \end{aligned} \tag{17}$$

$$\begin{aligned} p_{adv} &= \frac{\mathrm{p_i} \times \mathrm{E_1}}{\left(\oint \times e_1 + \oint^2 \times e_2 + \left(1 - \oint - \oint^2\right) \times e_3 \right)} \\ &= \frac{\mathrm{p_i}}{\left(\oint + \oint^2 \times e_2/e_1 + \left(1 - \oint - \oint^2\right) \times e_3/e_1 \right)} \end{aligned} \tag{18}$$

The threshold value of the nrm, int, and adv nodes are given as follows:

$$T(nrm_i) = \frac{p_{nrm_i}}{1 - p_{nrm_i}\left[rmod\left(\frac{1}{p_{nrm_i}}\right)\right]} \times \frac{\mathrm{Dist_{avg}}}{\mathrm{Dist_{BS_to_n}}} \times \mathrm{Dest_n} \times \frac{\mathrm{E_{resi}}}{\mathrm{E_{tot}}} \tag{19}$$

$$T(int_i) = \frac{p_{int_i}}{1 - p_{int_i}\left[rmod\left(\frac{1}{p_{int_i}}\right)\right]} \times \frac{\mathrm{Dist_{avg}}}{\mathrm{Dist_{BS_to_n}}} \times \mathrm{Dest_n} \times \frac{\mathrm{E_{resi}}}{\mathrm{E_{tot}}} \tag{20}$$

$$T(adv_i) = \frac{p_{adv_i}}{1 - p_{adv_i}\left[rmod\left(\frac{1}{p_{adv_i}}\right)\right]} \times \frac{\mathrm{Dist_{avg}}}{\mathrm{Dist_{BS_to_n}}} \times \mathrm{Dest_n} \times \frac{\mathrm{E_{resi}}}{\mathrm{E_{tot}}} \tag{21}$$

The $T(n_{nrm})$, $T(n_{int})$, and $T(n_{adv})$ indicate the threshold functional to nrm, int, and adv nodes, respectively. By using this procedure, cluster-efficient cluster nodes are selected for collecting the data.

Steady Phase: After setup, the clusters, and their respective cluster nodes the data collection process starts. In the data collection process, sensor nodes of all the clusters gathered data from the environment and send that information to the cluster heads. After getting the information from the sensor nodes, cluster nodes forward the collected information to the sink or BS using multi-hop communication. This

process continues for a certain amount of iterations called rounds. After a round setup phase again starts and follows the complete clustering process as discussed in the setup phase and then the data collection process follows the steady phase. This process continues till the dead of all the nodes in the networks.

5 Experimental Results and Discussions

In this paper, the existing protocols, namely DEEC [8], DDEEC [9], EDEEC [10], SEEC [11] protocols, and the proposed method are simulated by considering the different metrics such as the number of alive nodes, dead nodes, and packets sent to BS. The number of alive nodes refers to the time whenever a single node is alive in the network. The number of dead nodes defines as the time when all the sensor nodes are dead. The packet sent to BS refers to the number of packets sent to the base station per unit time. It is also called the threshold of the networks. This work discusses an efficient routing method for heterogeneous WSNs. In this work three clustering parameters such as cluster heads (CH) distance to sink, density of the cluster heads, and network residuals energy are considered for electing the cluster heads dynamically. The DEEC [8], DDEEC [9], EDEEC [10], SEEC [11] protocols and the proposed method are implemented in MATLAB with the 11th Gen Core (TM)i7 operating system and 16 GB RAM machine. In this network, 100 sensor nodes are deployed in the area of 100 m $\times$ 100 m. Three types of nodes are considered as nrm, int, and adv with 50, 30, and 20 in numbers and with their respective energy 0.5 J, 1 J, and 1.5 J, respectively. The message length is considered as 4000 bits, the threshold distance value is considered as 75 m, and 25 m is considered as cluster range. The depletion of the energy is 50 nJ/bits in management of the circuit, 50nJ/bit, 10pJ/bit/m^2 and 0.0013pJ/bit/m^4 to transmit the signal in shorter and long distances, respectively.

Figure 1 shows the simulation results in terms of number of rounds with respect to the number of alive nodes for the DEEC [8], DDEEC [9], EDEEC [10], SEEC [11] protocols and the proposed method. The proposed method is alive more number of rounds as compared to the DEEC [8], DDEEC [9], EDEEC [10], SEEC [11] protocols.

Also, the stability period of the proposed work is much higher than the existing protocols like DEEC [8], DDEEC [9], EDEEC [10], SEEC [11] because of the first node death after a long time in the network. The first node depicted their energy in the case of the proposed method and the SEEC [11], EDEEC [10], DDEEC [9], and DEEC [8] protocols are 2797, 1932, 1489, 1367, and 1134 rounds, respectively. It is evident from the result that the increment in the stability period is increased by 44.77% as compared to the SEEC [11]. The reason behind this gigantic improvement is not only the distance and residual energy parameters but also the node density which helps in selecting the effective cluster heads. This effective clustering reduces the distance among the cluster heads and the sink for collecting the data in inter and intra-cluster communication.

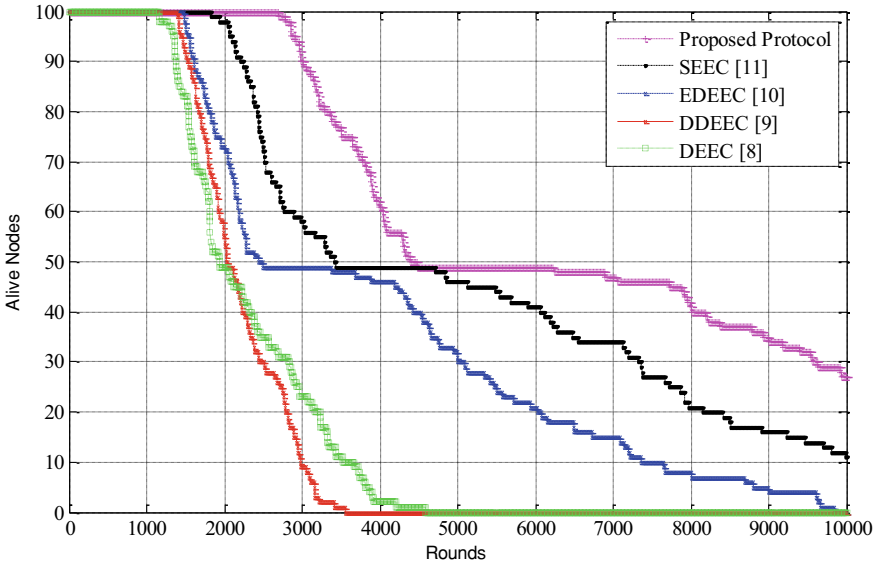


Fig. 1 Number of rounds versus alive nodes

Figure 2 shows the simulation results in terms of number of rounds with respect to the number of dead nodes for the DEEC [8], DDEEC [9], EDEEC [10], SEEC [11]

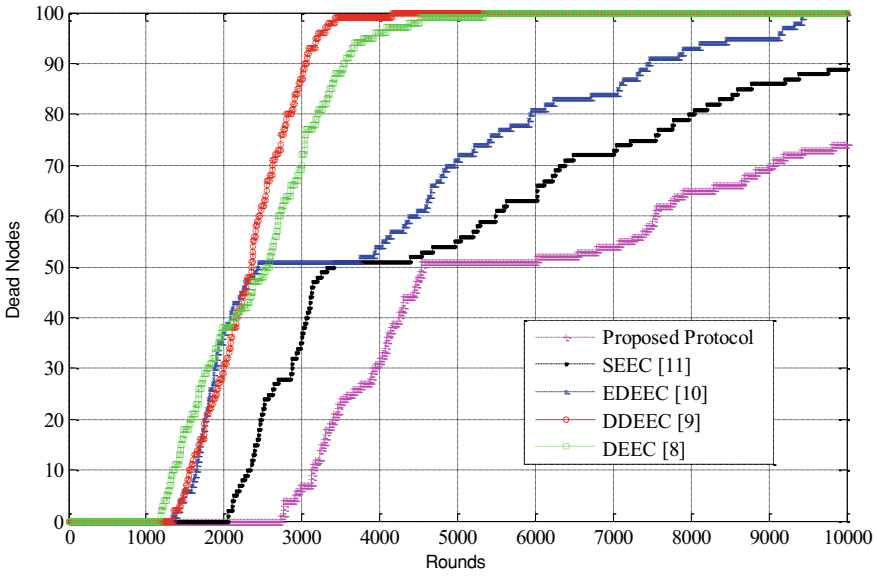


Fig. 2 Number of rounds versus dead nodes

protocols and the proposed method. The first node depicted dead after 2797, 1932, 1489, 1367, and 1134 rounds, half nodes dead after 5903, 4587, 3675, 2332, 2284 rounds, and 12,459, 10,983, 9043, 4284, and 3582 rounds for both the proposed and SEEC [11], EDEEC [10], DDEEC [9], and DEEC [8] protocols. This improvement in the results is because of the energy-efficient CH selection which reduces energy expenditure in the network. In addition to efficient CH election, the proposed also considered the average distance among the nodes and sink for transmitting the data. The average distance reduces the intra-cluster distance and inter-cluster distance which further reduces the communication cost of nodes. This improves the overall network lifetime due to its energy preservation.

Figure 3 shows the simulation results in terms of number of rounds with respect to the number of packets sent to BS for the DEEC [8], DDEEC [9], EDEEC [10], SEEC [11] protocols and the proposed method. The proposed method transmits more number of packets to the base station as compared to the existing methods such as DEEC [8], DDEEC [9], EDEEC [10], and SEEC [11]. The improvement is because of the reduction in the intra-cluster distance and inter-cluster distance which further decreases the communication cost of nodes. Because of the network longevity, more number of packets is transmitted to the BS. Single-hop communication also reduced energy consumption and increased the network lifetime.

We also analyze the simulation results in terms of number of rounds with respect to the number of cluster heads for the DEEC [8], DDEEC [9], EDEEC [10], SEEC [11] protocols and the proposed method. The cluster range for the DEEC [8], DDEEC [9], EDEEC [10], SEEC [11] protocols and the proposed method is varying between

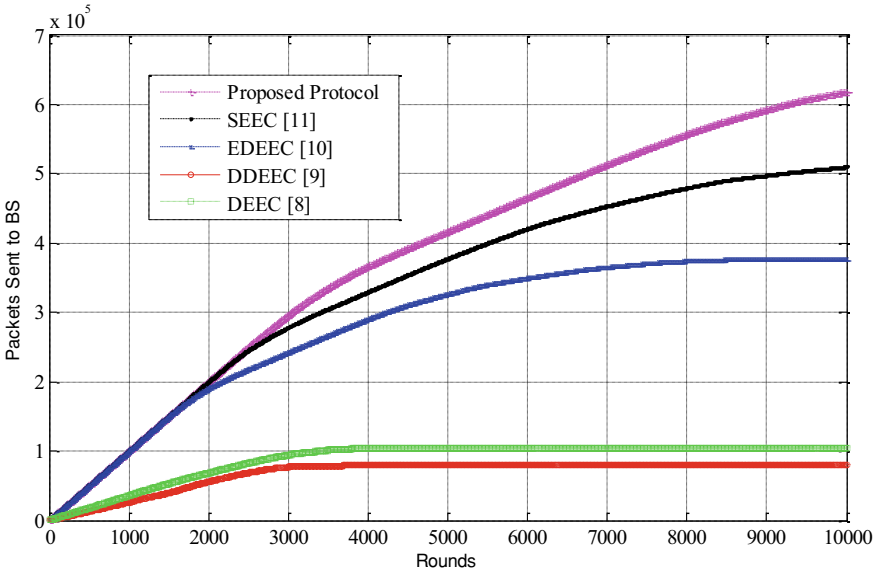


Fig. 3 Number of rounds versus packets sent to BS

0 and 10. The average number of cluster heads in the DEEC [8], DDEEC [9], EDEEC [10], SEEC [11] protocols is 8 or 9, whereas the average number of cluster heads in the proposed protocol is 5 or 6. In most of the rounds, the number of cluster heads is varying by 1. The proposed method works long as compared to the DEEC [8], DDEEC [9], EDEEC [10], and SEEC [11] protocols. Thus, approximately 10% nodes of the total number of nodes are more than sufficient for the cluster heads. If we have a very large number of cluster heads, then the lifetime of the networks will not be well as discussed in the other work. Therefore, the cluster heads should be defined in the well-structured range.

6 Conclusion

In this paper, an energy-efficient routing scheme for heterogeneous WSNs is discussed. The proposed work considers three clustering parameters such as cluster heads (CH) distance to the sink, density of the cluster heads, and network residual energy for electing the cluster heads dynamically. A threshold-based probability formula that correlates all three parameters is proposed by considering heterogeneous nodes for deploying in harsh environment monitoring. The first node depicted their energy in case of the proposed method and the SEEC [11], EDEEC [10], DDEEC [9], and DEEC [8], protocols is 2797, 1932, 1489, 1367, and 1134 rounds, respectively. It is evident from the result that the increment in the stability period is increased by 44.77% as compared to the SEEC [11]. The reason behind this gigantic improvement is not only the distance and residual energy parameters but also the node density which helps in selecting the effective cluster heads. This effective clustering reduces the distance among the cluster heads and the sink for collecting the data in inter- and intra-cluster communication.

References

1. W.B. Heinzelman, A.P. Chandrakasan, H. Balakrishnan, An application-specific protocol architecture for wireless microsensor networks. IEEE Trans. Wirel. Commun. **1**(4), 660–670 (2002)
2. S. Lindsey, C.S. Raghavendra, K.M. Sivalingam, Data gathering algorithms in sensor networks using energy metrics. IEEE Trans. Parallel Distrib. Syst. **13**(9), 924–935 (2002)
3. S. Chand, S. Singh, B. Kumar, Heterogeneous HEED protocol for wireless sensor networks. Wirel. Pers. Commun. **77**, 2117–2139 (2014)
4. S. Singh, S. Chand, B. Kumar, Multilevel heterogeneous network model for wireless sensor networks. Telecommun. Syst. **64**, 259–277 (2017)
5. S. Singh, P.K. Singh, Performance investigation of energy efficient HetSEP for prolonging lifetime in WSNs, in: *Futuristic Trends in Network and Communication Technologies. FTNCT 2018* , ed. by P. Singh, M. Paprzycki, B. Bhargava, J. Chhabra, N. Kaushal, Y. Kumar. *Communications in Computer and Information Science*, vol. 958 (Springer, Singapore, 2019)

6. S. Singh, An energy aware clustering and data gathering technique based on nature inspired optimization in WSNs. Peer-to-Peer Netw. Appl. **13**, 1357–1374 (2020)
7. R. Ramteke, S. Singh, Particle swarm optimization and genetic mutation based routing technique for IoT-based homogeneous software-defined WSNs, in *Recent Innovations in Computing, ICRIC 2020*, ed. by P.K. Singh, Y. Singh, M.H. Kolekar, A.K. Kar, J.K.Chhabra, A. Sen. Lecture Notes in Electrical Engineering, vol. 701 (Springer, Singapore, 2021)
8. L. Qing, Q. Zhu, M. Wang, Design of a distributed energy-efficient clustering algorithm for heterogeneous wireless sensor network. Comput. Commun. **29**, 2230–2237 (2006)
9. B. Elbhiri, R. Saadane, S. El Fkihi, D. Aboutajdine, Developed distributed energy-efficient clustering (DDEEC) for heterogeneous wireless sensor networks, in *5th International Symposium on I/V Communications and Mobile Network (ISVC)* (2010)
10. P. Saini, A.K. Sharma: E-DEEC—enhanced distributed energy efficient clustering scheme for heterogeneous WSN, in *2010 1st International Conference on Parallel, Distributed and Grid Computing (PDGC—2010)*
11. F.W. Zaki, R. Rizk, F. Farouk, Multi-level stable and energy-efficient clustering protocol in heterogeneous wireless sensor networks. IET Wirel. Sens. Syst. **4**, 159–169 (2014)
12. S. Sing, P.K. Singh, A. Malik, OSEP: an optimized stable election protocol in heterogeneous wireless sensor networks, in *Recent Innovations in Computing. ICRIC 2020,* ed. by P.K. Singh, Y. Singh, M.H. Kolekar, A.K. Kar, J.K. Chhabra, A. Sen. *Lecture Notes in Electrical Engineering,* vol. 701 (Springer, Singapore, 2021)
13. A. Malik, S. Singh, P.K. Singh, DACHE: A data aggregation-based effective and optimized cluster head election routing protocol for HWSNs, in *Recent Innovations in Computing. ICRIC 2020,* ed. by P.K. Singh, Y. Singh, M.H. Kolekar, A.K. Kar, J.K. Chhabra, A. Sen. Lecture Notes in Electrical Engineering, vol. 701 (Springer, Singapore, 2021)
14. S. Singh, A clustering-based optimized stable election protocol in wireless sensor networks, in *Applications in Ubiquitous Computing. EAI/Springer Innovations in Communication and Computing,* ed. by R. Kumar, S. Paiva (Springer, Cham, 2021)
15. S. Singh, R. Kumar, P.K. Singh, An effective analysis and performance investigation of energy heterogeneity in wireless sensor networks, in *Handbook of Wireless Sensor Networks: Issues and Challenges in Current Scenario's,* ed. by P. Singh, B. Bhargava, M. Paprzycki, N. Kaushal, W.C. Hong. Advances in Intelligent Systems and Computing, vol. 1132 (Springer, Cham, 2020)
16. S. Singh, P. Yadav, A. Malik, R. Agrawal, OCHEP: an optimized cluster head election protocol for heterogeneous WSNs, in *Advances in Smart Communication and Imaging Systems* ed. by R. Agrawal, C. Kishore Singh, A. Goyal. Lecture Notes in Electrical Engineering, vol. 721 (Springer, Singapore, 2021).
17. S. Singh, S. Chand, B. Kumar, Energy efficient clustering protocol using fuzzy logic for heterogeneous WSNs. Wirel. Pers. Commun. **86**, 451–475 (2016)
18. P.S. Mehra, M.N. Doja, B. Alam, Stable period enhancement for zonal (SPEZ)-based clustering in heterogeneous WSN. Smart Innov. Syst. Technol. **79**, 887–896 (2018)
19. P.S. Mehra, M.N. Doja, B. Alam, Stable period extension for heterogeneous model in wireless sensor network. Adv. Intell. Syst. Comput. **638**, 479–487 (2018)
20. R. Singh, A.K. Verma, Energy efficient cross layer based adaptive threshold routing protocol for WSN. AEU Int. J. Electron. Commun. **72**, 166–173 (2017). https://doi.org/10.1016/J.AEUE.2016.12.001
21. S. Manju Chand, B. Kumar, Selective α-coverage based heuristic in wireless sensor networks, wireless personal communication. Springer, **97**, 1623–1636 (2017)

Context-Enriched Machine Learning-Based Approach for Sentiment Analysis

Hamza Abubakar Kheruwala, Mohammed S. Ahmad, Jai Prakash Verma, Sudeep Tanwar, and Pradeep Kumar Singh

Abstract The advent of technology has given rise to a tremendous rise in people sharing information and voicing their opinions over different platforms. Several research studies have been focused on the analysis and prediction of political events through the "*sentiment*" perspective. Here, we present a comparative analysis of different approaches used by various researchers in this area. Also, an experiment-based analysis is presented with sentiment analysis for the situation of reservations in India with the consideration of public opinion on tweets, news reviews, and word of influential leaders with a huge number of public following to ensure that their opinions had a considerable impact with a bigger audience. We use a novel analysis model by employing *SVM with grid search* for sentiment analysis. The detailed discussion showcased is an intensive study of different methodologies adopted, which is of use to researchers looking to pursue this domain.

Keywords Text summarization · Text mining · Natural language processing · Machine learning · Big data

1 Introduction

Given the tremendous rise of social media since the inception of the twenty-first century, there has been a similar increase in the data available over various platforms [1]. People have shifted to the digital media to voice their opinions over countless

H. A. Kheruwala · J. P. Verma (✉) · S. Tanwar
Institute of Technology, Nirma University, Ahmedabad, India
e-mail: jaiprakash.verma@nirmauni.ac.in

S. Tanwar
e-mail: sudeep.tanwar@nirmauni.ac.in

M. S. Ahmad
Jamia Millia Islamia University, New Delhi, India

P. K. Singh
KIET Group of Instituions, Ghaziabad, Uttar Pradesh, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_6

numbers of ongoing issues and topics, flooding the Internet with reviews and user-generated content [2]. Various new platforms, blogs, and articles have paved the way for the average user to share their opinions and say on desired issues [3]. Likewise, the dynamic shift in technology has paved the way for techniques to mine this data and use the same for several useful practical applications [4]. This brings us to *sentiment analysis*, the interpretation, and mining of public opinions within textual data using several analysis techniques. The identification of public opinion toward political issues is essential to shape various decisions and reforms of the country. It can be achieved through opinion polls that have been in use for a long time, primarily due to unclear and biased questions that weaken the accuracy. Also, these apply only at the time these samples were taken. The need for a system that is dynamic, non-biased, and fine-grained is inherent to capture public opinions.

This paper is broadly divided into two main categories: The first category mainly discusses and analyzes several approaches adopted in this domain over the last few years. We discuss the methodologies adopted through various techniques (machine learning, Lexicon), giving readers the entire gist of sentiment analysis. Next, we discuss and propose our approach adopted here, for the situation of *Reservations in India* over the last 4–5 years [5].

The paper is divided into five sections. Section 1 gives the Introduction to this article (the need, findings, and the proposed system), Sect. 2 talks about the concept of sentiment analysis, and Sect. 3 discusses the related work on this domain of sentiment analysis (different approaches, findings, and analysis). Section 4 presents our model and the methodology adopted in our approach, Sect. 5 goes about our implementation strategy for the prescribed model, while Sect. 6 comes out with the results obtained and the comparison with similar model(s), and the novelty of our approach here.

2 Motivation

The most widely used text mining strategy, sentiment analysis, has widely been in use for the past few years. It is series of tools, methods, and techniques aimed at the identification, categorization, and calculation of sentiments in the form of subjective information over data expressed in mediums of text using computational analysis [6, 7]; in short words, identify the polarity of given data [8]. Both government and public/private organizations use sentiment analysis as a means of improving their products and services. Mainly, sentiment analysis can be performed on the document level, sentence level, and the aspect level of granularity [9]. The application of these levels adheres to the sentiment analysis of the document, sentence, and individual entity, respectively. Out of the many approaches used in sentiment analysis, the machine learning approach, Lexicon-based approach, and hybrid approach are the main categories. The Lexicon-based approach consists of a pre-defined polarity-based dictionary, where words are assigned scores based on their polarity; the machine learning approach [10] uses classifiers to train on datasets beforehand,

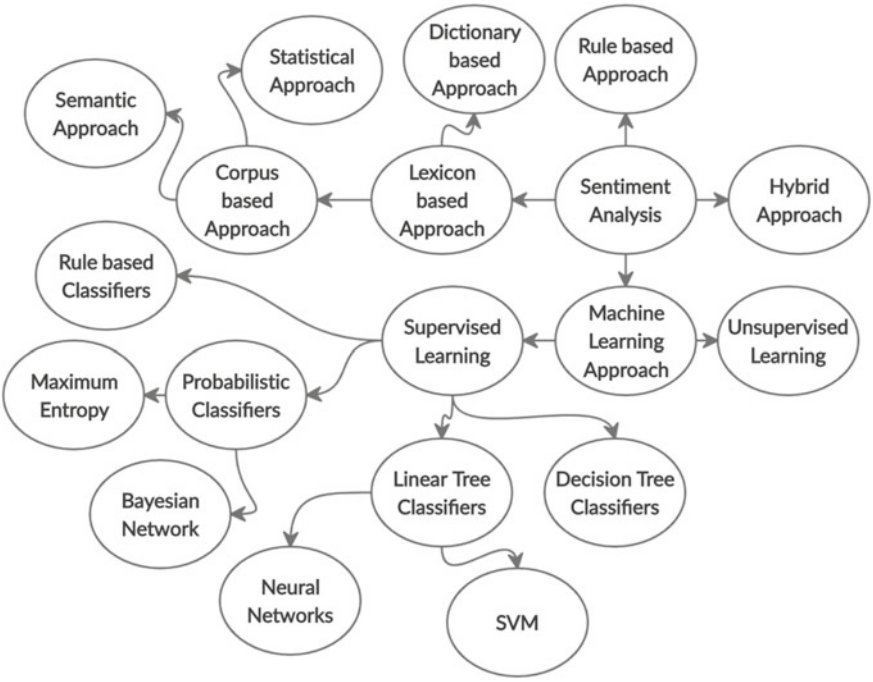


Fig. 1 Sentiment classification techniques

and then determine the document polarity. A combination of these two, the hybrid approach requires the models to be trained on the results of the Lexicon-based results [11]. Various techniques on the classification of sentiments, as explained by Medhat et al. [12], are given in Fig. 1.

From the various sources available for the extraction of data (Facebook, Instagram, Twitter), Twitter is widely preferred due to a limit in the number of characters allowed and its global use for the sharing of opinions on issues worldwide [13, 14]. The data from this social media platform is taken as the dataset for sentiment analysis covering different kinds of applications.

The rise of technology clubbed with the increased role of people has influenced major decisions taken by a nation. While this has been the case throughout generally, the extent of the approach varies in different countries. The nation where the opinions of the two go hand in hand proves to give away much better decisions at the biggest level. The invasion of Iraq in 2003 attracted the interests of the US and UK, who looked to disarm the country of "weapons of mass destruction". Following this, one of the biggest public demonstrations was held in the UK with the collective view of opposing the invasion. While the same was unable to influence the United Kingdom, it played an important role in influencing Canada's decision of abstaining from participation.

Comito et al. [15] used data from Twitter for the surveillance of Influenza in Italy for a period of two seasons; the results from this analysis were found to be significantly correlated to reports published on the InfluNet system. They further look to expand their reach of social media networking sites to obtain opinions from an even greater dataset. Alfarrarjeh et al. [16] combined datasets from Twitter and Flickr to perform the geospatial multimedia sentiment analysis on disaster-related content. This helped in the semantic abstraction of the situation then and also in making informed decisions during the crisis. In the future, they are applying the geo-sentiment framework in other applications and widening their reach of social media. Sentiment analysis can also be applied to the prediction and study of elections using related tweets; Cury [17] analyzed the oscillation of tweet sentiments concerning João Doria Jr. for election as Mayor. Tweets related to the subject were captured for periods before, during, and after the elections for a total of 12 months. A similar study by Chellia et al. [18] took to predicting the results of elections in India in the states of Haryana and Maharashtra. The results obtained were not quite promising, and the team looked to identify users based on age, gender, cast, etc. Dwivedi and Pant [19] focused on the sentiment analysis on big data based on a real-time update in reviews available online. Kušen et al. [20] focused on finding the shift in the emotional valence over the Internet over the spread of information. Son et al. [21] targeted the sarcasm behind textual data using bidirectional LSTM with convolutional network.

Targeting a different source of data, Yadav et al. [22] performed sentiment analysis on news, since the news was found to have an impact on the behavior of the financial trading industry. This approach was limited by the chaotic market behavior, though the results were found out to be at par with finding from prior studies. Fusing results from datasets obtained from various domains can lead to an even better semantic understanding of the situation at hand [23]. Based on recent studies, we find that the polarity of people's opinion is affected by the leaders they are following [24], and the positions they hold in the current information exchange [3, 25]. Rahab et al. [26] performed a sentimental analysis on tweets related to the Palestinian–Israeli conflict, where they found countries interested in this situation and calculated their sentiments during the conflict and the time after. The novel approach they adopted was the analysis of sentiments, both from Twitter users and from opinion leaders who influenced people. This helped in an even better understanding of the situation, as data was extracted from multiple sources [27].

Delving deeper into this strategy (Fig. 2), we propose a model where we target tweets from users, opinion leaders, and well data available from news agencies. Hence, we come to know the similarity/difference of opinion regarding selected topics between the general public and media houses. While many others focused on obtaining sentiments from one side of the situation, we target both sides, too (I) gain and complete understanding of the matter and (II) compare the difference of opinions in the form of sentiments.

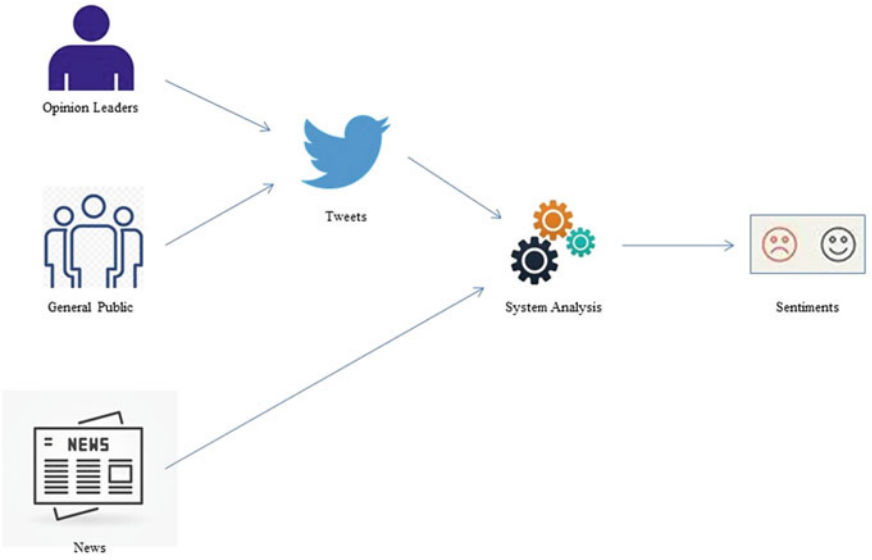


Fig. 2 General flow of the proposed system

3 Related Work

In this section, we are discussing the various approaches that have been applied by different researchers in this direction, followed by a novel approach used in our model.

Dragoni and Petrucci [28] propose a neural network-based approach to word embedding (NeuroSent tool), generated from an opinion-based corpus, to target individual words from raw sentences. They design a deep learning technique to exploit the generated word embedding to train the model and use multiple output layers to combine the scores of individual domains. Li et al. [29] propose the use of Recurrent Neural Networks (RNN) for the analysis of sentiments in their model. They combine the influence of different users on social media with a time series to filter the data obtained in these slots. This gives them a better efficiency than other models, with the incorporation of both time and user-related features.

To make the approach more favorable, Xia et al. [30] went a step ahead of sentiment analysis on isolated datasets and proposed a method toward using sentiment classifiers for lifelong learning of datasets. The knowledge from previous tasks could easily be incorporated into the prediction of newer analysis, and the general framework proposed in this model is suitable for any algorithm to perform lifelong sentiment learning. With a related ideology. Shunmuga Sundari and Shenbagavalli [31] performed temporal analysis to identify the change of people's sentiments with time. As part of this method, Lexicon-based BI-LSTM was used, which makes use of the widely used Lexicon and its features for the final representation of features.

While existing word embedding models failed to address the problems of sentiment embeddings, Yu et al. [32] proposed a strategy where they refine the existing trained words using sentiment analysis scores. This helped in getting vectors for words that are close enough in meaning both semantically and sentimentally, thereby giving a fine-grained analysis for each word in the dataset. Through experimental results, it is shown that the refined model proves to be better in places of the conventional word embeddings approach.

Nikitha Kumari et al. [33] target the sentiments of the Indian public behind the "MeToo" movement, using NLTK and TextBlob packages from Python to perform sentiment analysis, and the classification is done using the Naïve Bayes classifiers, thereby giving an accuracy of 86%. Sailunaz and Alhajj [34] also used the Naïve Bayes algorithm to classify the sentiments with a threefold, fivefold, and tenfold cross-validation on the text. Gupta et al. [35] used machine learning techniques to compute risks and compare different algorithms used for sentiment analysis and based on various attributes like precision, recall, F1-score, and accuracy. Data based on Naxal Activity in Maharashtra, India was collected from Twitter, and sentiment analysis was done using algorithms like Support Vector Machines (SVM), Naïve Bayes, K-nearest neighbor, and decision tree. SVM proved to give better results than the rest in terms of the given parameters.

Bhuvaneswari and Ravichandran [36] used the technique of assigning scores as in the conventional approach, in addition to the usage of adjectives and adverbs in the sentiment analysis. They showed how the use of these can affect the polarity of a sentence, and including these can help in an even better understanding of the sentiments behind datasets. Ray and Chakrabarti [37] propose a novel approach of combining the rule-based method's combination with deep learning to identify both the implicit and explicit aspects behind the given data. This gives a better understanding of the sentiments since many traditional methods get to identify only the words that are explicitly stated in the text. There is an increasing engagement of youth by the days on social media, who look forward to the use of emojis, emoticons, and slang to express their views and opinions. Part of the data pre-processing, the inclusion of these as features in sentiment analysis methods would help provide a better understanding since emoticons vaguely give the sentiment behind opinions expressed by users.

As stated by Senthil Kumar et al., the work of semantic analysis is challenged by dependency on the domain, spam, and redundant information in the dataset, explicit meanings, and more. Tripti Agrawal et al. [38] included the identification and score assignment of emojis and emoticons as part of the data pre-processing step in sentiment analysis apart from hashtags. This provided a more comprehensive study of sentiments than many widely used methods. While this method included the manual allotment of scores to emoticons, Bahri et al. [39] used an algorithm for the computation of scores for these; the scores were assigned to emoticons in the manner they were used in tweets. This method used here is both a necessity and an included advantage to the system, given that most users of the Internet today use these terms as part of their social media posts.

Table 1 Comparison of different approaches chosen for sentiment analysis

Approach	Methodology	Advantages	Disadvantages
Lexicon-based approach [16, 40]	Unsupervised learning	Prior training on the data is not required	Requires prior dictionary available for analysis (may not be always available)
Rule-based approach [41]	Supervised and unsupervised learning	Gives excellent performance in a narrow domain	Fails to perform for larger datasets and extremely labor intensive
Machine learning approach [2, 28, 26]	Supervised and unsupervised learning	Gives the highest accuracy in terms of sentiment classification so far, prior knowledge of words not required	Classifiers are not dynamic in terms of domains, which leads to overfitting in the case of a narrow domain
Deep learning approach [42, 42–44]	Unsupervised learning	Using a variety of features, things like explicit and vernacular language can be taken into account, outperforms machine learning approach in terms of accuracy	Domain specific and requires a huge amount of computing power

We are also presenting a comparative analysis of approaches adopted in the field of sentiment analysis (Table 1). While the machine learning approach yields the best overall results in comparison to the others, different methods are adopted for the fulfillment of different objectives. Once different approaches are compared, we can find and adopt the most suitable approach to be used for the fulfillment of our goal.

4 Proposed Research Work

In this paper, we target the system of reservations in India, a topic that has attracted severe debates over the years for the people of backward castes are some of those who have been actively fighting for their rights. Twitter is the major focus for opinion mining in this domain in addition to news headlines obtained over the web. Here, we perform a comparative study based on say the general public, news agencies, and major political leaders (government) have on the issue of reservations in India. This gives us an insight into this particular situation in the country, as well as the influence of sentiments of different parties over one another.

For the above-mentioned objective (sentiment analysis on Tweets), we focus on the use of the machine learning approach for sentiment analysis. This is the best currently outperforming the others on the combinational basis of accuracy and feasibility. The

stated is for the analysis of sentiments taken from social media, due to the huge amount of data available on the same, therefore making a requirement for an approach that can learn, adapt, and accommodate an increasing pool of words. Next, for the fulfillment of sentiment analysis on data obtained from news agencies, we will adopt the Lexicon-based approach, it will perform the best when having a closely related pool of word Lexicons to choose from concerning the problem statement. We will combine the results from these two approaches to come up with our conclusions. There are several available methodologies one can adopt in the field of machine learning for sentiment analysis. The most popularly used approaches include Naïve Bayes, KNN, and SVM. Out of these, we employ the *Support Vector Machines (SVM)* with minor optimization for sentiment analysis (Fig. 3).

While the problem of sentiment analysis can best be solved through the use of *Support Vector Machines (SVM)* in the machine learning domain, it can still be optimized further to obtain better results [45]. We employ *grid search* for further

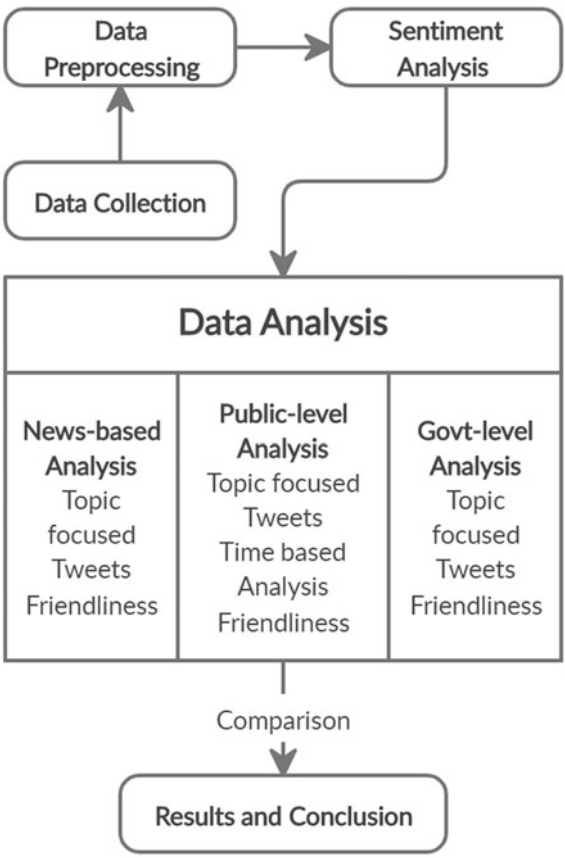


Fig. 3 Overview of proposed research work

optimization of the mentioned algorithm. As depicted in Fig. 3, we obtain the sentiments of opinion leaders (government) and the public from Twitter, and those of news from news websites. We then compare the results obtained, thereby coming to our conclusion over the difference of opinions between the government, the general public, and news media.

5 Implementation

5.1 Data Collection

The data we worked upon included the tweets about the reservation system at hand, where we scraped English tweets from *Twitter* for 4 years starting from January 2016. *Selenium* was used for data extraction, wherein relevant keywords and hashtags were used to specifically target the relevant information, for instance, India, reservation, and Dalits. All three sources (public, leaders, and news) were analyzed through these tweets. We picked many influential leaders with a huge number of public following to ensure that their opinions had a considerable impact with a bigger audience 2. The statistics for the public following of the chosen leaders are shown in Table 2. Most of the tweets originated from India; the content in addition to the respective date was recorded. While most people performed time-based analysis for a shorter period, we target the same for 4 years (2016–2019). We also discarded irrelevant tweets that might have come in during collection through the removal of those containing irrelevant words. The raw data here has to be processed further for filtration, which includes the removal of redundant and unnecessary information, in addition to other things discussed further in the next section.

5.2 Data Pre-processing

Tweets often have specific properties that make them different from general texts. Many inclusions such as URLs, hashtags, emoticons, emojis, and usernames have to be removed beforehand, else these will prove to be a hindrance to the proper processing of data. As given in Fig. 4, we have the above steps we use as part of pre-processing the tweets.

- **Filtering**: From the large volume of tweets available, we extracted relevant tweets with India as the country of origin, those within the scope of work; removal of non-English and re-tweets was also done to avoid redundancy.
- **Tagging and Tokenization**: Twitter has limits on the maximum length of tweets (140 characters), so users here tend to follow abbreviations and different grammar, with the use of least combinations of words. Here, the traditional tokenizers and POS taggers would fail to capture the exact meaning, and hence the use of

Table 2 Prominent Indian politicians on Twitter—statistics

SR no.	Name	Number of followers	SR no.	Name	Number of followers
1	Piyush Goyal	8,513,574	2	Nitin Gadkari	7,171,618
3	Raj Thackeray	572,270	4	Asaduddin Owaisi	1,421,950
5	Gautam Gambhir	10,658,343	6	Omar Abdullah	3,246,825
7	Mahesh Manjrekar	74,721	8	Mamata Banerjee	4,312,262
9	Naveen Jindal	291,379	10	Hema Malini	1,086,855
11	Sharad Pawar	1,770,138	12	Shivraj Chouhan	6,456,391
13	Akhilesh Yadav	11,975,646	14	Yogendra Yadav	1,528,652
15	Priya Dutt	172,055	16	S. M. Krishna	28,143
17	Rajeev Chan	253,169	18	Narendra Modi	57,791,321
19	Arvind Kejriwal	19,195,527	20	Arun Jaitey	15,203,657
21	Amit Shah	20,495,348	22	Rajnath Singh	17,238,808
23	Rahul Gandhi	14,415,484	24	Smriti Irani	11,110,643
25	Shashi Tharoor	7,571,960	26	Vijay Rupani	2,673,868
	Average	8,662,715		Std. deviation	1,170,512
	SDP	11,486,317		SDS	11,705,123

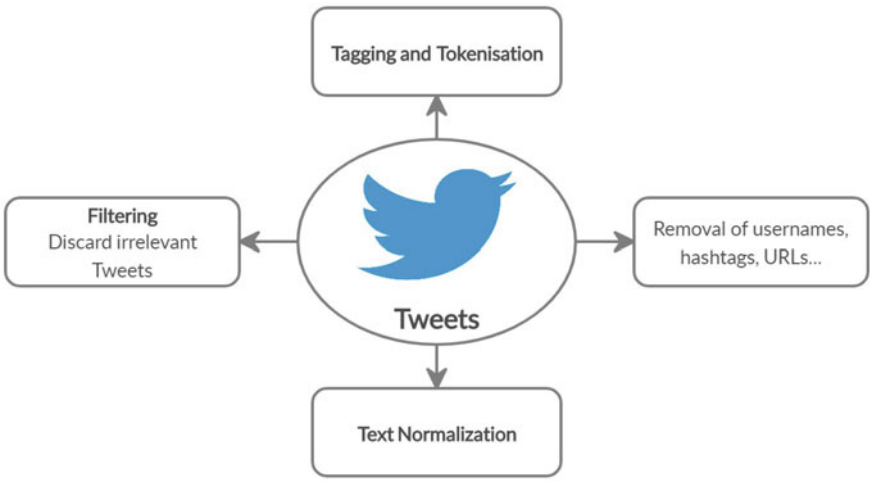


Fig. 4 Pre-processing of data

text processing library ArkTweetNLP is adopted to tokenize and tag the tweets. This library is primarily aimed at dealing with informal and conversational texts obtained from social media.

- **Removal of unwanted characters**: The Internet consists of a large number of symbols and nonstandard language as part of its data. The inclusion of such data might result in varied results, and hence have to be taken care of beforehand. Here, we cleaned the data through the removal of stop words, numbers, punctuation, URLs, and hashtags.
- **Text normalization**: From the large pool of words with related meanings, it is a good practice to club those to remove redundancy. For the same, a text normalization technique called Lemmatization is adapted to convert a variety of inflected forms of the same word into a single word, such that the base has the same meaning that covers all selected words.
- Here, we use the *NLTK* and *Textblob* from Python for the said correction and modification to our data. *Term Frequency-Inverse Document Frequency (TF-IDF)* is used to convert the textual data to the required representation matrix of numbers.

5.3 Sentiment Analysis

This is the core step toward identifying sentiments toward reservations in India. Once the tweets are identified, they have to be classified according to several features. As a result, the proper quality of the process is necessary for the desired results of subsequent steps. Every tweet is a representation in the form of an opinion which has to be classified from *negative* to *positive* on a scale of -2 to 2.

Algorithm 1 Algorithm adopted using Grid Search

```
Select the appropriate value of
Here, C = [0.1, 1, 10], gamma = [1, 0.1, 0.01, 0.001, 0.0001], Kernel = ['rbf'] i in C j
in gamma Train SVM classifier with required data set Obtain accuracy
Select parameters yielding rest parameters
Using obtained parameters, calculate Precision, Recall, F1 score
Generate Confusion Matrix
```

There are many pre-trained tools to perform sentiment analysis. Here, we are adopting a supervised approach; SVM is found to show the best results among the other classifiers used. Here, linearly inseparable data is mapped to a high-dimensional vector space for the categorization of data points. Using *tenfold cross-validation*, we utilize *SVM with grid search optimization* (Algorithm 1). Grid search helps with hyperparameter tuning to build and evaluate our model for each combination of parameters. The results in the next section support this methodology, as we obtain a higher accuracy than compared to the traditional approach.

Since the collected tweets do not come with prior sentiments, we figured out a way to train our model. We trained our model on the sentiment scores obtained from

Textblob. Tweets were assigned a score of $[-2, 2]$ depending on their variance from very negative to very positive, with 0 signifying a neutral attitude. We removed tweets signifying a neutral opinion (score of 0), since these disrupt our representations of sentiments, affecting factors like mean and variance.

Figure 5 depicts the sentiments obtained through tweets over a timeline of 4 years, from January 2016 to December 2019. This helps us to explore the change in public opinion over time, something that can be leveraged for the real-time depiction of the change of sentiments. Each tweet in the dataset has a unique timestamp that helps us identify meaningful patterns over time. We find that most tweets range in the early positive range between scores of 0.3 and 1. The sentiments of people have taken regular anomalies, with decreasing support over the years. Mostly, major drops in the scores can be attributed to major events and protests surrounding reservations in India. This shows how the citizens of the country have had a negative change in attitude toward reservations over the years, which is clear from the tweets posted by them on the platform. Figure 6 shows the comparison among the three different sources chosen for their attitude toward the problem of reservation in India. These are the most common words that have been used toward reservation system in India. We find decreasing scores of sentiments in public tweets, opinion leaders, and newspapers, showing friendliness in the respective approaches of these parties. From the given data, it is evident that the public has the most positive take on this situation, while the other two domains fail to stay on the same side and showcase decreasing sentiment scores. Next, we generated a Word Cloud to depict the most widely used words throughout the entire data. This is depicted in Fig. 7; the most widely used words are "*scheduled caste*" and "*scheduled tribe*". We can find multiple variables that help us understand the emotions of people behind these tweets. A majority of words are found to be attributed to facts and related vocabulary, and we find little usage of words that affect the sentiment score (Fig. 8).

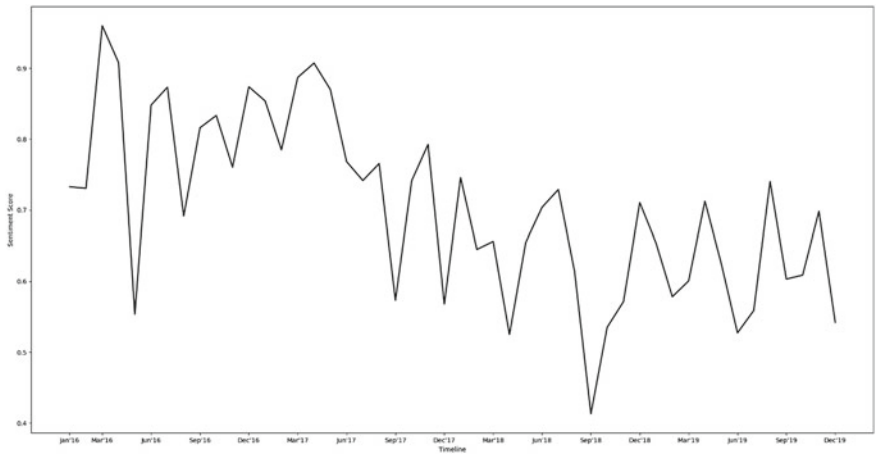


Fig. 5 Time series public sentiments

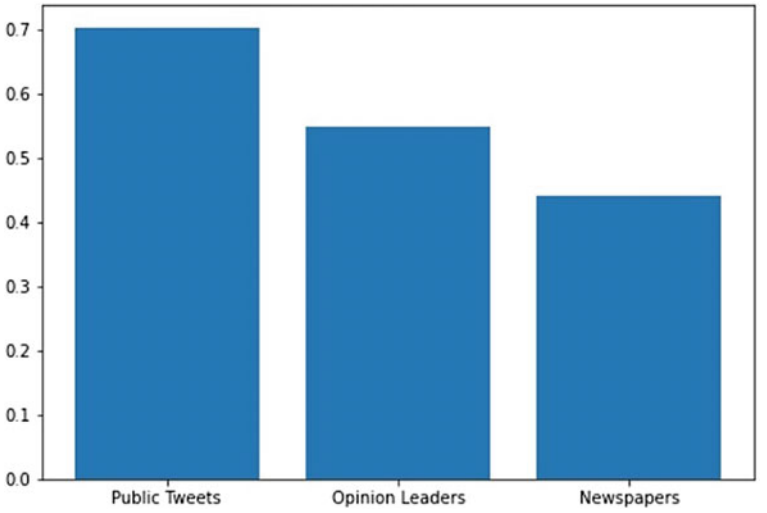


Fig. 6 Overall comparison of the domains

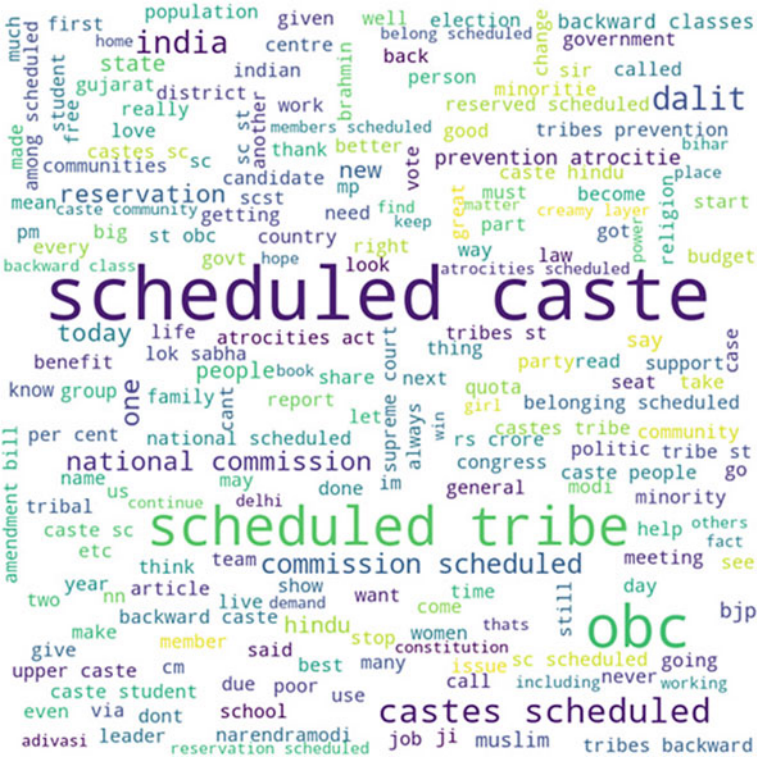


Fig. 7 World cloud depicting the tags used

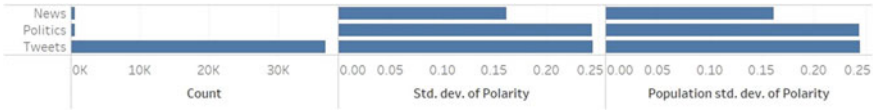


Fig. 8 Comparative study using different parameters

6 Results and Discussion

Equations (1) and (2) depict the formulae for the different parameters we have used to evaluate the model and generate the confusion matrix, given in Fig. 9. This comparison shows the sources that have been useful in the accumulation of the data. We experienced many neutral tweets arising out of the analysis; these were removed from the dataset as we experienced unwanted deviations from the mean.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad \text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad \text{F1 Score} = 2 * \frac{\text{Prescicsion} * \text{Recall}}{\text{Prescicsion} + \text{Recall}} \tag{2}$$

TP:TruePositive TN:TrueNegative FP:FalsePositive FP:FalseNegative

Traditional SVM is known to outperform other classifiers [46]; here we used grid search to fetch even better results. Sentiment analysis in this work has been done by both approaches and respective results have been shown in Table 3. The dataset used

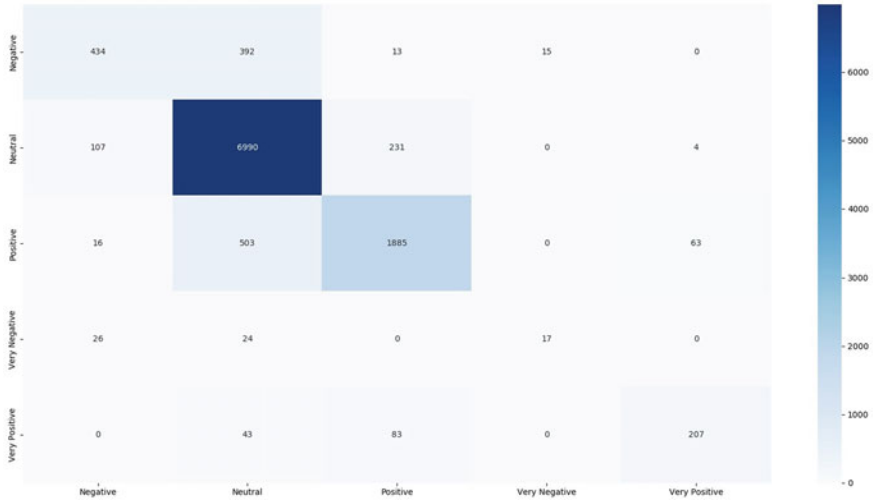


Fig. 9 Confusion matrix

Table 3 Comparison between sentiment classifiers used per the perspective of the interpreter

S. no.	Methodology	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
1	SVM	83	84	83	82
2	SVM with grid search	86	86	86	86

for the former was also used for the latter here. Pre-trained classifiers would perform remarkably poor on this testing data since they are designed to consider polarity alone, and ignore the differences arising out of political remarks [47]. Hence, traditional models may be inadequate in such political scenarios as Table 3.

In this approach, we used employed tenfold cross-validation; we divided the data collected into a training test and a testing set in the ratio of 70:30. The testing set was the same for both approaches employed here.

7 Conclusion

Sentiment analysis had transformed the huge volume of data available on the Internet into meaningful insight. In this paper, we presented a comparative analysis of research work done on this field, followed by a novel approach of using SVM with grid search to perform sentiment analysis on the issue of *reservations* for 4 years (starting 2016). We find that our model outperforms the traditional use of SVM alone, with better results. There are many directions where this work can extend in the future. First, we would like to train the sentiment analyzer with a greater number of tweets. Second, we would like to test the sentiments of political leaders from alternate sources, such as public speeches and rallies, giving us a precise understanding of the person's attitude toward the matter. Next, we would like to use SVM with Bayesian optimization for tuning hyperparameters for the set. Such studies can be used by a variety of researchers in different disciplines to identify the political situation in a country; similarities and/or differences may be used for important decision-making.

References

1. T. Singh, M. Kumari, Role of text pre-processing in Twitter sentiment analysis. Procedia Comput. Sci. **89**, 549–554 (2016). https://doi.org/10.1016/j.procs.2016.06.095, http://www.sciencedirect.com/science/article/pii/S1877050916311607, Twelfth International Conference on Communication Networks, ICCN 2016, August 19–21, 2016, Bangalore, India; Twelfth International Conference on Data Mining and Warehousing, ICDMW 2016, August 19–21, 2016, Bangalore, India; Twelfth International Conference on Image and Signal Processing, ICISP 2016, August 19–21, 2016, Bangalore, India
2. N.A. Abdullah, A. Feizollah, A. Sulaiman, N.B. Anuar, Challenges and recommended solutions in multi-source and multi-domain sentiment analysis. IEEE Access **7**, 144957–144971 (2019)

3. M. Miller, C. Sathi, D. Wiesenthal, J. Leskovec, C. Potts, Sentiment flow through hyperlink networks (Janury 2011)
4. R. Bhargava, Y. Sharma, G. Sharma, Atssi: abstractive text summarization using sentiment infusion. Procedia Comput. Sci. **89**, 404– 411 (2016). https://doi.org/10.1016/j.procs.2016.06.088, http://www.sciencedirect.com/science/article/pii/S187705091631153X, Twelfth International Conference on Communication Networks, ICCN 2016, August 19–21, 2016, Bangalore, India; Twelfth International Conference on Data Mining and Warehousing, ICDMW 2016, August 19–21, 2016, Bangalore, India; Twelfth International Conference on Image and Signal Processing, ICISP 2016, August 19–21, 2016, Bangalore, India
5. Z. Polkowski, J. Vora, S. Tanwar, S. Tyagi, P.K. Singh, Y. Singh, Machine learning-based software effort estimation: an analysis, in *2019 11th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)* (2019), pp. 1–6. https://doi.org/10.1109/ECAI46879.2019.9042031
6. R. Aasha, G.D.P.B. Rahul Jain, Sentimental analysis and detection of rumours for social media data using logistic regression. Int. J. Eng. Adv. Technol. (IJEAT) **9** (2019)
7. 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). https://doi.org/10.1016/j.cosrev.2017.10.002, http://www.sciencedirect.com/science/article/pii/S1574013717300606
8. M. Devika, C. Sunitha, A. Ganesh, Sentiment analysis: a comparative study on different approaches. Procedia Comput. Sci. **87**, 44–49 (2016). https://doi.org/10.1016/j.procs.2016.05.124, http://www.sciencedirect.com/science/article/pii/S187705091630463X, Fourth International Conference on Recent Trends in Computer Science Engineering (ICRTCSE 2016)
9. A.J.J. Mary, L. Arockiam, Jen-ton: a framework to enhance the accuracy of aspect level sentiment analysis in big data, pp. 452–457 (Nov 2017). https://doi.org/10.1109/ICICI.2017.8365391
10. S. Tanwar, Q. Bhatia, P. Patel, A. Kumari, P.K. Singh, W.C. Hong, Machine learning adoption in blockchain-based smart applications: the challenges, and a way forward. IEEE Access **8**, 474–488 (2020). https://doi.org/10.1109/ACCESS.2019.2961372
11. K. Elshakankery, M.F. Ahmed, HILATSA: a hybrid incremental learning approach for Arabic tweets sentiment analysis. Egypt. Inform. J. **20**(3), 163–171 (2019). https://doi.org/10.1016/j.eij.2019.03.002, http://www.sciencedirect.com/science/article/pii/S1110866518302123
12. W. Medhat, A. Hassan, H. Korashy, Sentiment analysis algorithms and applications: a survey. Ain Shams Eng. J. **5**(4), 1093–1113 (2014). https://doi.org/10.1016/j.asej.2014.04.011, http://www.sciencedirect.com/science/article/pii/S2090447914000550
13. A. Mars, M.S. Gouider, Big data analysis to features opinions extraction of customer. Procedia Comput. Sci. **112**, 906–916 (2017). https://doi.org/10.1016/j.procs.2017.08.114, http://www.sciencedirect.com/science/article/pii/S1877050917314709, Knowledge Based and Intelligent Information Engineering Systems: Proceedings of the 21st International Conference, KES-2017, 6–8 September 2017, Marseille, France
14. K. Sailunaz, R. Alhajj, Emotion and sentiment analysis from twitter text. J. Comput. Sci. **36**, 101003 (2019)
15. C. Comito, A. Forestiero, C. Pizzuti, Twitter-based influenza surveillance: analysis of the 2016–2017 and 2017–2018 seasons in Italy, pp. 175–182 (June 2018). https://doi.org/10.1145/3216122.3216128
16. A. Alfarrarjeh, S. Agrawal, S.H. Kim, C. Shahabi, Geo-spatial multimedia sentiment analysis in disasters, pp. 193–202, October 2017. https://doi.org/10.1109/DSAA.2017.77
17. R.M. Cury, Oscillation of tweet sentiments in the election of João Doria Jr. for mayor. J Big Data (2019). https://doi.org/10.1186/s40537-019-0208-1
18. B.J. Chellia, K. Srivastava, J. Panja, R. Paul, Sentiment analysis of Twitter for election prediction. Int. J. Eng. Adv. Technol. (IJEAT) **9** (2019)

19. A. Dwivedi, R. Pant, An algorithmic implementation of entropic ternary reduct soft sentiment set (ETRSSS) using soft computing technique on big data sentiment analysis (BDSA) for optimal selection of a decision based on real-time update in online reviews. J. King Saud Univ.—Comput. Inform. Sci. (2019). https://doi.org/10.1016/j.jksuci.2019.10.001, http://www.sciencedirect.com/science/article/pii/S1319157819310547
20. E. Kušen, M. Strembeck, G. Cascavilla, M. Conti, On the influence of emotional valence shifts on the spread of information in social networks (2017). https://doi.org/10.1145/3110025.3110031
21. L.H. Son, A. Kumar, S.R. Sangwan, A. Arora, A. Nayyar, M. Abdel-Basset, Sarcasm detection using soft attention-based bidirectional long short-term memory model with convolution network. IEEE Access **7**, 23319–23328 (2019). https://doi.org/10.1109/ACCESS.2019.2899260
22. R. Yadav, A.V. Kumar, A. Kumar, News-based supervised sentiment analysis for prediction of futures buying behaviour. IIMB Manag. Rev. **31**(2), 157–166 (2019). https://doi.org/10.1016/j.iimb.2019.03.006, http://www.sciencedirect.com/science/article/pii/S0970389619301569
23. L. Wang, J. Niu, S. Yu, Sentidiff: combining textual information and sentiment diffusion patterns for twitter sentiment analysis. IEEE Trans. Knowl. Data Eng., p. 1 (2019). https://doi.org/10.1109/TKDE.2019.2913641
24. X. Wang, C. Zhang, M. Wu, Sentiment classification analysis of Chinese microblog network, pp. 123–129 (2015)
25. H.A. Kheruwala, J.V. Shah, J.P. Verma, Comparative study of sentiment analysis and text summarization for commercial social networks, in *Emerging Technology Trends in Electronics, Communication and Networking*, ed. by S. Gupta, J.N. Sarvaiya (Springer Singapore, Singapore, 2020), pp. 213–224
26. H. Rahab, A. Zitouni, M. Djoudi, Sana: sentiment analysis on newspapers comments in Algeria. J. King Saud Univ.—Comput. Inform. Sci. (2019). https://doi.org/10.1016/j.jksuci.2019.04.012, http://www.sciencedirect.com/science/article/pii/S1319157819301557
27. A. Gupta, V. Dengre, H.A. Kheruwala, M. Shah, Comprehensive review of text-mining applications in finance. Financ. Innov. **6**(1), 1–25 (December 2020). https://doi.org/10.1186/s40854-020-00205, https://ideas.repec.org/a/spr/fininn/v6y2020i1d10.1186_s40854-020-00205-1.html
28. M. Dragoni, G. Petrucci, A neural word embeddings approach for multi-domain sentiment analysis. IEEE Trans. Affect. Comput. **8**(4), 457–470 (2017). https://doi.org/10.1109/TAFFC.2017.2717879
29. L. Li, Y. Wu, Y. Zhang, T. Zhao, Time+user dual attention based sentiment prediction for multiple social network texts with time series. IEEE Access **7**, 17644–17653 (2019). https://doi.org/10.1109/ACCESS.2019.2895897
30. R. Xia, J. Jiang, H. He, Distantly supervised lifelong learning for large-scale social media sentiment analysis. IEEE Trans. Affect. Comput. **8**(4), 480–491 (2017). https://doi.org/10.1109/TAFFC.2017.2771234
31. A. Shunmuga Sundari, R. Shenbagavalli, Dominant lexicon based Bi-LSTM for emotion prediction on a text. Int. J. Eng. Adv. Technol. (IJEAT) **8** (2019)
32. L. Yu, J. Wang, K.R. Lai, X. Zhang, Refining word embeddings using intensity scores for sentiment analysis. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(3), 671–681 (2018). https://doi.org/10.1109/TASLP.2017.2788182
33. P.K. Nikitha Kumari, Dynamic data analysis and decision making on twitter data. Int. J. Eng. Adv. Technol. (IJEAT) **9** (2019)
34. K. Sailunaz, R. Alhajj, Emotion and sentiment analysis from Twitter text. J. Comput. Sci. **36**, 101003 (2019). https://doi.org/10.1016/j.jocs.2019.05.009, http://www.sciencedirect.com/science/article/pii/S1877750318311037
35. R. Gupta, N. Gowalker, S.D. Joshi, S. Patil, Predicting risk in sentiment analysis using machine learning. Int. J. Eng. Adv. Technol. (IJEAT) **9** (2019)
36. R. Bhuvaneswari, S. Ravichandran, A New feature based scoring technique to discover sentiments present in the online product **9** (2019)

37. P. Ray, A. Chakrabarti, A mixed approach of deep learning method and rule based method to improve aspect level sentiment analysis. Appl. Comput. Inform. (2019). https://doi.org/10.1016/j.aci.2019.02.002, http://www.sciencedirect.com/science/article/pii/S2210832718303156
38. A.S. Tripti Agrawal, An effective knowledge-based pre-processing system with emojis and emoticons handling on Twitter and Google+. Int. J. Eng. Adv. Technol. (IJEAT) **8** (2019)
39. S. Bahri, P. Bahri, S. Lal: A novel approach of sentiment classification using emoticons. Procedia Comput. Sci. **132**, 669–678 (2018). https://doi.org/10.1016/j.procs.2018.05.067, http://www.sciencedirect.com/science/article/pii/S1877050918307993, International Conference on Computational Intelligence and Data Science
40. Devansh Nema, Devansh Bhudolia, R. Anitha, Sentiment analysis of movie using artificial neural network. Int. J. Civil Eng. Technol. (IJCIET) **8** (2017)
41. J. Cao, K. Zeng, H. Wang, J. Cheng, F. Qiao, D. Wen, Y. Gao, Web-based traffic sentiment analysis: methods and applications. IEEE Trans. Intell. Transp. Syst. **15**(2), 844–853 (2014). https://doi.org/10.1109/TITS.2013.2291241
42. A. Feizollah, S. Ainin, N.B. Anuar, N.A.B. Abdullah, M. Hazim, Halal products on twitter: data extraction and sentiment analysis using stack of deep learning algorithms. IEEE Access **7**, 83354–83362 (2019). https://doi.org/10.1109/ACCESS.2019.2923275
43. M. Usama, W. Xiao, B. Ahmad, J. Wan, M.M. Hassan, A. Alelaiwi, Deep learning based weighted feature fusion approach for sentiment analysis. IEEE Access **7**, 140252–140260 (2019). https://doi.org/10.1109/ACCESS.2019.2940051
44. B. Zhang, X. Xu, M. Yang, X. Chen, Y. Ye, Cross-domain sentiment classification by capsule network with semantic rules. IEEE Access **6**, 58284–58294 (2018). https://doi.org/10.1109/ACCESS.2018.2874623
45. M. Ahmad, S. Aftab, S. Bashir, N. Hameed, I. Ali, Z. Nawaz, SVM optimization for sentiment analysis. Int. J. Adv. Comput. Sci. Appl. **9**(4) (2018). https://doi.org/10.14569/IJACSA.2018.090455
46. N. Zainuddin, A. Selamat, Sentiment analysis using support vector machine. pp. 333–337 (Sept 2014). https://doi.org/10.1109/I4CT.2014.6914200
47. I. AlAgha, O. Dahrooj, Multi-level analysis of political sentiments using Twitter data: a case study of the Palestinian–Israeli conflict. Jordanian J. Comput. Inform. Technol. **5** (1), 12 (2019). https://doi.org/10.5455/jjcit.711562700251

Recommending Books Using RNN

Mala Saraswat, Rishi Saraswat, and Renu Bahuguna

Abstract Many recommendation algorithms have been developed using collaborative filtering such as latent matrix factorization. But most of these algorithms are based on other similar users rating of the books and have certain limitations. Reviews given to books provide a valuable source of information for recommender systems. Algorithms used in area of natural language processing are based on similarity between word vectors in the two texts or documents. Often context between these words are missing leading to poor recommendation as it requires semantics to be analyzed between texts from the given reviews. Deep learning has been widely used in natural language processing and has obtained a great progress in this field. In this paper, we use Recurrent Neural Networks (RNN) as a deep learning approach for recommending books. RNN is an improvement over the existing models as instead of each input unit being independent as in logistic regression and other neural networks each neuron can use its internal memory to maintain information about the previous units. Hence, the context between reviews is maintained by memory units and hence a more accurate classification can be made that leads to better recommendation.

Keywords Reviews · Recommender systems · Recurrent neural network · Deep learning · Matrix factorization

1 Introduction

Recommender system analyzes the online activities of users to anticipate their preferences among a set of items. Recommenders are broadly classified into two categories: collaborative-based and content-based recommender system. Collaborative model (CF) is based on similarity of items or users to make predictions about what a user may be interested in. In CF users with similar preferences and tastes are used as a measure to recommend items [1]. In conventional memory-based CF techniques

M. Saraswat (✉)
ABES Engineering College, Ghaziabad, UP, India

R. Saraswat · R. Bahuguna
Doon Institute of Engineering and Technology, Rishikesh, Uttrakhand, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_7

based on their ratings for different items similar users are found. User-based recommender system uses these similar users as the basis of their recommendations. In model-based CF, a rating pattern is examined and analyzed to estimate a model to recommend items [2].

Content-Based Recommender system (CBR) is based on analysis and examination of attributes of various items to make predictions. Content or features of items based on user profile is utilized in content-based recommender systems. Recommendations are based on user profile developed by user attributes [3]. Review-based recommender system uses user-generated content such as reviews to recommend items [4]. Review-based recommender system exploits various review features such as topics [5], frequent terms [6], sentiments [7], and emotions [8] for recommending items. In [9], Chakraverty et al. further used emotion for recommending items of target domain to user in source domain using cross-domain recommendations. In this paper, we propose an approach where we use deep learning model to process the book reviews and classify them into different genres. Based on this classification we recommend the books to different users. Conventional recommender systems based on reviews have certain shortcomings as they do not preserve the context between various reviews leading to ambiguous recommendations [1].

Deep learning improves the shortcomings offered by other methods and gives us more efficient recommender systems [10]. Deep learning methods have become more powerful in tackling recommender systems such as news, music, mobile apps recommendations, and fashion articles. To make strong recommendations, we need to take into consideration the context of the reviews that are being considered, so as to make a more coherent classification of genres. Hence, we have used a Recurrent Neural Network (RNN) to process our user reviews, so that a more accurate prediction of book can be made for a user based on its genres. RNN carries along with its memory units which preserves the old information, and hence an effective recommendation can be made [11]. This makes them applicable to tasks such as connected handwriting recognition or speech recognition. For instance, consider the phrase "This movie is bad", from simple recommendation and bag-of-words model we can easily recommend that user did not liked the movie. Let's now consider the phrase "This movie is not bad", by simple recommendation we may predict that user did not like the movie, but in actual user did like the movie. We overcome such problem by using recurrent neural network which takes into consideration the word "not" before the word bad, and hence is more effective in making predictions.

In this paper, we propose a review-based recommender system using RNN for recommending book items. Section 2 discusses the related work. In Sect. 3, we discuss the proposed approach. Experiments and results are discussed in Sect. 4.

2 Related Work

Recommender systems are one of the most successful and widespread applications of machine learning technologies in today's business world. There are many existing

algorithms which have been used for recommendation. Machine learning algorithms in recommender systems are typically classified into two categories—content-based and collaborative filtering methods. Many recommender systems combine both CF- and CBF-based approaches for better recommendations. Collaborative methods calculate similarity from interactions and content-based methods are based on similarity of item attributes.

2.1 User-Based Collaborative Filtering

The user-based collaborative filtering is very much similar to item-based collaborative filtering. In user-based collaborative filtering rather than measuring the similarity between two items, similarity between two users is considered. A correlation-based method of computing similarity between two users *u, v* is used which is the same as item based. Let us denote that similarity between user *u* and *v* as $s_{u,v}$. The recommendation for a user *u* on item *i* is computed by calculating weighted sum of most similar users as shown in Eq. (1) as follows:

$$P_{u,i} = \frac{\Sigma_v (r_{v,i} * s_{u,v})}{\Sigma_v s_{u,v}} \tag{1}$$

Here *v* represents similar user and *r* is the rating given by *v* to *i.*

2.2 Item Similarity Computation

The item-based collaborative filtering is almost identical to user-based collaborative filtering. In item-based collaborative filtering, similarity between two items is computed rather than calculating similarity between two users. The main idea of computing similarity between two products/items *i* and *j* is to differentiate those users who have rated both the item that are being considered at this point of time and then perform a similarity metric: $s_{i,j}$ to calculate the similarity between two items *i* and *j* and make recommendations for users.

$$P_{u,i} = \frac{\sum_{similar\ items\ to\ j} \left(r_{u,j} + s_{i,j}\right)}{\sum_{similar\ items\ to\ j} \left|s_{i,j}\right|} \tag{2}$$

Here *v* represents similar user and *r* is the rating given by *u* to *i.*

2.3 Matrix Factorization Method

Other important algorithm used for recommendation is matrix factorization which was established to reduce the problem dimensionality [12]. These algorithms are used for classification or regression and are much more computationally efficient on large sparse datasets. Also this algorithm is used to scale large datasets, like that of Netflix. This property of matrix factorization makes it widely used tool in recommendation where user count and item count are very large and actual number of recommendations is small.

2.4 Different Model Comparison and Deep Learning Approach

The user-based CF has some limitations. First, it is difficult to measure similarity between users than items and the second is the issue of scalability. The issue of scalability is when as the number of items or users increases computation time of algorithm also increases.

Since item-based CF computes similarity in an offline basis, it was suggested to overcome the scalability problem. An assumption is made, which is if a user likes a product or an item, he/she will also like the similar item that has been already bought by the user. The latent information about the relationship between items and users cannot be furnished by discrete values such as ratings. Some of the advantages of content-based recommender system is that first, user independence is achieved through ratings of active users, second, transparency is provided to users and third, also a new user can be recommended even if it has not recommended anything yet.

Some limitations offered by abovementioned algorithms are, first, it is difficult to generate features of items in certain areas, second, it suffers from overspecialization problem, and third, it is difficult to acquire recommendations from users, and hence recommendations are not validated.

2.5 Genre-Based Classification

In our approach, we have tried classifying book reviews and story plots among various genres. Genres play a very important role in learning user interests and recommending items for users. Books are part of entertainment domain and a common feature of most entertainment domains is their categorization into major genres. Both the literary genres and film genres are based on the concept of genres as prototypical narratives that invoke emotions [13]. Genres group books into various categories. Genres besides categorizing the books are also very important for users as they have in mind what is expected from the books while reading and whether to like it or not.

Genres further can be used in cross-domain recommendations, for instance, genres can be used as an intermediate for making recommendations for the liking of books based on movie likings.

Taking cues from all above different approaches, we recommend our proposed approach in next section that uses RNN for classifying the available reviews into genres and then using similar genre books are recommended.

3 Proposed Work

Our proposed solution is to perform a recurrent neural network on book reviews by using word embedding which is used to produce a dense representation of review vectors. These embeddings use a pre-trained model of glove embedding [14]. As inputs we have positive reviews (ratings >= 3) and as outputs we have multivalued, multiclass classification of review vector into genres. As a result, we can classify reviews into various genres based on user preferences. Further, Fig. 1 depicts the structure of the proposed approach review-based classification of books into genres for recommender system. The approach consists of four modules, namely, M1—preparation of Data, M2—pre-processing of data, M3—feeding of data into a RNN, and M4—classification into various genres. Each module is explained in following subsections.

The RNN-based genre classification approach illustrated in Fig. 1 can be extended to almost every domain of entertainment. Book reviews are used to exemplify our approach. From Bookcrossings dataset, book ratings were extracted and reviews were extracted from Amazon.com. Each segment as shown in Fig. 1 is elaborated on as given below.

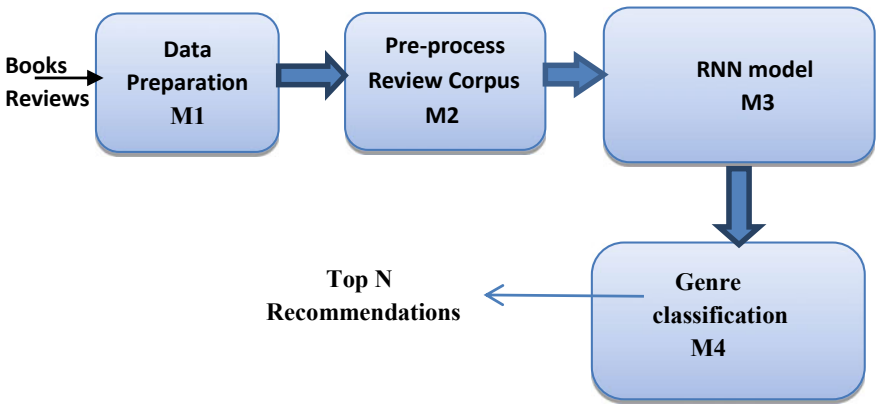


Fig. 1 Flow diagram of the proposed of genre classification of books' review

3.1 Data Preparation M1

Module M1, i.e., the data preparation module, extracts the books' title and books' reviews from website by multiple users from the dataset. Through reviews, users express their opinions, feelings, and experiences of items bought by them on e-commerce websites.

All reviews corresponding to particular books are appended to form a corpus of text. Hence, the result of this segment includes books' title and concatenated list of reviews for particular books.

3.2 Pre-processing of Review Corpus M2

Most of the times while reviewing the books' users use a relaxed or informal style of writing which results in noise and syntactic errors. Segment M2 refines the corpus of reviews by removing redundant and irrelevant content that do not provide any meaning to the review corpus. Following steps are incorporated in this segment.

(a) Removal of stop words. (b) Numeric characters removal and only alphabets are retained. (c) Compression of words such as "cuteeeeee" to "cute". (d) Stemming of words to their root words such as happiness, happily to happy.

3.3 Recurrent Neural Network Model M3

While recommendation systems have been widely used till now. In this work, we try to leverage RNN for books recommendations. Our main aim for is to recommend some books based on the reviews of books and user ratings for different books. Based on the ratings given by the user we first predict user ratings for all the items using a neural network running on the latent factors of matrix factorization. Neural networks consisting of hidden layers and using backward propagation to correct the weights based on the labeled data give an accuracy better than the existing matrix factorization and collaborative filtering techniques. Based on the user's rating for all the books in the dataset we could predict the dominant genres in which a user seems most interested in. Further, to recommend books for a particular user, we used various reviews and storylines of the books to classify into various genres. In their work, Patel et al. used sentiment analysis for movie review to analyze the information in the form of number of reviews where opinions are either positive or negative using RNN [15].

Huang et al. proposed a top N interactive recommender system that can be viewed as Markov decision processes, wherein the interactions between recommender system and user are simulated by the Recurrent Neural Network (RNN) [16].

Module M3 uses RNN to run on the long labeled reviews to generate the model. RNN is a widely used neural network and is different from other neural networks in a sense that instead of treating each input unit independently, each input is dependent on previous units and this is done with the use of memory called Long Short-Term Memory (LSTM) [17].

We took our datasets from Bookcrossing dataset and extracted reviews from Amazon website. After pre-processing, a word embedding that is a dense vector representation of reviews is created. Word embedding is again two-layered neural network which itself learns context of each word with respect to its use with neighboring terms in reviews. This dense representation of vectors helps us to make strong relation of the reviews to their corresponding outputs, and hence a more effective performance.

Our proposed approach divides the data into a training set and test set. RNN model consists of four layers. First is the embedding layer or the input layer, second and third layers are the LSTM (long short time memory) layers. LSTM layers are to prevent exploding and vanishing gradient in a recurrent neural network. Also two dropout layers (with a fraction of 0.3) were added to prevent overfitting of the RNN. Dropout basically drops some portion neurons so that network becomes less sensitive to specific weight of neurons. This results in a network that is capable of better generalizations and is less likely to overfit the data. Finally, the fourth layer is the output dense layer with sigmoid activation function added. Model was compiled with a RMS prop optimizer, loss binary cross entropy, and accuracy metrics. Our model was run on 20 epochs and a batch size of 256. The optimized weights for the model are saved to be used later so as to train the network once only.

3.4 Genre Classification M4

Module M4 generates the genres of any book based on its review and storylines. In this module, reviews corresponding to a book are given as input based on the already trained model generated by module M3. Genres are generated for that book based on its reviews. This module then classifies the books based on these genres. Books belonging to same set of genres are recommended. These predictions have been verified by comparing with genres of actually rated books by the user.

4 Experimental Results

This section presents the experimental results conducted on using Bookcrossing datasets. We have processed the books' dataset as discussed in Sect. 3.1. After pre-processing the dataset, we divide our data into training and test set, where specified labels are multivalued multiclass classification of genres. One hot encoding is applied to the dataset and feature scaling is done after that to improve the training process.

Then, once the data is modeled, it is fed into an RNN architecture consisting of input layer, hidden LSTM layer, and an output layer. Appropriate parameters are chosen such as learning rate = 0.001 and number of epochs = 50. After applying the RNN model as discussed, we classify our books' reviews into 18 genres. From the results elaborated in Table 1, we can clearly see that recurrent neural network-based recommendation has precision of 0.82 as compared to 0.77 for matrix factorization approach for top 5 recommendations. Figure 2 illustrates the results graphically.

As can be seen in Fig. 2, RNN has best F-measure of 0.80 compared to other approaches using user-based CF, item-based CF, or matrix factorization. From above observations, we infer that using a Deep Learning model using recurrent neural network increases our accuracy, precision, recall, and overall F-measure and decreases the validation loss, and hence the root mean squared error.

Table 2 shows the variation of precision, accuracy, recall, and F-measures for Top-N recommendations. Figure 3 also compares accuracy, precision, and recall for RNN-based recommendations in book domain for Top-N recommendations graphically. The result shows accuracy and precision increases with N, i.e., number of

Table 1 Performance comparison of RNN network for TOP 5 recommendations

S. no.	Accuracy	Precision	Recall	F-measure
User-based CF	0.75	0.72	0.73	0.725
Item-based CF	0.71	0.69	0.71	0.699
Matrix factorization	0.79	0.77	0.78	0.775
RNN	0.84	0.82	0.78	0.80

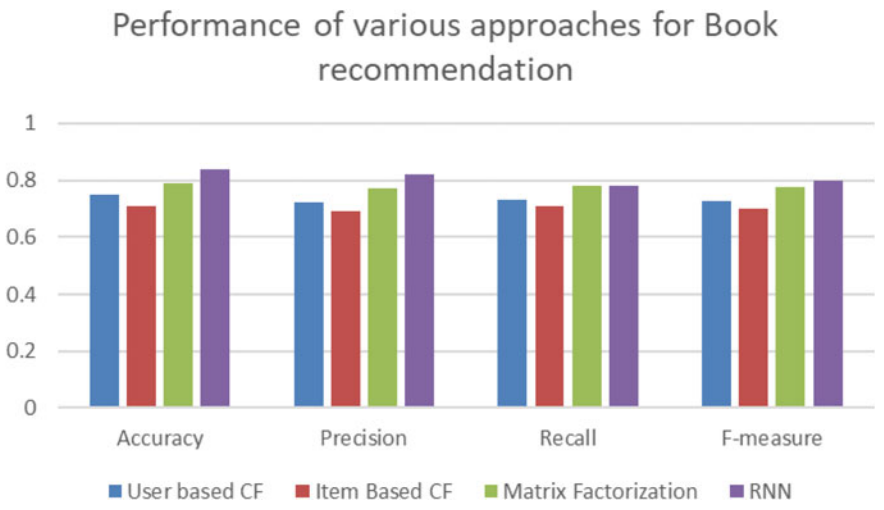


Fig. 2 Performance comparison of different book recommendation approaches

Table 2 Performance of book leveraging RNN network on book reviews for genre classification

Top-N recommendation	Accuracy	Precision	Recall	F-measure
Top 5	0.77	0.79	0.81	0.79
Top 10	0.84	0.82	0.79	0.80
Top 15	0.81	0.84	0.77	0.803
Top 20	0.82	0.85	0.77	0.808

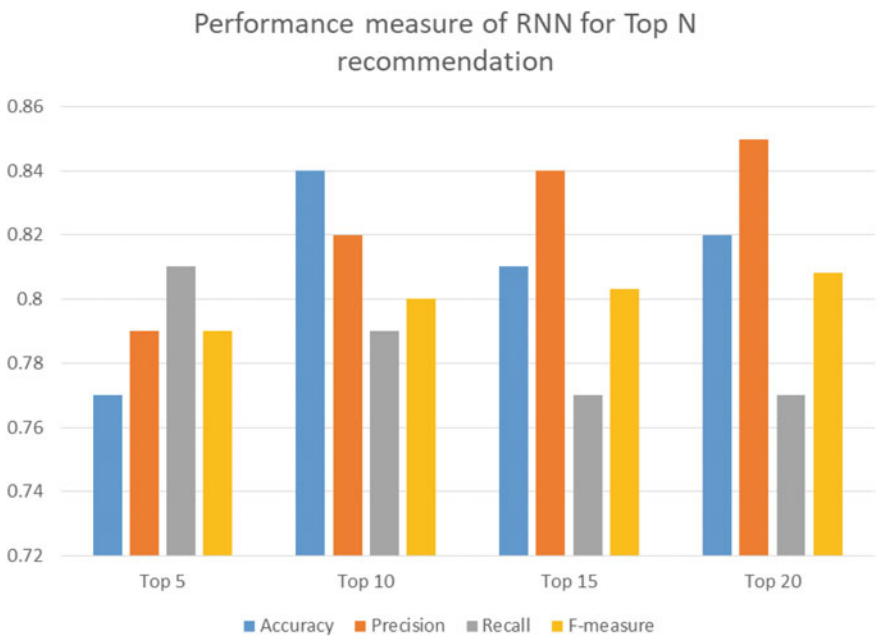


Fig. 3 Performance measures for Top- N recommendation using RNN in Books domain

recommendation. As number of N increases, F-measure also increases to 0.808 for Top-20 recommendation.

5 Conclusions

This paper presents and scrutinizes recommendation model based on genre classification of the books' reviews by various users using deep learning approach. We examined our genre classification based on books' reviews. The results achieved show that review-based genre classification provides high-accuracy recommendations using recurrent neural network than any other neural networks. An improved user model is generated as user reviews for the books are used as intermediate source

to boost recommendations. In future, we extract more features from user-generated content for recommender system and use different deep learning approaches. We can also delve deep to use deep learning for cross-domain recommendations.

References

1. G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. **6**, 734–749 (2005)
2. J. Ben Schafer et al., Collaborative filtering recommender systems, in *The Adaptive Web* (Springer, Berlin, Heidelberg, 2007), pp. 291–324
3. M.J. Pazzani, D. Billsus, Content-based recommendation systems, in *The Adaptive Web* (Springer, Berlin, Heidelberg, 2007), pp. 325–341
4. L. Chen, G. Chen, F. Wang, Recommender systems based on user reviews: the state of the art. User Model User-Adapt. Interact. **25**(2), 99–154 (2015)
5. C.C. Musat, Y. Liang, B. Faltings, Recommendation using textual opinions, in *Twenty-Third International Joint Conference on Artificial Intelligence*, June 2013
6. S.G. Esparza, M.P. O'Mahony, B. Smyth, Effective product recommendation using the real-time web, in *International Conference on Innovative Techniques and Applications of Artificial Intelligence.* (Springer, London, December 2010), pp. 5–18
7. Š. Pero, T. Horváth, Opinion-driven matrix factorization for rating prediction, in *International Conference on User Modeling, Adaptation, and Personalization.* (Springer, Berlin, Heidelberg, June 2013), pp. 1–13
8. Y. Moshfeghi, B. Piwowarski, J.M. Jose, Handling data sparsity in collaborative filtering using emotion and semantic based features, in *Proceedings of the 34th International ACM Sigir Conference on Research and Development in Information Retrieval.* (ACM, July 2011), pp. 625–634
9. S. Chakraverty, M. Saraswat, Review based emotion profiles for cross domain recommendation. Multimed. Tools Appl. **76**(24), 25827–25850 (2017)
10. A. Singhal, P. Sinha, R. Pant, Use of deep learning in modern recommendation system: a summary of recent works, in *Proceedings of the 2017 International Journal of Computer Applications* (0975–8887), vol. 180, no. 7 (2017)
11 T. Mikolov, M. Karafiát, L. Burget, J. Černocký, S. Khudanpur, Recurrent neural network based language model, in *Eleventh Annual Conference of the International Speech Communication Association* (2010)
12. Y. Koren, R. Bell, C. Volinsky, Matrix factorization techniques for recommender systems. Computer **8**, 30–37 (2009)
13. http://www.findmeanauthor.com. Accessed Jan 2021
14. J. Pennington, R. Socher, C. D. Manning, Glove: global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), (2014, October), pp. 1532–1543
15. P. Patel, D. Patel, C. Naik, Sentiment analysis on movie review using deep learning RNN method, in: *Intelligent Data Engineering and Analytics* ed. by S. Satapathy, Y.D. Zhang, V. Bhateja, R. Majhi. Advances in Intelligent Systems and Computing, vol. 1177 (Springer, Singapore, 2021). https://doi.org/10.1007/978-981-15-5679-1_15
16. L. Huang, M. Fu, F. Li, H. Qu, Y. Liu, W. Chen, A deep reinforcement learning based long-term recommender system. Knowl.-Based Syst. **213**, 106706 (2021)
17. J. Kim, J. Kim, H.L.T. Thu, H. Kim, Long short term memory recurrent neural network classifier for intrusion detection, in *Proceedings of the 2016 International Conference on Platform Technology and Service (PlatCon)*

A Survey on Applications of Unmanned Aerial Vehicles (UAVs)

Ritu Dewan and Khandakar Faridar Rahman

Abstract The demand and need of Unmanned Aerial Vehicles is rising day by day. The Unmanned Aerial Vehicles are used in various areas including, search and rescue, delivery of goods, remote sensing, monitoring of traffic, surveillance in the construction and infrastructure. In the past pandemic situation of COVID-19, Unmanned Aerial Vehicles played an important role by delivering of medicines and essential goods to the needy people. It also helped to check for the congestion in a particular area. We present the key areas where Unmanned Aerial Vehicles can be used and can be an important asset in the future.

Keywords Unmanned Aerial Vehicles · Search and rescue · Transportation of goods · Surveillance · Monitoring of traffic · MANET · VANET

1 Introduction

Unmanned Aerial Vehicle can be applied in many areas of science because it can be easily deployed, cost of maintenance is low, mobility and ability to hover is high [1]. Unmanned Aerial Vehicles are being used for delivery of goods, traffic monitoring in real time, providing Wi-Fi capabilities in remote areas, remote sensing, search and rescue missions, precision agriculture, and civil infrastructure inspection.

Hayat et al. [1] presented the qualities and necessities of networks of Unmanned Aerial Vehicles for common usage areas during the duration 2000–2015. They overview the nature of administration necessities, network-pertinent mission boundaries, information prerequisites, and the base information to be communicated throughout the organization for common uses. It also examines systems administration-related prerequisites, for example, network, versatility, safety, protection, security, and adaptability. At last, they present exploratory outcomes from numerous activities and research the appropriateness of existing interchanges advances to help solid flying organizations. In [2], they endeavors to research in the regions of steering, consistent handover and energy effectiveness. To start with,

R. Dewan (✉) · K. F. Rahman
Banasthali Vidyapith, Vanasthali, Rajasthan, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_8

they recognize foundation and impromptu UAV organizations, application zones to which Unmanned Aerial Vehicles go about like workers or like customers, star or work UAV organizations and whether the sending is solidified against deferrals and interruptions. At that point, they center around the principle issues of steering, consistent handover, and energy productivity in UAV organizations.

FANETs are the Adhoc networks associated with the Unmanned Aerial Vehicles. In [3] initially explained the contrasts between MANET, Flying Adhoc Networks and VANET. They present the fundamental Flying Adhoc Networks configuration issues and talk about open examination challenges. In [3] the author gives a detailed survey on Unmanned Aerial Vehicles, with their potential use in providing Internet of Things services from the sky. This paper [4], gives an exhaustive report on the utilization of Unmanned Aerial Vehicles in remote organizations. They examine two fundamental use instances of Unmanned Aerial Vehicles; in particular, aeronautical base stations and cell associated clients. For each utilization instance of Unmanned Aerial Vehicles, they present key difficulties, applications, and principal open issues. Also, they depict numerical ways and procedures needed for meeting Unmanned Aerial Vehicles challenges like going for Unmanned Aerial Vehicles in remote organizations.

In [6], the creators give a diagram of UAV-aided remote interchanges by presenting the essential organization engineering and fundamental channel qualities. They feature the key plan contemplations just as the new chances to be investigated.

In [4] authors present a review of inheritance and developing user security technologies along with the range designation of security use over all the groups in the US. They presume that the utilization of Unmanned Aerial Vehicles on the side of user communications are covered by protection issues and absence of comprehensive strategies, guidelines, and administration for Unmanned Aerial Vehicles. In [5], they reviewed the applications of Unmanned Aerial Vehicles in the different ways like tracking, object detection method, and general purpose distributed processing applications. In [7], the authors presented the classification also includes surveillance, data collection, path planning, navigation, collision avoidance, coordination, environmental monitoring. Be it may be that, this overview doesn't consider the difficulties confronting Unmanned Aerial Vehicles in the areas and the possible part of upcoming innovations in Unmanned Aerial Vehicles employments.

In [9] they give an exhaustive study on accessible Air to Ground channel estimation missions, big and trivial-scale fading channel models, their impediments, and subsequent scope for Unmanned Aerial Vehicles in corresponding situations. In [10], it gives a review on the estimation techniques for UAV channel utilizing low altitudes and considers different channel allocation. They likewise audit the contemporary viewpoint of Unmanned Aerial Vehicles showcasing ways and show some future challenges in this area.

Unmanned Aerial Vehicles are extended to be an efficient deliverer of common ways in numerous fields like cultivating, surveillance of traffic, and search and rescue. In this paper, we survey a few Unmanned Aerial Vehicles common applications. The motivation to embrace this overview is the absence of a study in on these issues.

Unmanned Aerial Vehicles present a key role in today's market. The market payload covers all hardware which are conveyed by Unmanned Aerial Vehicles, for example, cameras, sensors, radars, LIDARs correspondences gear. Additionally the market estimation of Unmanned Aerial Vehicles employments.

- It provides a grouping of Unmanned Aerial Vehicles dependent on Unmanned Aerial Vehicles continuance, most fuel type, extreme height, fuel type, payload, and applications.
- The Unmanned Aerial Vehicles common applications shrouded in this review include: constant checking of road traffic, giving remote inclusion, distant detection, search and rescue, conveyance of products, Precision agriculture, and remote sensing.

Unmanned Aerial Vehicles are used for giving wireless coverage in inaccessible areas, for example, Facebook's Aquila Unmanned Aerial Vehicles [11]. In this application, Unmanned Aerial Vehicles need to come back to the ground station for charging, because of their battery limit. To solve this, solar boards were mounted on Unmanned Aerial Vehicles where electrical energy was generated from the solar energy for longer flight journeys [12]. Laser power is the forthcoming innovation to charging even around night time when sun-based energy isn't possible or when it is insignificant during chilly winter. This would provide such Unmanned Aerial Vehicles capability to have longer flights months without landing [13].

1.1 Applications of Unmanned Aerial Vehicles

1.1.1 Search and Rescue (SAR)

In the era of new scientific developments, future capability of UAV has been developed for public and civil applications. Unmanned Aerial Vehicles are considered to be of great benefit in these areas, particularly for public safety, disaster management, and search and rescue operations. If there occurs any natural calamity or human-made calamities like floods, forest fires, or attacks by the terrorists, essential infrastructure including food, water utilities, electricity, transportation, and communication systems are partially or completely influenced by the calamity. This initiates faster support of rescue operations [1]. When the telecommunications is suspended, warnings related to the disaster and various other instructions for increasing the speed of rescue and recovery operations can be provided through Unmanned Aerial Vehicles timely. Medical stuff to affected areas that are targeted as not reachable can also be carried out by Unmanned Aerial Vehicles as shown in Fig. 1 [53]. In certain disastrous situations like Tsunami, forest fires, avalanches, and search scavenging for lost people, Unmanned Aerial Vehicles are utilized to play the key role to speed up the efficiency of SAR operation [15]. Also, it helps to cover larger extent w/o endangering the safety of the person comprised.

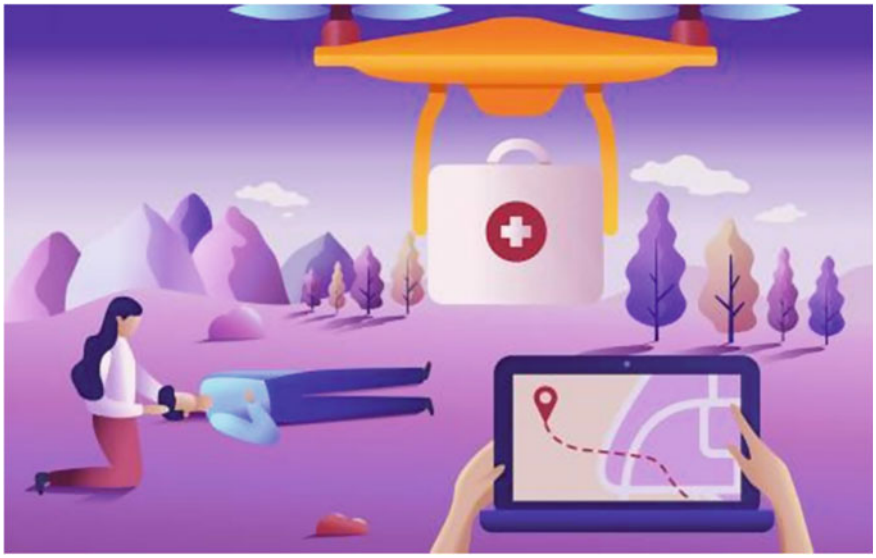


Fig. 1 Medical facilities delivered using drones

SAR operations are now one of the areas of interest for research. UAVs are used for SAR missions with following steps:

(1) Good Quality photos and video streams are captured using on-board cameras during survey of a specified area. The photos and videos help to evaluate the amount of damage that has occurred in that area. And on the basis of the data received, rescue teams can easily find the search area and proceed with the search operations [16].
(2) Lightweight quadrotor Unmanned Aerial Vehicles fitted with GPS is the prototype was developed to help in finding the missing people in the Alcedo project [17]. In snow avalanche scenarios Unmanned Aerial Vehicles were used to support SAR operations. In capstone project Geographic Information System and Thermal infrared imaging was used by UAVs to find the missing people [18].
(3) Delivery of medicine, food and water to the injured can also be done using UAV. In [19], Unmanned Aerial Vehicle having potential VTOL capability was designed.
(4) Unmanned Aerial Vehicle help to provide communication in the target area after the disaster.

IMAGE PROCESSING IN SAR

Unmanned Aerial Vehicles (UAVs) use techniques of image processing to find the targeted objects. With the help of aerial images, we can know the location of the target object. For the detection of targets Various thermal and vision cameras are used. The heat profile of missing people can be recovered through Thermal cameras (e.g., IR

cameras). Similarly, Vision cameras are also used in the detection of target objects and people. Many methods and algorithms are used which use both the cameras.

1.1.2 Remote Sensing

Unmanned Aerial Vehicle's captured data can be collected from ground sensors can be used to collect data through ground sensors and give the available data to ground base stations [20]. Unmanned Aerial Vehicles have sensors which are used for monitoring the environment and disaster management. Various datasets emerging from Unmanned Aerial Vehicles remote sensing is helping the research teams to work for various applications like drought monitoring, water quality monitoring, crop monitoring, yield estimation, tree species, and disease detection, etc. [21]. There are two kinds of Remote Sensing System: passive and active [22]. The origin of energy to find various objects, sensors are accountable in active remote sensing system.

The sensor communicates radiations toward the target to be investigated, at that point where the sensor identifies and computes the amount of energy which is getting reflected via object. Remote system which is active in nature is used for propagation in the atmosphere [22]. Remote systems which are active in nature have radar, LiDAR, radar, scatter meter, and sounder.

Active Remote Sensing System is different from passive remote sensing system because of radiations that are produced or are reflected by the object in Active Remote Sensing System. Most of passive sensors work in the visible, infrared, thermal infrared, and microwave parts of the electromagnetic range [22]. In passive remote sensing systems, there are various devices like accelerometer, hyper-spectral radiometer, imaging radiometer, radiometer, sounder, spectrometer, and spectroradiometer.

LiDAR and radar are the examples of active sensors. Laser beam is put on the earth's surface and calculate the distance with the help of time between incoming and outgoing light pulses in a LiDAR sensor 0.2 D image of the surface is produced using a radar sensor by finding the range of reflected energy from various objects. Spectrometer is the most common passive sensor which helps in the detection, measurement, and analysis of spectral data of incident electromagnetic radiations.

IMAGE PROCESSING IN REMOTE SENSING SYSTEM

In the image processing for a remotely sensed networks, they get the images captured by Unmanned Aerial Vehicles and find the position and direction of image. After this to get the location of the object, aerial triangulation algorithms are applied which tries to find the exact location and orientations of the images from the above. A lot of tie points are generated which occur across multiple images. With the help of least square adjustment method. These tie points are then used by bundle-block adjustment algorithm to find the object's position by getting a no. redundant values and applying least square adjustments.

1.1.3 Construction and Inspection

Nearly 45% of UAV market is covered by the various applications of Construction and infrastructure inspection [23, 24] like monitoring of construction projects and inspection of power lines and gas pipelines.

Unmanned Aerial Vehicles are utilized for monitoring of construction project sites and also in construction and infrastructure inspection applications [25]. Also, the site can be accessed by the project managers with the help of Unmanned Aerial Vehicles with better vision without visiting the site.

In the power transmission line's inspection with high voltage, Unmanned Aerial Vehicles can also be utilized. Unmanned Aerial Vehicles are utilized for detection, inspection, and diagnosis of defects in the power line infrastructure.

Authors in [26] designed a full autonomous Unmanned Aerial Vehicles system for inspection of power line. Moreover, Unmanned Aerial Vehicles are used to capture images of the regions close to the power lines, it also calculates the distance between the adjoining trees, buildings, and power lines. Sometimes with the help of small-UAV (Unmanned Aerial Vehicles) equipped with a gas controller unit it detects air and gas content. The system had a remote sensing to detect gas leaks in oil and gas pipelines.

IMAGE PROCESSING AND CONSTRUCTION IN INFRASTRUCTURE INSPECTION

For Unmanned Aerial Vehicles in Construction and infrastructure inspection there occurs cameras and sensors mounted on it. Image Processing helps to monitor and assess the projects that go on construction sites along with the checking of the infrastructure like survey of construction sites, monitoring work progress, detection of damage, inspection of highways, in the construction and degradation in the surfaces.

There occur different methods for inspections of and real-time structural health monitoring and infrastructures. After this GCS will receive the real-time data. Then the image processing will be done by the units present in the GCS. The authors also proposed a mechanism for crack detection, they merge Hat transforms and HSV thresholding. With the help of color camera, TIR camera, and transmitter, captured images are sent to the GCS which are thereby used for temperature monitoring in power lines.

1.1.4 Precision Agriculture

In precision agriculture (PA) Unmanned Aerial Vehicles can be utilized for crop management and monitoring [27, 28], irrigation scheduling [30], weed detection, gathering data from ground sensors about moisture, soil properties, etc., [32], disease detection and pesticide spraying. The PA becomes economical and improved technology with the deployment of Unmanned Aerial Vehicles and helps to improve crop yields, productivity, and yield more profit in farming systems.

At low altitudes in order to grow small crops, Unmanned Aerial Vehicles can be deployed with more accuracy and less-cost in comparison to the traditional way. Moreover, High resolution images of crop can be captured by Unmanned Aerial Vehicles that are used for the management of crop and detection of variation of crop response to irrigation, management of weed, and reduction in the number of herbicides [27, 28]. The images which are captured by Unmanned Aerial Vehicle are used for the appropriate dispensation of soil moisture on the surface. It also helps to calculate the water stress index which helps to determine water deficient areas [30].

Unmanned Aerial Vehicles can also be used to check the crop height with the help of images obtained by the Unmanned Aerial Vehicles. However, weather played an important role as wind can harm the kind of images obtained by the Unmanned Aerial Vehicles [33].

Unmanned Aerial Vehicles can be used to check the various phases of crop diseases during different development of disease. It can be checked using Unmanned Aerial Vehicles, e.g., detection of soil-borne fungus can be detected using images captured in UAV. For example, aerial thermal images are utilized for detection of faster stage development of soil-borne fungus [34]. To quantify soil texture at a regional scale Unmanned Aerial Vehicle's thermal images can be used to measure the difference in surface temperature in a relatively homogeneous climatic condition [35]. These images captured are also used for Crop yield mapping with which they decide for cash flow budgeting, crop insurance, harvest planning, and storage requirements. Unmanned Aerial Vehicle images are utilized to estimate the yield and total biomass of rice crop in Thailand.

IMAGE PROCESSING IN PRECISION AGRICULTURE

Image Processing can be highly used to capture images of crops and farms. The crop yield can be predicted using the vegetation Indices (VI) generated by processing of the images. This VI also helps to detect diseases and is also used in management of weed. These Vis can also be used by the government to check for the vegetative cover of an area.

1.1.5 Delivery of Goods

Food, packages, and other goods [36–39] can be transported with the help of Unmanned Aerial Vehicles. Ambulance drones are used to deliver medicines, immunizations, and blood, to unreachable locations in the healthcare field. Medical instruments can be rapidly transported to a local after cardiac arrests. Live video streaming facility also allows doctors to remotely see and tell on-scene individuals about the usability of medical instruments. For the goods delivery quadrotor drones can be used like Fig. 2 [54]. Commercial Delivery Drones will be used in future for delivery of goods [40]. When the packet is delivered by an Unmanned Aerial Vehicle, the control processor checks the receiver of the packet. If the code sent by the sender and receiver matches then the packet is delivered. Else a message is sent to the receiver

Fig. 2 Drones used for delivery of goods

about non-delivery of the packet [41]. The request may contain the device identifier, or network address of the package docking device.

IMAGE PROCESSING IN DELIVERY OF GOODS

During the delivery of goods, whenever a packet is to be delivered it uses the GPS co-ordinates to reach the place and a QR code to identify the correct recipient of the packet. On its arrival at the GPS location the control processor matches the QR code if the code matches the device it transfers the packet and confirmation message is sent to the receiver [51].

1.1.6 Real-Time Monitoring of Road Traffic

Unmanned Aerial Vehicles are used in the automation of different things like tasks of field support teams, traffic police, road surveyors, and rescue teams.

Unmanned Aerial Vehicles is known as a new method to monitor traffic where it collects information about traffic on the road as depicted in Fig. 3 [55]. In comparison to the old monitoring devices like surveillance using video cameras and microwave sensors, Unmanned Aerial Vehicles can cover larger targeted areas in a cost-effective manner [42].

When certain calamity occurs it damages computing, communication setup or power supply. With such failures there occurs inability to check and collect data about

Fig. 3 UAV is used by police to catch traffic violators

the transportation system. A system was proposed to detect vehicle, it consisted of three basic steps. It firstly checks the pre-classification, with boosted classifier than with the detection of blob and then finally SVM doing classification.

In [43] Unmanned Aerial Vehicle cameras were used for detecting moving vehicles. They proposed a method to identify vehicles in real time using CNN. In the proposed method they applied CNN to identify vehicles more precisely and in real time. The method constitutes of few steps to identify moving vehicle, firstly Adjoining frames are matched. Then, it is decided whether these frame pixels are candidate frames or not. In the last candidate frames are trained with deep convolutional neural network to classify them into vehicles or background. 90% accuracy was attained during their evaluation using the dataset of CATEC UAV.

In this author introduced detection of vehicles tracking system used to collect images captured by UAV. Consecutive frames are used to get the information like the speed and location of the object. It comprises four stages image registration, image feature extraction, vehicle shape detection, and vehicle tracking. A method is proposed in [42] to collect data of speed, density, and flow of traffic as depicted in Fig. 4. The recommended method works in various steps: (1) A Estimation of multi-directional traffic flow parameters using aerial videos. (2) KLT tracker and k-means clustering methods were combined for vehicle detection and counting. (3) Real-time information of road traffic is collected. (4) Proposed method works in daylight and night light settings. It also is insensitive to activities of UAV (like vibration, change in speed, and hovering).

For the real-time detection and tracking of an object on road area in low and mid heighted place a system was proposed in [44]. This system is utilized for monitoring,

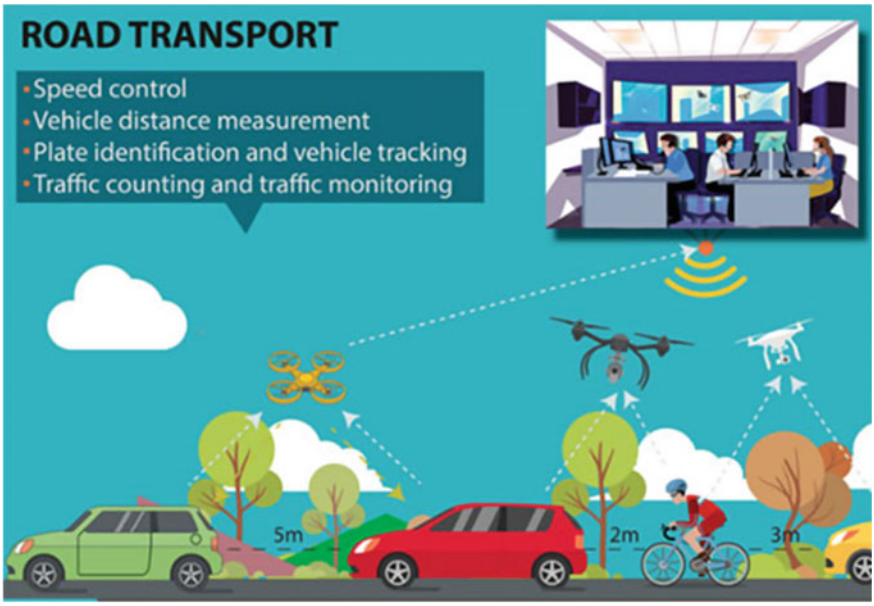


Fig. 4 Drones helping in the monitoring of traffic

inspection, autonomous navigation, surveillance of traffic, and monitoring. It takes the concept of Graph Cut algorithm for road detection as it provides systematic and well-organized performance in segmentation of 2-D color images.

In [44], the authors proposed a method where they had integrated video streams from Unmanned Aerial Vehicles to increase real-time monitoring and controlling of traffic. In [45] the authors proposed a mechanism for the security issues while monitoring road traffic using Unmanned Aerial Vehicles. The UAV is trained to check real-time traffic along with the security issues. It then accordingly acts for the traffic management control by re-routing the traffic.

For the use in traffic on road Kansas Department of Transportation (KDOT), in [46] authors represented the novel work where Unmanned Aerial Vehicles can be used for road traffic monitoring and controlling but they can't replace the existing system as the Unmanned Aerial Vehicle has battery issues so 24 h surveillance is not possible. Apeltauer et al. [47] proposed a method for detection of moving vehicles and tracking by capturing images through UAVs. Bor-Yaliniz et al. [53] proposed that image processing along with the vision algorithms are used for monitoring of traffic. With the images extracted investigation of accidents and damage can be assessed easily.

IMAGE PROCESSING REAL TIME MONITORING OF ROAD TRAFFIC

Detecting many vehicles at one time is a big challenge, algorithm is under process where with the help of video background images are extracted to find objects

and distinguish objects with their reflection. In the swarm's operations the image processing and collision avoidance together play a very major role. Therefore, to go for this resource like cameras, location sensors, gyroscopes, etc. are needed.

1.1.7 Surveillance Applications of UAV

Unmanned Aerial Vehicles help in surveillance as it covers a wider area. A multi-UAV coordination method was proposed by the authors of [48] under the AWARE project (distributed decision-making architecture used for multi-UAV coordination). Basically, in this there occurs different tasks which does a given task in a systematic manner according to a proper plan. These tasks are the small steps to achieve a big goal.

The authors in [49], did analysis for UAVs deployments in war-areas (like Pakistan, Iraq & Afghanistan), border-zones, and urban areas in the USA. This analysis was very helpful as it reduces the chances of human error from various geographical, political, and cultural backgrounds.

The use of UAVs for surveillance has a problem in exact target identification. However, it also focusses on the benefits of covering larger area and the implementation of using UAV for the purpose of surveillance as depicted in Fig. 5.

The authors in [50] sum up the influence of surveillance done by UAV in civil applications. This paper emphasizes the law enforcement policy for privacy through Unmanned Aerial Vehicles for surveillance.

IMAGE PROCESSING SURVELEINCE APPLICATIONS OF UAV

In surveillance using Unmanned Aerial Vehicles specially in military services where UAVs can check for any type of physical attack and also help to secretly patrol a given area. Also, during the course of any pandemic like COVID-19 UAVs can check for the congestion and amount of people in an area. The best feature is there can be nano UAVs which are small in size can be used for security purpose at a particular area.

The images captured by UAVs during the surveillance at a target area can be used for findings and can help to achieve a task.

1.1.8 Providing Wireless Coverage

Unmanned Aerial Vehicles are utilized in emergency problems for providing wireless coverage. In this UAV acts as an aerial wireless base station when there occurs no cellular network. They also help to increase the data speed and can also help in increasing the area covered by the base station. The authors in [8] presented various use cases of aerial wireless base stations where Unmanned Aerial Vehicles are used to provide wireless network at a specific location.

Unmanned Aerial Vehicles give very fast recovery in service whenever any calamity occurs [8]. In order to provide cellular network use of Fixed wing, Balloon

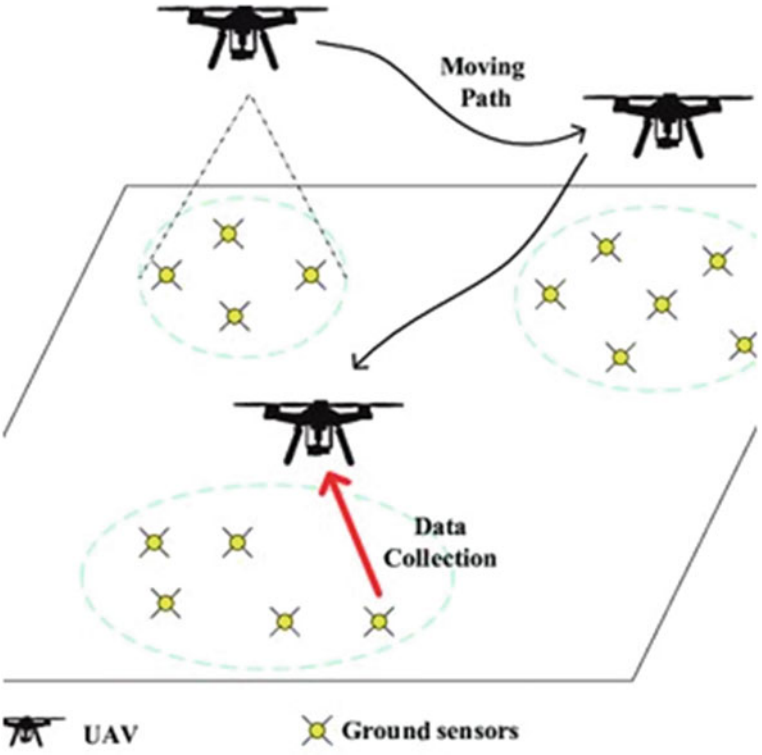


Fig. 5 UAVs for data collection

and VTOL Unmanned Aerial Vehicles can be done example Viking aircraft [33–35], Tethered Heli kite Balloon [30], and the FALCON Quadrotor [32]. In sparsely connected areas, Unmanned Aerial Vehicles are used as gateway nodes in order to give connectivity to the base station and it also helps to increase the speed of incoming and outgoing data packets. Unmanned Aerial Vehicles are also used to provide the indirect connectivity between the nodes where these nodes at as relay nodes. In [51], authors proposed a mechanism where the relay node is positioned for providing connectivity is on the Unmanned Aerial Vehicle. This results in maximizing the throughput by maximizing the relay trajectory and the allocation of source-relay. Wireless sensors are the best example in precision agriculture applications. This paper [53] focuses on the communication of UAVs with wireless networks. It discusses both the aspects firstly about the use of wireless networks for personal or professional UAVs and secondly about the performance of network when UAVs were using the network. For this the categorization was made into mobile-enabled drones (MEDs) and wireless infrastructure drones (WIDs).

2 Conclusion

The concept of Unmanned Aerial Vehicles is used in every field nowadays. From small application of delivery of goods to a major application of surveillance or monitoring of traffic, Unmanned Aerial Vehicles are used everywhere. In Search and Rescue operation Unmanned Aerial Vehicles helps in increasing the speed of up of rescue and recovery operations. In Remote Sensing, Unmanned Aerial Vehicles are used to share the data packets from ground sensors to ground base station. In Construction and Infrastructure Inspection, Unmanned Aerial Vehicles helps in the real-time monitoring of project site. Unmanned Aerial Vehicles are used for detection, inspection, and diagnoses of the defects of any infrastructure. In Delivery of Goods, Unmanned Aerial Vehicles help to deliver the goods be it be medicines, grocery, parcels, or medical instruments. In Real-Time Monitoring of Road Traffic, Unmanned Aerial Vehicles can help to automate the traffic system. These aerial devices help to monitor road traffic and collect information about road traffic and road accidents. In Surveillance, UAVs help to check for the enforcement of law in a particular area. The images captured by Unmanned Aerial Vehicles in the field are used to find various things like crop height, detection of disease, Gathering of soil information. UAVs also help to provide wireless coverage in the remote areas. Unmanned Aerial Vehicles when merged with image processing and 4G can help in real-time processing of data.

References

1. S. Hayat, E. Yanmaz, R. Muzaffar, Survey on unmanned aerial vehicle networks for civil applications: a communications viewpoint. IEEE Commun. Surv. Tuts. **18**(4), 2624–2661, 4th Quart. (2016)
2. L. Gupta, R. Jain, G. Vaszkun, Survey of important issues in UAV communication networks. IEEE Commun. Surv. Tuts. **18**(2), 1123–1152, 2nd Quart. (2016)
3. I. Bekmezci, O.K. Sahingoz, Ş Temel, Flying ad-hoc networks (FANETs): a survey. Ad Hoc Netw. **11**(3), 1254–1270 (2013)
4. A. Kumbhar, F. Koohifar, I. Güvenç, B. Mueller, A survey on legacy and emerging technologies for public safety communications. IEEE Commun. Surv. Tuts. **19**(1), 97–124, 1st Quart. (2017)
5. G. Chmaj, H. Selvaraj, Distributed processing applications for UAV/drones: a survey, in *Progress in Systems Engineering* (Springer, 2015), pp. 449–454
6. Y. Zeng, R. Zhang, T.J. Lim, Wireless communications with unmanned aerial vehicles: opportunities and challenges. IEEE Commun. Mag. **54**(5), 36–42 (2016)
7. N.H. Motlagh, T. Taleb, O. Arouk, Low-altitude unmanned aerial vehicles-based Internet of Things services: comprehensive survey and future perspectives. IEEE Internet Things J. **3**(6), 899–922 (2016)
8. M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, M. Debbah, A tutorial on unmanned aerial vehicles for wireless networks: applications, challenges, and open problems (2018), https://arxiv.org/abs/1803.00680
9. W. Khawaja, I. Guvenc, D. Matolak, U.-C. Fiebig, N. Schneckenberger, A survey of air-to-ground propagation channel modelling for unmanned aerial vehicles (2018), https://arxiv.org/abs/1801.01656
10. A.A. Khuwaja, Y. Chen, N. Zhao, M.-S. louini, P. Dobbins, A survey of channel modelling for UAV communications. IEEE Commun. Surv. Tuts., 20(4), 2804–2821, 4th Quart. (2018)

11. M. Zuckerberg, *Connecting the World from the Sky*. Facebook, Cambridge, MA, USA, Technical Report 249 (2014)
12. Y. Sun, D.W.K. Ng, D. Xu, L. Dai, R. Schober, Resource allocation for solar powered UAV communication systems, in *Proceedings of IEEE 19th International Workshop Signal Process. Advances Wireless Communication*, June 2018, pp. 1–5
13. NASA Armstrong Fact Sheet (2014) *Beamed Laser Power for Unmanned Aerial Vehicles*, http://www.nasa.gov/centers/armstrong/news/Factsheets/FS-087-DFRC.html
14. T. Kelly, *The Booming Demand for Commercial Drone Pilots*, https://www.theatlantic.com/technology/archive/2017/01/drone-pilot-school/515022. Accessed Feb 2018
15. M. Silvagni, A. Tonoli, E. Zenerino, M. Chiaberge, Multipurpose UAV for search and rescue operations in mountain avalanche events. Geomat. Nat. Haz. Risk **8**(1), 18–33 (2017)
16. M.A.R. Estrada, How unmanned aerial vehicles–UAV's–(or Drones) can help in case of natural disasters response and humanitarian relief aid? Technical Report (2017)
17. *Alcedo*, http://www.alcedo.ethz.ch/. Accessed Feb 28, 2018
18. J. Joern, Examining the use of unmanned aerial systems and thermal infrared imaging for search and rescue efforts beneath snowpack. Ph.D. dissertation, University of Denver, Denver, CO, USA (2015)
19. D. Jo, Y. Kwon, Development of rescue material transport UAV (unmanned aerial vehicle). World J. Eng. Technol. **5**(4), 720 (2017)
20. E. Tuyishimire, A. Bagula, S. Rekhis, N. Boudriga, Cooperative data muling from ground sensors to base stations using unmanned aerial vehicles, in *Proceedings of IEEE Symposium Computer Communication (ISCC)*, July 2017, pp. 35–41
21. Vito, *Datasets and Products for a Broad Range of Applications*, https://remotesensing.vito.be/data-products-services/using-remote-sensing-drones. Accessed Dec 2017
22. NASA, *Remote Sensors*, https://earthdata.nasa.gov/user-resources/remote-sensors. Accessed Dec 2017
23. C. Deng, S. Wang, Z. Huang, Z. Tan, J. Liu, Unmanned aerial vehicles for power line inspection: a cooperative way in platforms and communications. J. Commun. **9**(9), 687–692 (2014)
24. M. Gheisari, J. Irizarry, B.N. Walker, UAS4SAFETY: the potential of unmanned aerial systems for construction safety applications, in *Proceedings of Construction Research Congress, Construction Global Network* (2014), pp. 1801–1810
25. C. Sampedro, C. Martinez, A. Chauhan, P. Campoy, A supervised approach to electric tower detection and classification for power line inspection, in *Proceedings of International Joint Conference Neural Network (IJCNN)*, July 2014, pp. 1970–1977
26. J.I. Larrauri, G. Sorrosal, M. González, Automatic system for overhead power line inspection using an Unmanned Aerial Vehicle— RELIFO project, in *Proceedings of International Conference Unmanned Aircraft Systems (ICUAS)*, May 2013, pp. 244–252
27. Y. Huang, S.J. Thomson, W.C. Hoffmann, Y. Lan, B.K. Fritz, Development and prospect of unmanned aerial vehicle technologies for agricultural production management. Int. J. Agricult. Biol. Eng. **6**(3), 1–10 (2013)
28. N. Muchiri, S. Kimathi, A review of applications and potential applications of UAV, in *Proceedings of Sustainable Responsible Innovation Conference* (2016), pp. 280–283
29. W. Kazmi, M. Bisgaard, F.J. Garcia-Ruiz, K.D. Hansen, A. la Cour-Harbo, Adaptive surveying and early treatment of crops with a team of autonomous vehicles, in *Proceedings of ECMR* (2011), pp. 253–258
30. V. Gonzalez-Dugo, P. Zarco-Tejada, E. Nicolás, P. Nortes, J. Alarcón, D. Intrigliolo, E. Fereres, Using high resolution UAV thermal imagery to assess the variability in the water status of five fruit tree species within a commercial orchard. Precis. Agricult. **14**(6), 660–678 (2013)
31. F. Garcia-Ruiz, S. Sankaran, J.M. Maja, W.S. Lee, J. Rasmussen, R. Ehsani, Comparison of two aerial imaging platforms for identification of Huanglongbing-infected citrus trees. Comput. Electron. Agricult. **91**, 106–115 (2013)
32. P. Mathur, R.H. Nielsen, N.R. Prasad, R. Prasad, Data collection using miniature aerial vehicles in wireless sensor networks. IET Wirel. Sens. Syst. **6**(1), 17–25 (2016)

33. D. Belton, P. Helmholz, J. Long et al., Crop height monitoring using a consumer-grade camera and UAV technology. PFG **87**, 249–262 (2019). https://doi.org/10.1007/s41064-019-00087-8
34. R. Calderón, J.A. Navas-Cortés, C. Lucena, P.J. Zarco-Tejada, High-resolution airborne hyperspectral and thermal imagery for early detection of Verticillium wilt of olive using fluorescence, temperature and narrow-band spectral indices. Remote Sens. Environ. **139**, 231–245 (2013)
35. D.-C. Wang, G.-L. Zhang, X.-Z. Pan, Y.-G. Zhao, M.-S. Zhao, D.-C. Wang et al., Retrieval and mapping of soil texture based on land surface diurnal temperature range data from MODIS. PLoS One **10**(6), Art. no. e0129977 (2015)
36. G.N. Fandetti, Method of drone delivery using aircraft. U.S. Patent 14 817 356, February 9, 2017
37. PwC, *Besides Blockchain, What's Missing From IoT?* http://usblogs.pwc.com/emerging-technology/. Accessed Dec. 2017
38. The Future of Distribution, https://redstagfulfillment.com/the-future-of-distribution/. Accessed Dec. 2017
39. The Future of Distribution—Part II: Product Distribution in Emerging Markets, https://redstagfulfillment.com. Accessed Dec. 2017
40. Unmanned Cargo, *Drones Going Postal a Summary of Postal Service Delivery Drone Trials.* http://unmannedcargo.org/drones-going-postal-summary-postal-service-delivery-drone-trials/. Accessed Dec. 2017
41. G. Hoareau, J.J. Liebenberg, J.G. Musial, T.R. Whitman, Package transport by unmanned aerial vehicles. U.S. Patent 9 731 821, August 15, 2017
42. R. Ke, Z. Li, S. Kim, J. Ash, Z. Cui, Y. Wang, Real-time bidirectional traffic flow parameter estimation from aerial videos. IEEE Trans. Intell. Transp. Syst. **18**(4), 890–901 (2017)
43. Y. Qu, L. Jiang, X. Guo, Moving vehicle detection with convolutional networks in UAV videos, in *Proceedings of 2nd International Conference, Control, Automation Robotices (ICCAR)*, April 2016, pp. 225–229
44. H. Zhou, H. Kong, L. Wei, D. Creighton, S. Nahavandi, Efficient road detection and tracking for unmanned aerial vehicle. IEEE Trans. Intell. Transp. Syst. **16**(1), 297–309 (2015)
45. R. Reshma, T. Ramesh, P. Sathish Kumar, Security situational aware intelligent road traffic monitoring using unmanned aerial vehicles, in *Proceedings of International Conferences VLSI Systems, Architecture, Technology and Applications (VLSI-SATA)*, January 2016, pp. 1–6
46. P.A.R. Melissa McGuire, M. Rys, A study of how unmanned aircraft systems can support the Kansas Department of Transportation's efforts to improve efficiency, safety, and cost reduction. Kansas State University Transportation Center, Manhattan, KS, USA, Technical Report, August 2016
47. J. Apeltauer, A. Babinec, D. Herman, T. Apeltauer, Automatic vehicle trajectory extraction for traffic analysis from aerial video data. Int. Arch. Photogram. Remote Sens. Spat. Inf. Sci. **40**(3), 9 (2015)
48. I. Maza, F. Caballero, J. Capitán, J.R. Martínez-de-Dios, A. Ollero, Experimental results in multi-UAV coordination for disaster management and civil security applications. J. Intell. Robot. Syst. **61**(1), 563–585 (2011)
49. T. Wall, T. Monahan, Surveillance and violence from afar: the politics of drones and liminal security-scapes. Theor. Criminol. **15**(3), 239–254 (2011)
50. R.L. Finn, D. Wright, Unmanned aircraft systems: surveillance, ethics and privacy in civil applications. Comput. Law Secur. Rev. **28**(2), 184–194 (2012)
51. Y. Zeng et al., Throughput maximization for UAV-enabled mobile relaying systems. IEEE Trans. Commun. **64**(12), 4983–4996 (2016)
52. F. Outay, H.A. Mengash, M. Adnan, Applications of unmanned aerial vehicle (UAV) in road safety, traffic and highway infrastructure management: recent advances and challenges. Transp. Res. Part A: Policy Pract. **141**, 116–129 (2020)
53. I. Bor-Yaliniz, M. Salem, G. Senarath, H. Yanikomeroglu, Is 5G ready for drones: a look into contemporary and prospective wireless networks from a standardization perspective. IEEE Wirel. Commun. **26**(1), 18–27 (2019). https://doi.org/10.1109/MWC.2018.1800229
54. https://builtin.com/drones

55. https://www.alamy.com/people-with-quadrupter-sending-and-receiving-goods-drone-delivery-service-vector-3d-isometric-concept-illustration-of-drone-technology-delivery-container-image354454977.html, http://labyrinth2020.eu/the-project/
56. F. Author, S. Author, Title of a proceedings paper, in *Conference 2016, LNCS*, vol. 9999, ed. by F. Editor, S. Editor (Springer, Heidelberg, 2016), pp. 1–13
57. F. Author, S. Author, T. Author, *Book title*, 2nd edn. (Publisher, Location, 1999)
58. Author, F.: Contribution title, in: *9th International Proceedings on Proceedings* (Publisher, Location, 2010), pp. 1–2
59. LNCS Homepage, http://www.springer.com/lncs. Accessed Nov 21, 2016

Early Detection of Influenza Using Machine Learning Techniques

Sajal Maheshwari, Anushka Sharma, Ranjan Kumar, and Pratyush

Abstract Influenza is an infectious disease that rapidly affects all living beings and the effect of this outbreak can be seen as infection in thousands of people yearly across the world. For example, the Spanish Flu is detected as one of the most devastating influenza outbreak in 1918. Recently, the usage of machine learning techniques has been increased in medical research fields which include early diagnosis of a disease, pathology, and classification of various diseases. In this study, the dataset is extracted Human Surveillance records from the Influenza Research Database. Initially, 15,651 records extracted over the period of 2006–2017 for data pre-processing and then 9548 records have been selected for further analysis. The next step follows the classification and analysis of data based on the four different machine learning methods, support vector machines, K-nearest neighbors, artificial neural networks, and random forest. Their performances are evaluated in terms of sensitivity, specificity, and accuracy. It is evident from the experiments that random forest is one of the best machine learning techniques for early detection of the influenza.

Keywords Artificial neural network · Classification · Influenza · K-NN · Random forest · Sensitivity · Specificity · SVM

1 Introduction

Influenza is an infectious disease caused by RNA viruses from the Orthomyxoviridae family which infects the respiratory tract of animals, birds, and humans. It is found that usually Type A, Type B, and Type C of the influenza viruses affect human beings, while the Type D influenza virus too has the potential to affect human beings. Influenza spreads around the world in yearly outbreaks, resulting in about three to five million cases of severe illness and causing about 290,000–650,000 deaths. "Spanish flu" resulted in 15–20 million deaths [1]. "Asian Influenza" caused 115,700 deaths and the "Hong Kong Influenza" pandemic result 112,000 deaths [2]. In 1997, some other influenza strains like H5N1 recorded 18 cases which caused 6 deaths in Hong

S. Maheshwari (✉) · A. Sharma · R. Kumar · Pratyush
Department of Computer Science, Aryabhatta College, University of Delhi, Delhi, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_9

Kong. Later, in 2003, it reappeared in Hong Kong and mainland China, causing great loss to poultry farms and other bird species. In 2003, another strain, H7N7, appeared in Holland with around 90 cases and 1 death in humans [3]. The 2009 H1N1 pandemic strain has around 30,000 cases reported across 74 countries by June 11, 2009 [4]. During the 2019–20 influenza seasons, influenza was associated with 405,000 hospitalizations and 22,000 deaths [5]. The rapid evolution and mutation of influenza virus has created a generation of vaccine-resistant and antiviral medication-resistant variants, thereby the influenza virus infection remains a major public health threat [6].

Machine learning techniques are playing a significant role in detecting, predicting, and classifying many medical issues [7–10]. Different machine learning techniques like Decision Trees (DT), Random Forests (RF), Artificial Neural Networks (ANN), Bayesian Networks (Bayes Network), Support Vector Machines (SVMs), and Gaussian Processes (GPs) have been used for many medical purposes [11–13]. Influenza virus can be predicted in a patient by implementing machine learning techniques on the patient's dataset which includes the symptoms.

Many sources provide access to the Influenza Human Surveillance dataset. IRD's (Influenza Research Database) Influenza Human Surveillance Record [14] consists of 15,561 records out of which 9,611 records possess information regarding the Flu Test Status. There are several other databases available such as Health Data [15], Data.World [16], WHO [17], and CDC [18]. IRD's Influenza Human Surveillance Record has been selected over these datasets as it has many attributes like symptoms, vaccination status, pre-medical conditions, diagnosis status, post-medication status, and many other useful clinical attributes many of which are not present in the other databases.

Initially, four different machine learning techniques, viz., random forests, neural networks, support vector machine with linear kernel, and k-NN algorithm are used for predictive analysis of the influenza test status of the patient. This knowledge can be used by the physician/healthcare worker to diagnose the patient more accurately and economically. Although some Point-of-Care (POC) test like RIDTs (rapid influenza diagnostic tests), DIAs, or NAATs (nucleic acid amplification tests) tests give results in <15 min with high sensitivity percentage [19], recently it has been observed that it is not enough to predict flu with high accuracy. This is because certain impediments which are economic, regulatory, and policy related, as well as due to user's perceptions and other cultural barriers [20], prevent the use of many highly accurate POC tests.

2 Materials and Methods

Figure 1 describes the schematics of the methodology followed by the study. The data for this study is taken from Influenza Research Database (IRD) [14]. After pre-processing the data, symptoms and other clinical data are the main attributes. We

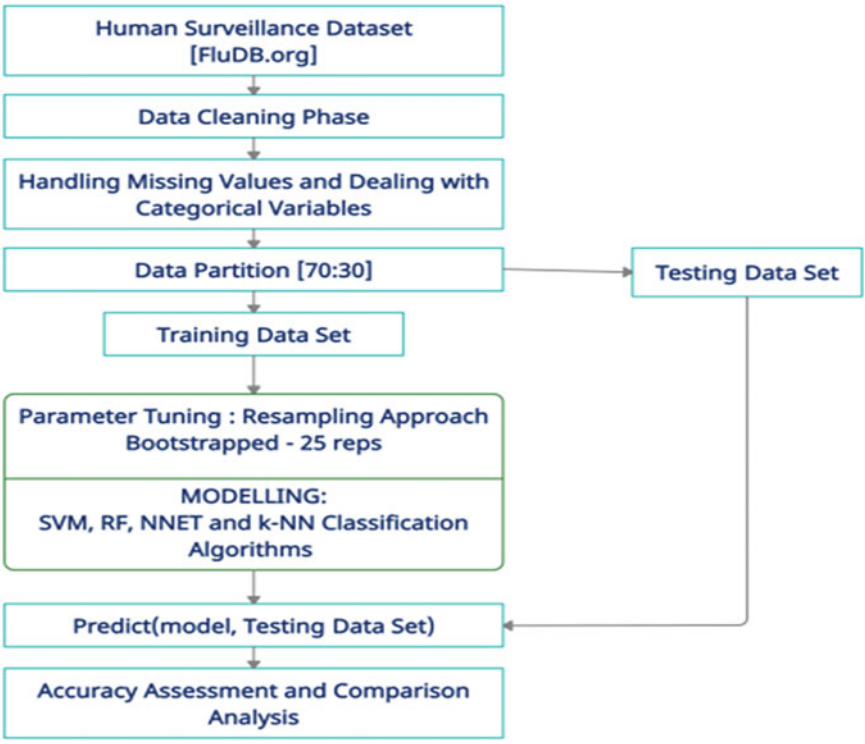


Fig. 1 Flowchart of the study

have implemented four different machine learning techniques for predictive analysis on the pre-processed data.

2.1 Study Area

A stepwise logistic regression was performed by Monto et al. [21] to determine the basic symptoms and patient characteristics which predicts an influenza infection. Their study determined that the best predictors of influenza infection were cough and fever, having a positive predictive value of 79% and patients with influenza like illness having both cough and fever within 48 h of symptom onset are likely to have influenza. Further, in the paper by Marquez and Barrón [22], the goal was to create an artificial intelligence system to support a diagnosis of influenza on the basis of a data of the Mexican population wherein relevant factors such as clinical symptoms of influenza, patient's clinical record, and other factors were utilized. It has been also determined in [23] that machine learning methods perform better than the expert-built Bayesian model for the detection of infectious diseases such as influenza. Proposed

Fig. 2 Country-wise distribution of records in dataset

MSDII-FFNN, a feed-forward neural network-based model, predicted with the 90% accuracy in predicting the influenza pandemic using different parameters [24].

The present study is based on influenza's database of the Human Surveillance Records accessed from the Influenza Research Database website [14]. This database contains 15,651 records of 22 different health institutions of different countries. For the influenza research community, the source has been very helpful in terms of projects involving sequence and structure analysis, comparative genomics, and virus phenotype studies among others. However, among all the projects, only 38 projects/papers are based on machine learning prediction, and out of these projects, there is no paper/project which has an objective similar to the present investigation. Figure 2 shows the distribution of records/samples of patients in different countries in the data used in the present study.

2.2 Data Used

Influenza Human Surveillance Data acquired for the years 2006–2017 was downloaded from the Influenza Research Database website [14]. Initially, the dataset consisted of 58 attributes with 15,651 records. Figure 3 shows the year-wise and gender-wise distribution of the records in the data.

Certain columns/attributes related to the drug resistance, proteins sequence, strain name, contact email address, pre-visit medications, post-visit medications, test sample type, disease outcome, miscellaneous comments, ARDS, and hospitalization duration were removed in the beginning itself as these attributes were not

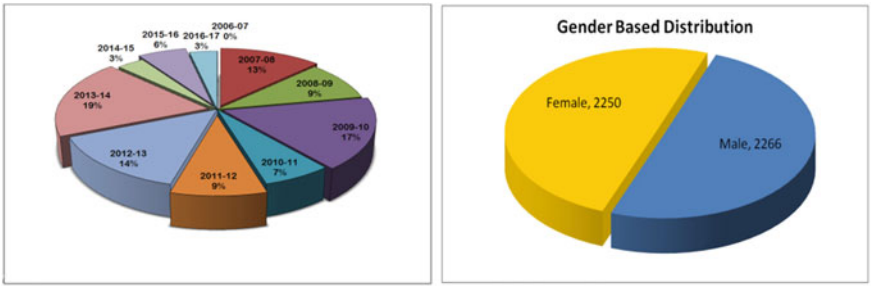


Fig. 3 **(Left)** Year-wise distribution. **(Right)** Gender-wise distribution of "Positive" class

relevant for our work. For our work, we are primarily interested in the symptoms, vaccination status, and medical conditions of the patient for the purpose of early prediction of the flu in the patient. The total relevant records for the positive flu cases in the data are almost equally distributed gender-wise, i.e., female (48.74%) and male (51.26%) approximately. However, in the data the distribution based on vaccination status, the records with "Not Vaccinated" status (87.28%) dominates heavily over "Vaccinated" (12.72%) records.

2.3 Data Pre-processing

In order to extract the relevant information for the analysis, several pre-processing techniques have been implemented on the dataset. Initially, "Flu test status" is considered the target variable in which the positive label (1) denotes "Positive to influenza infection" while negative label (2) denotes "Negative to influenza infection". "Symptoms" attribute consists of 24 different symptoms with their state. For each record, 24 or less symptoms are provided. To identify each symptom individually for each record, we divide the "Symptoms" attribute in 24 "Symptom names" attributes that hold the binary values 0 (for "No" or "empty" values) and 1(for "Yes"). Table 1 shows the distribution of the "Symptoms".

Initially, there are 71 unique combinations of medical diseases for records in data. To make it relevant for the data, four most repeating factors were identified and based on these a power set is defined over these 4 factors, dividing the factors in 16 unique combinations. Table 2 of medical condition defines the value substitution in the data.

Most of the missing values are handled by substituting a value of 0 for them. Here it is assumed that the "NA" values means that either that parameter is not known while collecting the data or since that value had no positive response for the particular patient record, i.e., left empty by the data manager of the source. We cannot remove or substitute these values with any other positive value. Hence, we directly replaced these values with 0 in our data. In order to handle 63 records with unknown values of gender, record is removed from the data because values assigned to "Gender" are not

Table 1 Symptoms distribution

Symptom name	Values
Running nose	0, 1
Cough	0, 1
Myalgia	0, 1
Headache	0, 1
Throat	0, 1
Fever	0, 1
Fatigue	0, 1
Temperature	0, 1
Diarrhea	0, 1
Nausea	0, 1
Sudden onset	0, 1
Short breath	0, 1
Malaise	0, 1
Wheezing	0, 1
Chills	0, 1
Vomiting	0, 1
Loss of appetite	0, 1
Arthralgia	0, 1
Ear ache	0, 1
Aches	0, 1
Sinus congestion	0, 1
Rash	0, 1
Chest pain	0, 1
High_temp_home	0, 1

binary values. The gender, “Female” is represented by “1” and “Male” is represented by “2”. At this stage, 14,592 records are present in the pre-processed data, with 5255 records stating flu test status as “Positive” and 9333 records stating flu test status as “Negative”.

Finally, the improper ratio (5255:9333) of the flu test status “Positive” records to “Negative” values needs to be balanced. A careful observation of the 5040 records yields the information that among the data there are patient’s with flu test status “Negative” with all the “Symptom” attributes as “0” and vaccinated status as “1”. Clearly, these entries are not relevant for our analysis, thus removing these makes the positive-to-negative ratio as 5255:4293, which is almost a balanced one.

Table 2 Medical conditions distribution

Factors combination	Values
Other	0
Immuno	1
Asthma	2
Diarrhea	3
Chronic	4
Immuno, Asthma	5
Immuno, Diarrhea	6
Immuno, Chronic	7
Asthma, Diarrhea	8
Asthma, Chronic	9
Diarrhea, Chronic	10
Immuno, Asthma, Diarrhea	11
Immuno, Asthma, Chronic	12
Immuno, Diarrhea, Chronic	13
Asthma, Diarrhea, Chronic	14
Immuno, Asthma, Diarrhea, Chronic	15

2.4 Training and Testing Sample Datasets

For the analysis purposes, standard R tools are used. Training samples and testing samples were acquired in 70:30 ratios with 6685 training samples and 2863 testing samples. Distribution of the dependent variable "Flu Test Status" in the training sample divided as "Positive" (55.03%) and "Negative" (44.97%).

Formula defined for our analysis[1]:

$$\text{Flu Test Status} \sim \text{Gender} + \text{Vaccination Status} + (\text{All the 24 Symptoms}) + \text{Medical Condtions}$$

2.5 Classification Algorithm

Support Vector Machines (SVM)

Support Vector Machine is a machine learning algorithm that is mainly used for binary classification. In SVM, outliers are weighed down, if it is not possible to separate the classes using a linear classifier in the space, thereby projecting the data points into another dimension (usually higher). There are several tuning parameters in

[1] For the symptoms attribute (Table 1).

SVM, depending on the type of kernel used. These parameters include regularization parameter (C), kernel, degree, and gamma. Authors are interested in the "linear" kernel.

Mathematically,

$$d_H(\phi(x_0)) = \frac{\left|w^T(\phi(x_0)) + b\right|}{\|w\|_2}$$

where "w" is the vector normal to the hyperplane, b is the intercept or bias term for the hyperplane, and d_H is the distance of the point from the line separating the hyperplane. The above equation gives the prediction value for the positive class (if $d_H > 0$) and negative class (if $d_H < 0$). In the train() method, method = 'svmLinear' sets the kernel to "Linear" value and the default value of C = 1 for classification.

k-Nearest Neighbor (k-NN)

k-NN or k-nearest neighbor algorithm is a non-parametric method that is used for estimation, prediction, and classification. k-NN classifies a new unclassified record by comparing it to the most similar set of records in the training set. To achieve this, the training dataset is stored. For estimation and prediction, a locally weighted averaging method is used. "k" in the k-NN defines the number of nearest neighbors included in majority of the classification process [25]. A common distance function used is Euclidean distance:

$$d_{\text{Euclidean}}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_i (x_i - y_i)^2}$$

For estimation and prediction, locally weighted averaging method is used. In locally weighted averaging, the estimated target value $\hat{y}$ is

$$\hat{y}_{\text{new}} = \frac{\sum_i w_i y_i}{\sum_i w_i}$$

By default, k equals sqrt (N), where N is the total number of samples. Since the createDataPartition() method of the "Caret" package in R is used for creating partitioning, it uses the bootstrapped sampling, i.e., the values selected for k equal to 5, 7, and 9.

Neural Networks

As pointed out by Ripley [26], neural networks consist of units, which are further arranged in layers. These layers consist of input layer, hidden layers, and output layer. All these layers are linked using a particular weight (w) which multiplies the signal traveling along with them by that factor. Also, each unit sums its input and adds a bias (constant) as a total input (x), and applies a function ϕ to input (x) to return as output (y).

$$y_k = \phi_o\left(\alpha_k + \sum_{i\to k} w_{ik}x_i + \sum_{j\to k} w_{jk}\phi_h\left(\alpha_j + \sum_{i\to j} w_{ik}x_i\right)\right)$$

The default number of iterations "maxit" in nnet is 100. The "nnet" package of the R provides the default value for decay as 0 and it allows the skip-layer units by initializing the size to 0 but it is a required argument. Also, the train() of "Caret" package in R also provides an implementation of "nnet" method. In this, size and decay tuning parameters are optional and set as size as 1, 3, and 5, and decay as 0.0001, 0.1, and 0.

Random Forests (RF)

According to Breiman [27], a random forest is defined as "A classifier consisting of a collection of tree-structured classifiers $\{h(x, k), k = 1,\ldots\}$ where the $\{k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x". For classification, the default value of x is $\lfloor\sqrt{p}\rfloor$ and minimum node size is 1, where p is number of variables in x. The advantage of using random forest is that it is claimed to "cannot over fit the data" in most cases. Increasing the tree samples will not over fit the random forest sequence [28]. The train function in R allows implementing random forest with the tuning parameter 'mtry' = sqrt(p). Hence, the values of "mtry" are 2, 14, and 27.

2.6 Accuracy Assessment and Comparisons

Parameters used for evaluation are as provided in the work of Zhu et al. [29].

Accuracy: Accuracy is defined as the number of correct cases predicted to the total number of cases present in the dataset. It measures the degree of accuracy of a case on a condition.

$$\text{Accuracy} = (\text{TN} + \text{TP})/(\text{TN} + \text{TP} + \text{FN} + \text{FP})$$

Sensitivity[2]: Sensitivity is defined as the proportion of the true positive cases that are discovered accurately by the diagnosis (diagnostic test). It is the percentage of people having infection from the dataset. Sensitivity is also called as True Positive Rate (TPR).

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN})$$

[2] TP (True Positive): number of True Positive values
TN (True Negative): number of True Negative values
FP (False Positive): number of False Positive values
FN (False Negative): number of False Negative values.

Specificity: Specificity is defined as the true negative cases that are discovered accurately from the diagnosis. It is the percentage of people having no infection from the dataset. Specificity is also called as True Negative Rate (TNR).

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP})$$

F1-Score: F1-score combines precision and sensitivity (recall) in one metric. It calculates the harmonic mean between precision and sensitivity. Range of F1-score is [0, 1].

$$\text{F1-Score} = 2 * (\text{Sensitivity} * \text{Precision})/(\text{Sensitivity} + \text{Precision})$$

Here, Precision = TP/ (TP + FP).

ROC curve: ROC curve or receiver operating characteristic curve plots sensitivity versus specificity at different classification thresholds, where FPR = FP/(FP + TN).

AUC: AUC or area under the ROC curve provides a performance measurement for the classification problems. It measures how well predictions are ranked and the quality of the technique's predictions. It calculates the entire two-dimensional area under the ROC curve. The higher the value of AUC, the better the techniques are at predicting cases with infection or no infection.

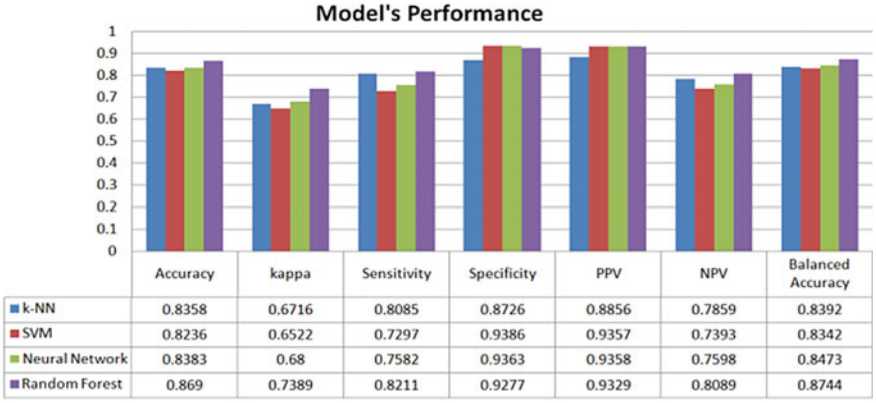
3 Results

Initially, the performance measures selected for the models are accuracy, sensitivity, and specificity. Table 3 provides the results for these models. The results are shown in ranges as the performance measure varies for different samples for training and testing. The average mean of the accuracy with the worst to best training samples acquired is 5.185%, i.e., based on the selected samples for training and testing, accuracy may vary by 5% (approximately).

From Fig. 4, it can be discerned that the method of the random forests performs better analysis than other three techniques. On an average, random forest gives 86.9% accuracy which is 3.64% more accurate than other results based on the parameter of accuracy. Although the difference in sensitivities of k-NN and random forest is

Table 3 Value band of performance measures for different machine learning techniques

Methods	Accuracy	Sensitivity	Specificity
SVM with linear kernel	76.08–82.36	68.44–72.97	85.76–93.86
k-nearest neighbor	79.75–83.58	66.44–80.85	87.26–96.6
Neural network	79.85–83.83	67.00–75.82	93.63–96.12
Random forest	**80.24–86.9**	**67.13–82.11**	**92.77–96.84**



	Accuracy	kappa	Sensitivity	Specificity	PPV	NPV	Balanced Accuracy
k-NN	0.8358	0.6716	0.8085	0.8726	0.8856	0.7859	0.8392
SVM	0.8236	0.6522	0.7297	0.9386	0.9357	0.7393	0.8342
Neural Network	0.8383	0.68	0.7582	0.9363	0.9358	0.7598	0.8473
Random Forest	0.869	0.7389	0.8211	0.9277	0.9329	0.8089	0.8744

Fig. 4 Performance measures for different machine learning techniques

<2%, but vis-a-vis the other two models it is higher by more than 5%. Similarly, the specificities of these two models, k-NN and random forest, are less than the specificities of the other two models.

Again Figs. 5 and 6 clearly represent a better performance for the random forest over the other techniques. On the other hand, the AUC graph shows a better result for the neural networks compared to the k-NN technique but the performance is opposite in the F1-score graph. This prediction occurs due to the higher variance in the values with respect to sensitivity of the model.

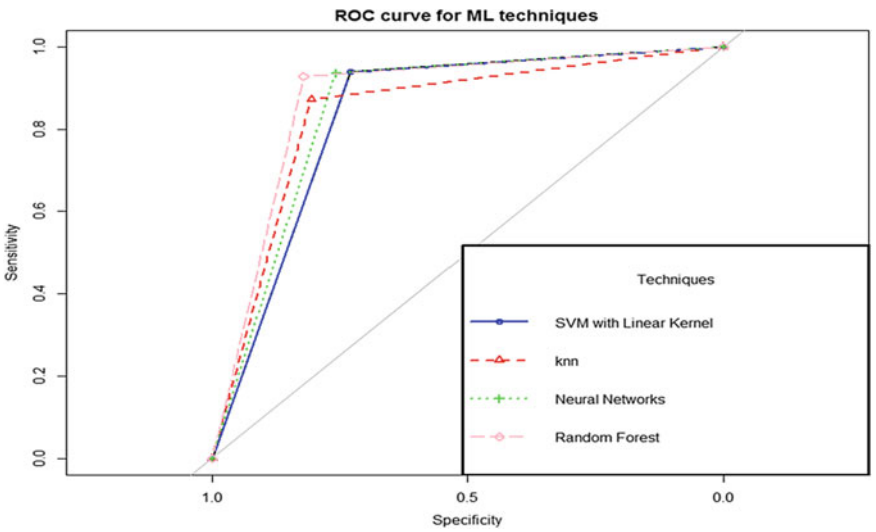


Fig. 5 ROC curve of the machine learning techniques

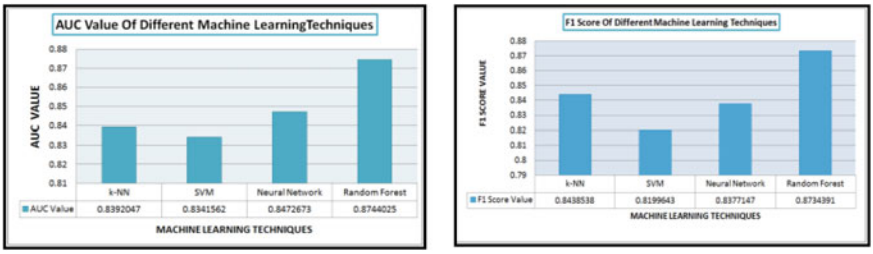


Fig. 6 **(Left)** AUC graph of different machine learning techniques. **(Right)** F1-score graph of different machine learning techniques

As in [30], it is clearly verified from the AUC graph that all the selected techniques fit on the data model and out of these, random forest shows the best results.

From the above results, currently, random forest model is the best performing model and should be chosen for developing system for the same. But as the data will grow, it might be possible that neural network (nnet) can overrule the random forest accuracy and predict more efficiently. Due to the categorical data and large number of training examples, SVM with linear kernel and k-NN performances are lower than other two techniques.

4 Conclusion

In this paper, four different well-established machine learning techniques have been implemented over the analytic data of Human Surveillance Records accessed from the Influenza Research Database. Out of these techniques, random forest returns better results with accuracy ranging from 82 to 87% as compared to SVM, k-NN, and neural networks which show accuracies of 77–82%, 79–83.5%, and 79–84%, respectively. On the basis of ROC curve, AUC graph, and F1-score graph, random forest dominates over other techniques. It is also found that the methods of neural networks and k-NN perform almost similar to each other.

From this study, it is clearly identified that any medical tests including RIDTs and DIAs give accuracy <80% of the positive prediction. But this model using any of the four techniques will guarantee to give better results, i.e., >80%. And also, compared to NAATs, it will be more economically feasible because its features are the common medical tests and information from the patient's medical history.

5 Future Scope

To further improve the accuracy of the models, feature selections, identify top features, and sampling using statistical parameters will be done to acquire much better predictions. A mixed strategy of top techniques or a new modeling technique may also develop to improve the efficiency of the machine. Several other parameters will also be included to train the model at the much deeper levels based on strain type, age category, geographical distributions, etc.

References

1. P. Spreeuwenberg, M. Kroneman, J. Paget, Reassessing the global mortality burden of the 1918 influenza pandemic. Am. J. Epidemiol. **187**(12), 2561–2567 (2018). https://doi.org/10.1093/aje/kwy191
2. W. Paul Glezen, Emerging infections: pandemic influenza. Epidemiol. Rev. **18**(1), 64–76 (1996). https://doi.org/10.1093/oxfordjournals.epirev.a017917
3. J.S. Malik Peiris, M.D. de Jong, Y. Guan, Avian Influenza Virus (H5N1): a threat to human health. Clin. Microbiol. Rev. **20**(2), 243–267 (2007). https://doi.org/10.1128/CMR.00037-06
4. S.J. Sullivan et al., 2009 H1N1 influenza. Mayo Clin. Proc. **85**(1), 64–76 (2010). https://doi.org/10.4065/mcp.2009.0588
5. Centers for Disease Controls and Prevention (CDC), https://www.cdc.gov/flu/about/burden/2019-2020.html. Accessed Mar 5, 2021
6. R.B. Squires, J. Noronha, V. Hunt, García-Sastre, et al.: Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. Influenza Other Respiratory Viruses **6**(6), 404–416 (2012). https://doi.org/10.1111/j.1750-2659.2011.00331.x
7. N. Lavrač, Selected techniques for data mining in medicine. Artif. Intell. Med. **16**(1), 3–23 (1999). https://doi.org/10.1016/S0933-3657(98)00062-1
8. A.C. Tan, D. Gilbert, Ensemble machine learning on gene expression data for cancer classification, in: *Proceedings of New Zealand Bioinformatics Conference* (University of Glasgow, Te Papa, Wellington, New Zealand, 2003), pp. 13–14
9. G. Battineni, G.G. Sagaro, N. Chinatalapudi, F. Amenta, Applications of machine learning predictive models in the chronic disease diagnosis. J. Pers. Med. **10**(2), 21 (2020). https://doi.org/10.3390/jpm10020021
10. V. Rajalakshmi, D. Sasikala, A. Kala, A predictive analysis for heart disease using machine learning, in *Intelligent Computing and Applications. Advances in Intelligent Systems and Computing*, vol. 1172 (Springer, 2021). https://doi.org/10.1007/978-981-15-5566-4_42
11. G. Meyfroidt, F. Güiza, J. Ramon, M. Bruynooghe, Machine learning techniques to examine large patient databases. Best Pract. Res. Clin. Anaesthesiol. **23**(1), 127–143 (2009). https://doi.org/10.1016/j.bpa.2008.09.003
12. H. Li, F. Sun, Comparative studies of alignment, alignment-free and SVM based approaches for predicting the hosts of viruses based on viral sequences. Sci. Rep. **8**, Article no.10032 (2018). https://doi.org/10.1038/s41598-018-28308-x
13. J. Li, S. Zhang, B. Li, Y. Hu et al., Machine learning methods for predicting human-adaptive influenza a viruses based on viral nucleotide compositions. Molec. Biol. Evol. **4**(37), 1224–1236 (2020). https://doi.org/10.1093/molbev/msz276
14. Influenza Research Database (Human Surveillance Record), https://www.fludb.org/brc/influenza_humanSurveillanceData_search.spg?method=ShowCleanSearch&decorator=influenza. Accessed Jan 28, 2021
15. Health Data, https://healthdata.gov/dataset/influenza-surveillance. Accessed Mar 5, 2021

16. Data.World, https://data.world/datasets/influenza. Accessed Mar 5, 2021
17. WHO (World Health Organization), https://www.who.int/influenza/gisrs_laboratory/flunet/en/, last accessed 2021/03/05.
18. CDC (Centers for Disease Control and Prevention), https://www.cdc.gov/flu/weekly/index.htm. Accessed Mar 5, 2021
19. J. Merckx, R. Wali, I. Schiller, C. Caya et al., Diagnostic accuracy of novel and traditional rapid tests for influenza infection compared with reverse transcriptase polymerase chain reaction: a systematic review and meta-analysis. Ann. Intern. Med. **167**(6), 394–409 (2017). https://doi.org/10.7326/M17-0848
20. N.P. Pai, C. Vadnais, C. Denkinger, N. Engel, M. Pai, Point-of-care testing for infectious diseases: diversity, complexity, and barriers in low- and middle-income countries. PLOS Med. **9**(9) (2012). https://doi.org/10.1371/journal.pmed.1001306
21. A.S. Monto, S. Gravenstein, M. Elliott, M. Colopy, J. Schweinle, Clinical signs and symptoms predicting influenza infection. Arch. Intern. Med. **160**(21), 3243–3247 (2000). https://doi.org/10.1001/archinte.160.21.3243
22. E. Marquez, V. Barrón, Artificial intelligence system to support the clinical decision for influenza, in *IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)* (Ixtapa, Mexico, 2019), pp. 1–5. https://doi.org/10.1109/ROPEC48299.2019.9057056
23. A.L. Pineda, Y. Ye, S. Visweswaran, F. Gregory et al., Comparison of machine learning classifiers for influenza detection from emergency department free-text reports. J. Biomed. Inform. **58**, 60–69 (2015). https://doi.org/10.1016/j.jbi.2015.08.019
24. M.A. Khan, W.U. Abidi, M.A. Al Ghamdi, S.H. Almotiri, S. Saqib, et al., Forecast the influenza pandemic using machine learning. Comput. Mater. Continua **66**(1), 331–340 (2021). https://doi.org/10.32604/cmc.2020.012148
25. D.T. Larose, C.D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*, 2nd edn. (Wiley, 2014), pp. 149–164. https://doi.org/10.1002/9781118874059.ch7
26. B. Ripley, *Pattern Recognition and Neural Networks* (Cambridge University Press, 1996), pp. 143–180. https://doi.org/10.1017/CBO9780511812651.006
27. L. Breiman, Random Forests. Mach. Learn. **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
28. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*. Springer Series in Statistics (Springer, New York, NY, 2009). https://doi.org/10.1007/978-0-387-84858-7_15
29. W. Zhu, N. Zeng, N. Wang, Sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical SASÒ implementations, in *NESUG Proceedings: Health Care and LifeSciences*, Baltimore, Maryland (2010)
30. J.V. Carter, J. Pan, S.N. Rai, S. Galandiuk, ROC-ing along: evaluation and interpretation of receiver operating characteristic curves. Surgery **159**(6), 1638–1645 (2016). https://doi.org/10.1016/j.surg.2015.12.029

Fuzzy Time-Series Models Based on Intuitionistic Fuzzy, Rough Set Fuzzy, and Differential Evolution

Partha Pratim Deb, Diptendu Bhattacharya, and Indranath Chatterjee

Abstract Time-series forecasting is a vital concern for any data having temporal variations. Comparing with the other conventional time-series methodologies, the fuzzy time-series (FTS) proved its superiority. Substantial research using time-series forecasting to predict the stock index data has been found in the earlier works. The fuzzy sets approach alone cannot explain the data thoroughly. In this article, we have proposed three different methods of time-series forecasting. The first method is based on a rough set of FTS, a rule induction-based method; the second method is based on intuitionistic FTS. The last method is the extension of the second method using differential evolution. In the first model, a fuzzy algorithm based on rules is used to derive prediction rules from the time-series data and adopt an adaptive expectation model that replaces the fuzzy logical relationships or groups. In the second method, to split the universe of discourse into a non-uniform interval, a clustering algorithm-based intuitionistic fuzzy approach is used, taking care of the membership and non-membership function. Finally, the last method has been tuned for a better outcome using differential evolution. To examine the results, contrast analyses on the Taiwan stock exchange data and daily cases of COVID-19 pandemic prediction have been carried out. The outcome of the proposed approaches validates that the first and second techniques, showing promising results. However, the third method outperforms the other methods and the present techniques concerning the root-mean-square error metric.

Keywords Rough set · Fuzzy set · Intuitionistic set · Time series · Stock index · Differential evolution · COVID-19

P. P. Deb · D. Bhattacharya
Department of Computer Science and Engineering, National Institute of Technology, Agartala, Tripura 799046, India

I. Chatterjee (✉)
Department of Computer Engineering, Tongmyong University, Busan 48520, South Korea

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_10

1 Introduction

The time-series (TS) expectation hypothesis assumes an essential part in society, economy, assets, and nature. In any case, traditional determining approaches are generally found on measurable translation, for example, the ARMA model and the ARIMA model. These procedures have two deficiencies: right off the bat, they require loads of past information meeting a few necessities; besides, they cannot deal with the phonetic qualities and uncertain information. Thus, Song and Chissom [1–3] presented the fuzzy time-series (FTS) forecast model, which could effectively oversee fuzzy information with the joining of fuzzy sets and fuzzy rationale. The central thought of FTS is that customary information is meant as fuzzy sets, and the series varieties are addressed as fuzzy connections. Information is anticipated by fuzzy legitimate thinking when there isn't much verifiable information or loose information. Loads of exploration work have been finished by utilizing the thought of FTS as of recently. Chen and Hwang [4] proposed another FTS model called the two-figured time-variation FTS model to manage expectation issues. Melody et al. [5] built up a fuzzy stochastic FTS model focusing on a specific sort of customary fuzzy information whose possibilities are also fuzzy sets. The varieties of chronicled information are utilized by Hwang et al. [6] as opposed to the actual data to shape a time-variation FTS model. This FTS model is unique concerning Song and Chissom's model, yet it gives solid outcomes. Jiang et al. [7] proposed a half-breed expectation system for gauging wind speed information utilizing FTS and MODE algorithm. Some limitations have been found after a detailed literature review, that (a)most earlier TS models used a single variable to predict stock index price, (b) optimal interval length is considered to be a serious issue that influences the forecasting result, (c) system performance validation is one of the most critical concerns which validates the accuracy of the model.

The learning may upgrade the deficiencies of initial FTS models from examples such as module, version 2 (LEM2) algorithm (learning from examples module, version 2, [8]). The first suggested method is a rough set rule induction-based FTS model that is put forward here to forecast stock index data.

In the intuitionistic fuzzy set (IFS) [9], three indicators illustrate data: the membership, the non-membership, and an intuitionistic index, which makes it more exact correct in fuzzy information classification. Castillo et al. [10] coupled the IFS with TS analysis and has used an intuitionistic fuzzy reasoning scheme for data forecast. Hence, to minimize the limitations, we will employ IFS as a second method.

Differential evolution (DE) is a kind of evolutionary calculation strategy that improves an issue by repeatedly attempting to build up a competitor arrangement concerning a given model of value. Das et al. [11] broke down specialized possibilities, applications, and hybridization viewpoints of PSO and differential evolution calculations. Das et al. [12] utilized the differential evolution technique in a mutation-based area.

2 Rough Set Theory

Rough Set Theory (RST) is an approach that has roots in set theory, which has no parameters or which is not parametric and holds a substantial procedure to decision-based difficulties [13]. This technique is of specific importance in artificial intelligence (AI) research, particularly in machine learning (ML), an understanding examination from databases (DB), support systems, knowledge discovery in databases (KDD), analytical reasoning, and design recognition. LEM2 (Learning from Examples Module Version 2, [8]), rule selection algorithm, is frequently used, and it allows the significant consequence in maximum case. LEM2 inspects the inspection space of the attribute-value matches. In a general perception, LEM2 (Learning from Examples Module Version 2) computes a limited coverage. Hereafter, it converts it into a ruleset. LEM2 grasps a particular rule set along with grasps the least set of minimum rules describing a notion.

3 Adaptive Expectation Model

It has been found that the adaptive expectation model [14] is found to be a feasible prediction model for different applications but especially in stock index forecasting. The prediction formula is generated by the following equation mentioned as Eq. (6):

$$\text{Forecast}_t = C_{t-1} + \lambda * \Delta(t-1), \tag{1}$$

where Forecast_t and C_{t-1} are the prediction of stock index price at time t and $(t-1)$, respectively, the value of λ is used here as an error-adjustment parameter for $\Delta(t-1)$, and $\Delta(t-1) = \text{Forecast}_{t-1} - C_{t-1}$ is supplied here for the prediction error found at time t.

4 Differential Evolution Algorithm

The standard differential evolution (DE) algorithm [11, 15–17] incorporates four steps:

(i) initialize experiment outcomes,
(ii) mutation,
(iii) crossover, and
(iv) selection.

5 Proposed Approach

Method 1: Prediction using Rough Set Fuzzy

For a quick understanding of the suggested model, the procedure is alienated into four segments. The first segment, which contains Step 1–Step 3, is used for data pre-processing. Phase 2, comprising of Step 4, is used to generate rules using the LEM2 algorithm, and phase 3 consists of Steps 5 and 6 represent the initial prediction. Adaptive adjustment for prediction, respectively, and in the last phase, we have compared the result using Step 7.

Step 1: Some important variables are considered here from earlier studies [18], which are lag periods, variation, and gradient in the statistical method. LAG1, LAG2, and LAG3 are represented as lagged period of first order (close_{t-1}), second order (close_{t-2}), and third order (close_{t-3}), respectively. Variation of Close price (VAR) where

$$\text{VAR}_t = \text{close}_t - \text{close}_{t-1}, \quad \text{Gradient}(G) = G_i = \frac{\text{VAR}_i - \text{VAR}_{t-1}}{t - (t-1)}, \quad \text{and}$$

Close (C) = The closing value for business day t.

Step 2: In this step, we have stated the universe of discourse (UoD) and the linguistic intervals. Let $\text{D}_{\min}$ be the nominal and $\text{D}_{\max}$ be the utmost value of a specific past dataset, respectively. Then, to ensure the least and largest interval of UoD, U can be defined as $\text{U} \rightarrow [\text{Interval}_{\text{Least}}, \text{Interval}_{\text{Largest}}]$, where

$$\text{Interval}_{\text{Least}} = D_{\min} - D_1, \quad \text{Interval}_{\text{Largest}} = D_{\max} - D_2,$$

and D_1 and D_2 are the two real values considered positive and used to partition the UoD into the intervals of n equal length, such as u_1, u_2, …and u_n.

To express the intervals of linguistic values, we will follow the number of categories of individual short memory function from the study of Miller [19], which is 7 ± 2. So, according to this, the UoD is partitioned into the equal length of seven linguistic intervals with u_1, u_2, u_3, u_4, u_5, u_6, u_7 (seven used as an example), and the length I is defined as follows:

$$I = \frac{\left[\text{Interval}_{\text{Least}}, \text{Interval}_{\text{Largest}}\right]}{n} = \frac{[D_{\min} - D_1, D_{\max} - D_2]}{n}. \tag{2}$$

Step 3: Here, we define the factors by their symbolic names first.

L1 ← LAG1, L2 ← LAG2, L3 ← LAG3, VAR_t ← Variation, G_t ← Gradient, C ← Close price.

Fuzzy set L1 is partitioned into L1_1, L1_2, L1_3, L1_4, L1_5, L1_6, and L1_7. Each L1_i is represented by the intervals u_i.

$$\begin{aligned} L1_1 &= \left\{ \frac{1}{u_1} + \frac{0.5}{u_2} + \frac{1}{u_3} + \frac{1}{u_4} + \frac{1}{u_5} + \frac{1}{u_6} + \frac{1}{u_7} \right\} \\ &\vdots \\ L1_7 &= \left\{ \frac{1}{u_1} + \frac{1}{u_2} + \frac{1}{u_3} + \frac{1}{u_4} + \frac{1}{u_5} + \frac{0.5}{u_6} + \frac{1}{u_7} \right\} \end{aligned} \tag{3}$$

where "+" is used as a union operator.

Step 4: This step intends to render a set of significant rules to predict stock index values. Here, we use the LEM2 algorithm to generate "if–then" decision rules: the composition of different fuzzy conditional values, decision values, and "supports" that represent the amount of information in the dataset matching with the produced decision rules.

Step 5: In this step, we perform forecasting. The generated rules from step 4 are used to find the decision values pi of the matching rules as per Table 1. In trading day t, if no linguistic value has been found for the testing dataset, this study creates and adds a new linguistic interval or value for each of the fuzzy variables for more appropriate forecasting. Suppose there are no linguistic values in fuzzy variables of the testing data of trading Day t. In that case, this study will add a new linguistic value (interval) for each fuzzy variable to a more reasonable forecast consideration.

Then we apply Adaptive Expectation Model (AEM) to forecast the values like.

$\text{Forecast}_t = C_{t-1} + \lambda * \Delta(t-1)$. The prediction of stock price at time t C_{t-1} is the actual stock price value at time $(t-1)$, λ is the error-adjustment parameter for $\Delta(t-1)$, and $\Delta(t-1) = \text{Forecast}_{t-1} - C_{t-1}$ is the prediction error at time t.

Step 6: In this step, we performed the defuzzification. This step is performed by taking the middle point of the linguistic decision values C_i as the defuzzified value.

Step 7: To compare the prediction performance of the suggested model with the listing illustrations, this step utilizes the daily TAIEX prices as the verification dataset. The experimental dataset is split into two different subsets, where first set consists of 10-month data from January to October as the training data, and November to December is used as testing data. We have taken the models [20–23] as a reference for comparison to forecast TAIEX testing data. Here, RMSE is used as an evolution criterion.

Table 1 Rules generated for TAIEX dataset used in the model

No.	Rules
1	If (LAG1 = $L1_3$) And (LAG2 = $L2_3$) And (LAG3 = $L3_3$) And ($VAR_t = VAR_5$) Then (Class = C_3)
2	If (LAG1 = $L1_5$) And (LAG2 = $L2_5$) And (LAG3 = $L3_5$) And ($VAR_t = VAR_5$) Then (Class = C_5)
3	If (LAG1 = $L1_6$) And ($VAR_t = VAR_5$) Then (Class = C_6)
4	If (LAG1 = $L1_8$) And (LAG2 = $L2_8$) And (LAG3 = $L3_8$) And ($VAR_t = VAR_5$) And (G = G_4) Then (Class = C_8)
5	If (LAG1 = $L1_6$) And (LAG2 = $L2_6$) And (LAG3 = $L3_6$) And ($VAR_t = VAR_5$) Then (Class = C_6)

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\text{forecasted}(\text{value}_i) - \text{actual}(\text{value}_i))^2}{n}}, \tag{4}$$

where n indicates the prediction of the number of trading days. After the error-adjustment parameter (λ) is achieved, the initial prediction of the testing data is enhanced by Eq. (1).

Method 2: Prediction using Intuitionistic Fuzzy Set

Step 1: The UoD is defined as $U = [a_{\min} - P_1, a_{\max} - P_2]$, where $a_{\min}$ is minimum and $a_{\max}$ is the maximum historical data. P_1 and P_2 are two positive numbers. Normally, P_1 and P_2 are used to round-off the values of $a_{\min}$ and $a_{\max}$ to proper integer numbers.

Step 2: Let $A = \{a_1, a_2, \ldots, a_1\}$, where $a_j = \{a_{j1}, a_{j2} \ldots, a_{jm}\}$. For standardizing historical data, the elements must lie between 0 and 1. Data in different forms must be transformed into the interval [0, 1], to meet the standards of similarity matrix R [24]. Here, we use two types of transformations.

(1) Z-score transformation
(2) Range transformation

Step 3: Let s_{jl} be the similarity between $\hat{a}_j$ and $\hat{a}_l$, then the fuzzy similarity matrix will be $R = (s_{jl})_{\text{nxm}}$. Though there are different ways to calculate s_{jl}, Euclidean distance is usually used in similarity matrix formation [24].

Step 4: In this step, prims algorithm [25] is used to generate a maximum spanning tree (MST). Arbitrarily select a starting vertex a_1 in A. find that vertex $a_2 \in A - \{a_1\}$ which is most similar to the set $F(a_1, a)$ where F is the fuzzy relation over a $\in A - \{a_1\}$. We will put an edge (a_1, a_2) to form the graph. The vertices of the associated branch produce a group.

Step 5: In this step, the calculation of δ is determined. The value of δ ranges from 0 to 1, and the best δ among them leads to the fittest classification. F_{stat} is used as widely used statistics to find the best δ:

$$F_{\text{stat}} = \frac{\text{distance between the groups}}{\text{distance within the groups}}, \quad F_{\text{stat}} = \frac{\text{dist}_{\text{between}}}{\text{dist}_{\text{within}}} \tag{5}$$

where $\text{dist}_{\text{between}}$ and $\text{dist}_{\text{within}}$ g are the numbers of groups for δ.

n_i is the number of objects in group i ($i = 1, 2, \ldots, g$). $F_{\text{stat}} \sim F_{\text{stat}}(g-1, n-g)$.

Here when the F_{stat} is bigger, then the classification will be better.

Step 6: We define g IFSs corresponding to the above g intervals:

$$C_i = \left\{ \left\langle a, \mu_{C_i}(a), \delta_{C_i}(a) \right\rangle | a \in A \right\}, i = 1, 2, \ldots, g. \tag{6}$$

Here, membership and non-membership function has been created.

For objective analysis, two rules depend on the location: (i) middle of an interval and (ii) boundary of the interval.

$$a_{\text{middle}} = \frac{(p_{i-1} + p_i)}{2},\ a_{\text{boundary}} = p_i,$$

Rules:

If a_{middle} Then $\mu_{C_i}(a_{middle}) = 1$ and $\delta_{C_i}(a_{middle}) = 0$.

If a_{boundary} Then $\mu_{C_i}(a_{\text{boundary}}) = \delta_{C_i}(a_{\text{boundary}})$ [$\because$ intuitionistic index (IIX) = maximum value].

After this, membership and non-membership both defined as Gaussian Function

$$\mu_{C_i} = \left(\exp\left(-\frac{(a - a_{\text{middle}})^2}{2\gamma_{\mu_i}^2}\right)\right) \tag{7}$$

$$\delta_{C_i} = \left(1 - \exp\left(-\frac{(a - a_{middle})^2}{2\gamma_{\delta_i}^2}\right)\right) \tag{8}$$

Step 7: In this step, reasoning model is generated. Let C_i ($i = 1, 2, \ldots, n$) and C^* be IFSs in universe U and D_i ($i = 1, 2, \ldots, n$) and D^* be IFSs in universe V, based on the intuitionistic fuzzy relation [32].

Inspired by the intuitionistic fuzzy multiple modus ponens, we exchanged historical data (I_j) and IFSs (C_i) in the IFTS model.

The rules as follows:

IF a is I_1 THEN b *is* I_2
IF a is I_2 THEN b *is* I_3
$\vdots$
IF a is I_j THEN b *is* I_{j+1}
$\vdots$
IF a is I_{t-1} THEN b *is* I_t
Input or facts: IF a is I
Output or consequence: b *is* I_{t+1}^*

Step 8: In this step, we have to choose from different implication operators and choose the one with better performance. Since the Mamdani implication operator shows better performance [26], we use it here.

The reason output is

$$I_{t+1}^* = I_t \circ R \tag{9}$$

Step 9: Though different defuzzification algorithms exist, such as gravity algorithm, maximum truth-value algorithm, and weighted average algorithm [26], we use the gravity algorithm, which gives smoother output than others even when there are few changes [26].

The calculation is

$$Q = \frac{\int_U q(\mu_I(q) + (1/2)\pi_I(q))dq}{\int_U (\mu_I(q) + (1/2)\pi_I(q))dq} = \frac{\int_U q(1 + \mu_I(q) + \delta_I(q))dq}{\int_U (1 + \mu_I(q) + \delta_I(q))dq} \tag{10}$$

where U is the output domain and I is the IFS in U.

Step 10: Here, we will use two different datasets for TAIEX and COVID-19 Case prediction, respectively. The TAIEX is a conventional financial data set widely applied in fuzzy time series (FTS) prediction. This analysis has taken 10 months of data from January 1 to October 31 as a training set, and 20 months of data from November 1 to December 31, 2004 are utilized as testing data.

We have collected the COVID-19 data from WHO website (https://covid19.who.int/table) and dataset website Kaggle (https://www.kaggle.com/gpreda/coronavirus-2019ncov) and extracted the data. We have taken the models [2, 20–23, 32–36] as a reference for comparison to forecast TAIEX and models [30, 31] as a reference for forecasting COVID-19 Daily Case up to November 10, 2020. Here, RMSE is used as an evolution criterion.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\text{forecasted}(\text{value}_i) - \text{actual}(\text{value}_i))^2}{n}}, \tag{11}$$

where n indicates the prediction of the number of trading days.

Method 3: Prediction using IFTS and Differential Evolution

Till now, we have used the fixed structure of membership functions in the above-proposed methods. But, in practice, it is found that the fixed structure of a membership function cannot give accurate results in most of the cases and hence the RMSE value will be greater. To handle such problems, this method uses the adaptation of membership functions. The transformation of MFs is achieved by ideally choosing the base-width of the triangular MFs (type-1) to change the slopes of the MFs and it can be made by employing an evolutionary algorithm.

Though PSO or evolutionary algorithm can serve the objective, differential evolution (DE) algorithm is utilized here for its higher accuracy and speed of convergence. The effectiveness of DE is already tested with our experience of practicing it for numerous years [11, 15].

The steps of the adaptation algorithm are named below according to [11, 15]:

1. Initialization
2. Fitness evaluation
3. Adaptation
4. Convergence

5. Reporting
6. Looping

In the sixth step of the adaptation algorithm, repetition from step 2 by increasing day by 15 days is needed (i.e., day = day + 15) until day > 45 (i.e., December 15).

6 Experimental Results

Experimental result shows that our proposed models outperform other reference models. Tables 2 and 3 show the comparison of RMSE values for TAIEX 2004 data and COVID-19 data. Figures 1, 2, 3, and 4 show the graph of actual TAIEX data vs. proposed models 1, 2, 3, and all the proposed models with other reference models. Figures 5, 6, 7, and 8 show the graph of actual Corona confirmation data versus proposed models 1–3 and all the proposed models with other reference models.

Table 2 Comparison of RMSEs obtained by the proposed techniques with existing techniques for TAIEX forecasting

Methods	Year
	2004
Model [2]	61.17
Model [35]	52.63
Model [33]	53.63
Model [34]	50.27
Model [20]	68.82
Model [21]	71.63
Model [22]	83.39
Model [23]	85.3
Model [36]	82.90
Model [32]	54.50
Proposed Method 1	58.6
Proposed Method 2	43.23
Proposed Method 3	41.50

Table 3 Comparison of RMSEs obtained by the proposed technique with existing techniques for daily cases of COVID-19 forecasting

Methods	Year
	2020
Model [30]	0.33
Model [31]	0.26
Proposed Method 1	0.17
Proposed Method 2	0.11
Proposed Method 3	0.02

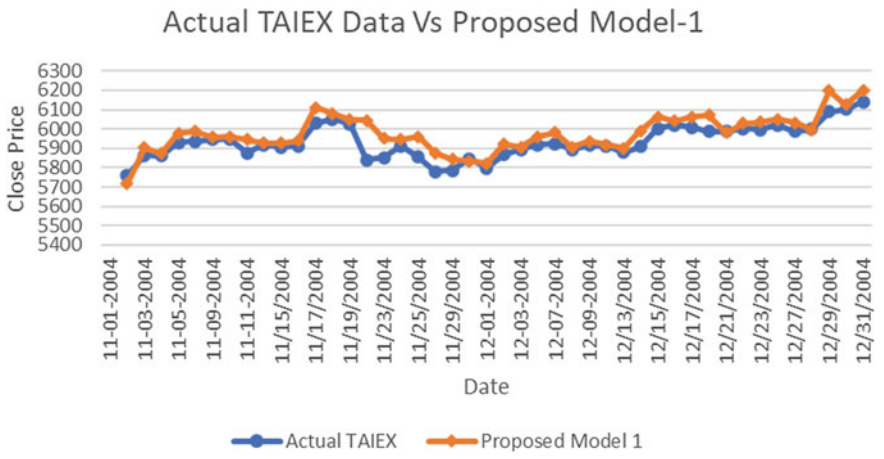


Fig. 1 Actual TAIEX data versus Proposed Model 1

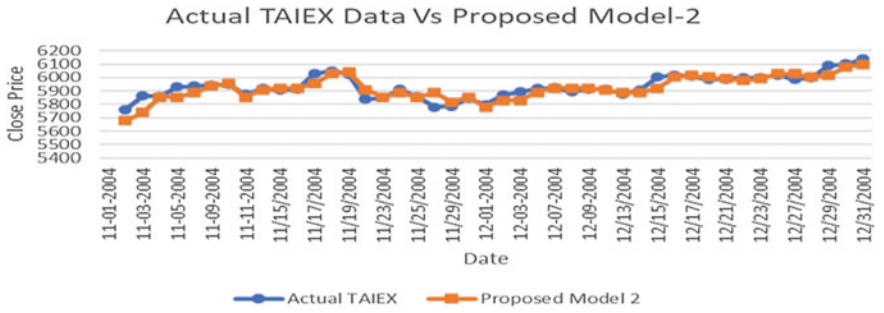


Fig. 2 Actual TAIEX data versus Proposed Model 2

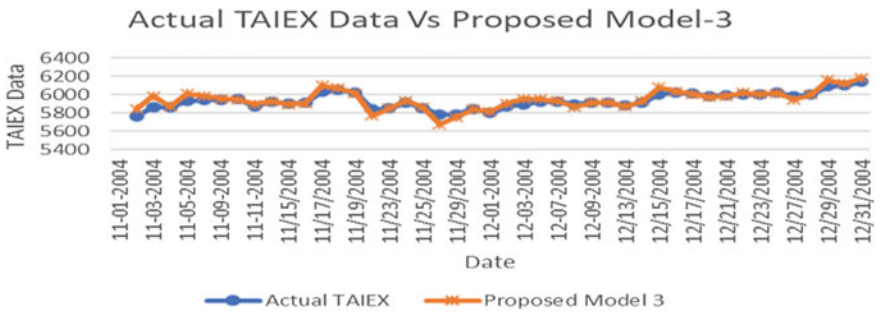


Fig. 3 Actual TAIEX data versus Proposed Model 3

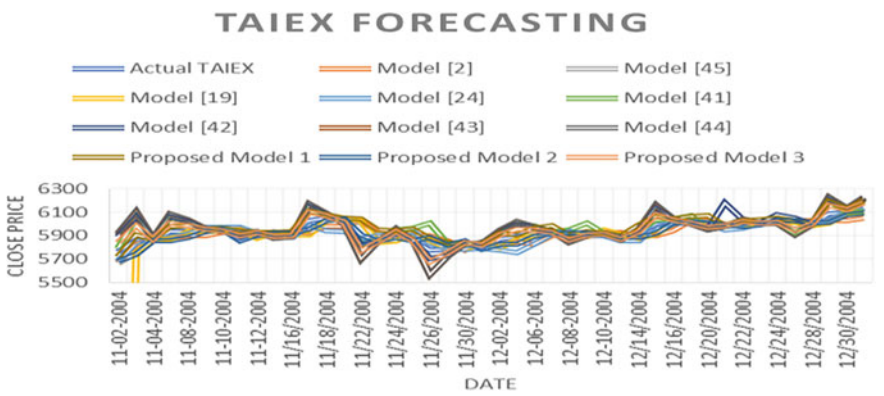


Fig. 4 TAIEX forecasting for Proposed Models versus Reference Models

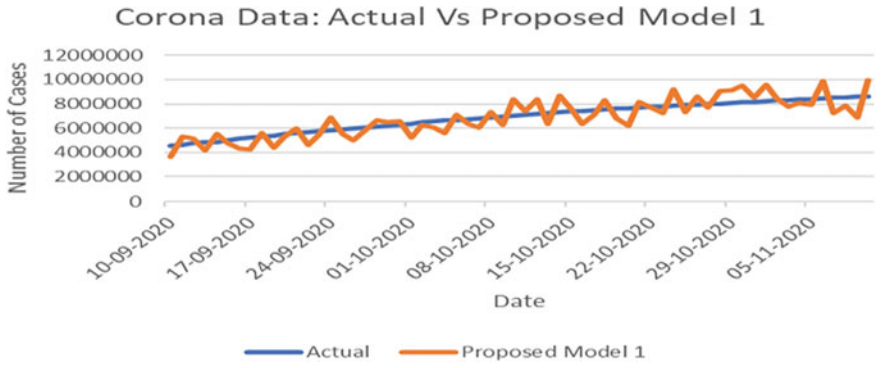


Fig. 5 Actual Corona data versus Proposed Model 1

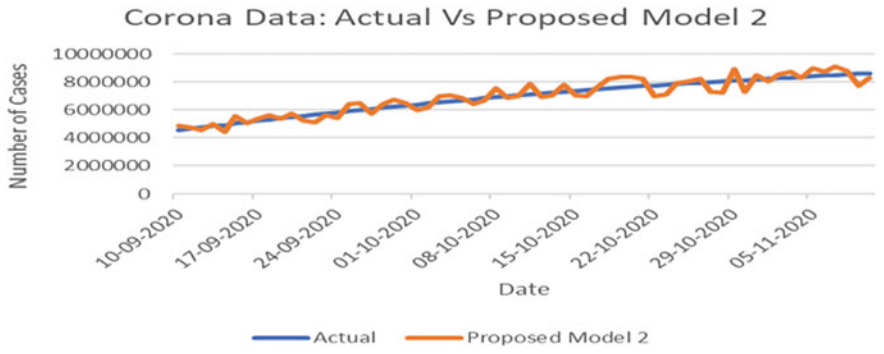


Fig. 6 Actual Corona data versus Proposed Model 2

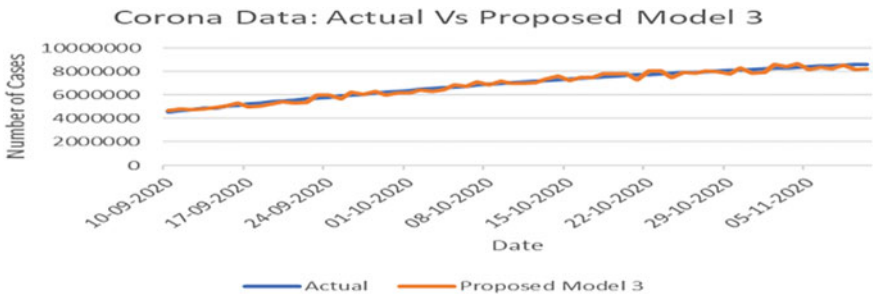


Fig. 7 Actual Corona data versus Proposed Model 3

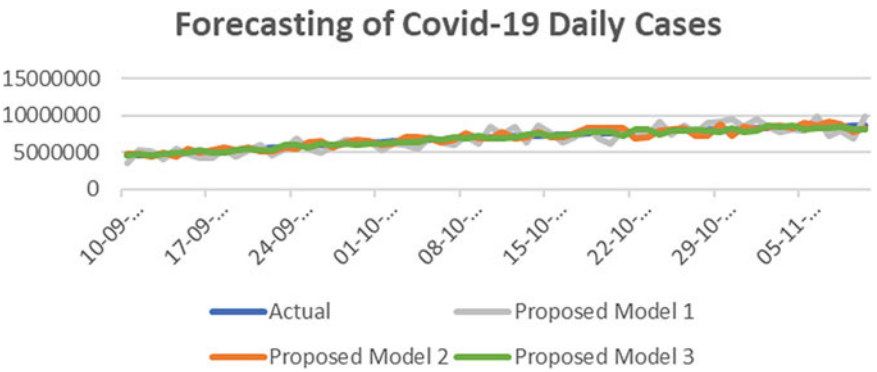


Fig. 8 Actual Corona data versus All the Proposed Models

7 Discussion

The first proposed model is a novel FTS method based on a fuzzy rough set which is a rule induction method for forecasting stock index and daily cases of COVID-19 pandemic. This model employed a rough set LEM2 algorithm to produce prediction rules.

The second proposed model is a novel IFTS model used along with MST-based fuzzy clustering to partition the UoD into unequal intervals. To determine the membership and non-membership function of this intuitionistic fuzzy set, a more accurate approach is proposed to handle the characteristics of the partitioned data. Intuitionistic fuzzy reasoning is used to build prediction rules and this causes the model more perceptive to the fuzzy variation of unknown data. The result based on the TAIEX and COVID-19 dataset reveal that this model has forecasted better than the 1st proposed model with an RMSE of 43.23 and 0.11 for TAIEX forecasting and COVID-19 daily case forecasting, respectively.

The third method has provisions for MF tuning in the analysis stage. Such tuning is lacked to adapt the forecasting algorithm by structural modifications in the time

series. The third method produces the most reliable RMSE value, outperforming all other mentioned algorithms, including the three proposed models concerning RMSE measure, which is 41.50 for TAIEX forecasting and 0.02 for COVID-19 daily forecasting.

8 Conclusion and Future Work

In this article, we have proposed three different methods of time-series forecasting. The first method is based on a rough set of FTS, a rule induction-based method; the second method is based on intuitionistic FTS. The last method is the extension of the second method using differential evolution. The outcome of the proposed approaches validates that the first and second techniques, showing promising results. However, the third method outperforms the other methods and the present techniques concerning the root-mean-square error metric. Several issues can be addressed by the above methods using more optimized algorithms. Other stock index-based databases like NASDAQ or Dow Jones may be implemented. From COVID-19-based datasets zone or state-wise confirmed, recovered and death cases can be forecasted too. Different kinds of fields can also be utilized with different machine intelligence-based methods.

References

1. Q. Song, B.S. Chissom, Fuzzy time series and its models. Fuzzy Sets Syst. **54**(3), 269–277 (1993)
2. Q. Song, B.S. Chissom, Forecasting enrolments with fuzzy time series—part I. Fuzzy Sets Syst. **54**(1), 1–9 (1993)
3. Q. Song, B.S. Chissom, Forecasting enrolments with fuzzy time series—part II. Fuzzy Sets Syst. **62**(1), 1–8 (1994)
4. S.M. Chen, J.R. Hwang, Temperature prediction using fuzzy time series. IEEE Trans. Syst. Man Cybern. B Cybern. **30**(2), 263–275 (2000)
5. Q. Song, R.P. Leland, B.S. Chissom, Fuzzy stochastic fuzzy time series and its models. Fuzzy Sets Syst. **88**(3), 333–341 (1997)
6. J.-R. Hwang, S.-M. Chen, C.-H. Lee, Handling forecasting problems using fuzzy time series. Fuzzy Sets Syst. **100**(1–3), 217–228 (1998)
7. P. Jiang, H. Yang, J. Heng, A hybrid forecasting system based on fuzzy time series and multi-objective optimization for wind speed forecasting. Appl. Energy **235**, 786–801 (2019). https://doi.org/10.1016/j.apenergy.2018.11.012
8. J.W. Grzymala-Busse, A new version of the rule induction system LERS. Fundam. Inform. **31**, 27–39 (1997)
9. K.T. Atanassov, Intuitionistic fuzzy sets. Fuzzy Sets Syst. **20**(1), 87–96 (1986)
10. O. Castillo, A. Alanis, M. Garcia, H. Arias, An intuitionistic fuzzy system for time series analysis in plant monitoring and diagnosis. Appl. Soft Comput. **7**(4), 1227–1233 (2007)
11. S. Das et al., Particle swarm optimization and differential evolution algorithms: technical analysis, applications and hybridization perspectives. Stud. Comput. Intell. SCI **116**, 1–38 (2008)

12. S. Das, A. Abraham, U.K. Chakraborty, A. Konar, Differential evolution using a neighborhood-based mutation operator. IEEE Trans. Evol. Comput. **13**(3), 526–553 (2009)
13. A. Wakulicz-Deja, P. Paszek, International diagnose progressive encephalopathy applying the rough set approach, JFMI **46**, 119–127 (1997)
14. J. Kmenta, *Elements of Econometrics*, 2nd edn., (MacMillan, 1986)
15. S. Das, P.N. Suganthan, Differential evolution: a survey of the state-of-the-art. IEEE Trans. Evol. Comput. **15**(1), 4–31 (2011)
16. A. Halder, A. Konar, R. Mandal, A. Chakraborty, P. Bhowmik, N.R. Pal, A.K. Nagar, General and interval type-2 fuzzy face-space approach to emotion recognition. IEEE Trans. Syst. Man Cybern. Syst. **43**(3), 587–605 (2013)
17. K.V. Price, R.M. Storn, J.A. Lampinen (eds.), *Differential Evolution: A Practical Approach* (Springer, New York, 2005)
18. S.M. Chen, P.Y. Kao, TAIEX forecasting based on fuzzy time-series, particle swarm optimization techniques and support vector machines. Inf. Sci. **247**, 62–71 (2013)
19. G.A. Miller, The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol. Rev. **63**, 81–97 (1956)
20. J.L. Elman, Finding structure in time. Cogn. Sci. **14**, 179–211 (1990)
21. H.-K. Yu, A refined fuzzy time-series model for forecasting. Physica A **346**(3/4), 657–681 (2005)
22. Y. Ren, P.N. Suganthan, N. Srikanth, A novel empirical mode decomposition with support vector regression for wind speed forecasting. IEEE Trans. Neural Netw. Learn. Syst. **27**(8), 1–6 (2016)
23. Y. Wan, Y.-W. Si, Adaptive neuro fuzzy inference system for chart pattern matching in financial time series. Appl. Soft Comput. **57**, 1–18 (2017)
24. B.S. Liang, D.L. Cao, *Fuzzy Mathematics and Applications* (Science Press, Beijing, China, 2007)
25. J.C. Dunn, A Graph Theoretic analysis of pattern classification via Tamura's fuzzy relation. IEEE Trans. Syst. Man Cybern. **SMC-4**(3), 310–313 (1974). https://doi.org/10.1109/tsmc.1974.5409141
26. Y.J. Lei, J. Zhao, Y.L. Lu, Y. Wang, Y. Lei, Z.H. Shi, *Theories and Applications of Intuitionistic Fuzzy Set* (Science Press, Beijing, China, 2014)
27. Z.M. Yaseen, M.I. Ghareb, I. Ebtehaj, H. Bonakdari, R. Siddique, S. Heddam, R. Deo, Rainfall pattern forecasting using novel hybrid intelligent model based ANFIS-FFA. Water Resour. Manag. **32**(1), 105–122 (2017). https://doi.org/10.1007/s11269-017-1797-0
28. W. Lu, X.Y. Chen, W. Pedrycz, X.D. Liu, J. Yang, Using interval information granules to improve forecasting in fuzzy time series. Int. J. Approx. Reason. **57**(1), 1–18 (2015)
29. L.A. Zadeh, Fuzzy sets. Inf. Control **8**(3), 338–353 (1965)
30. D. Parbat, O. Chakraborty, A Python based support vector regression model for prediction of Covid19 cases in India. Chaos Solitons Fractals 109942 (2020). https://doi.org/10.1016/j.chaos.2020.109942
31. N. Poonia, S. Azad, Short-term forecasts of COVID-19 spread across Indian states until 1 May 2020, arXiv preprint arXiv:2004.13538, https://arxiv.org/abs/2004.13538
32. S. Panigrahi, H.S. Behera, A study on leading machine learning techniques for high order fuzzy time series forecasting. Eng. Appl. Artif. Intell. **87**, 103245 (2020)
33. S.-M. Chen, S.-W. Chen, Fuzzy forecasting based on two factors second-order fuzzy-trend logical relationship groups and the probabilities of trends of fuzzy logical relationships. IEEE Trans. Cybern. **45**(3), 405–417 (2015)
34. B.P. Joshi, S. Kumar, Intuitionistic fuzzy sets based method for fuzzy time series forecasting. Cybern. Syst. **43**(1), 34–47 (2012)
35. S.-M. Chen, H.-P. Chu, T.-W. Sheu, TAIEX forecasting using fuzzy time series and automatically generated weights of multiple factors. IEEE Trans. Syst. Man Cybern. Part A: Syst. Hum. **42**(6), 1485–1495 (2012)
36. T. Zhan, F. Xiao. A fast evidential approach for stock forecasting. arXiv:2104.05204, https://arxiv.org/abs/2104.05204

Genetic Algorithm Application on 3D Pipe Routing: A Review

Vivechana Maan and Aruna Malik

Abstract The problem of routing paths has been studied systematically for several years now. The problem is quite deterministic when there are only one start and ending point, and there are many well-known pathfinding algorithms like Dijkstra's, A*, Maze Algorithm with their improved versions are there for decades to solve the problem. But for these algorithms, both efficiency and correctness fall terribly when applied on more than two connection points (pair of connection points). To overcome these problems, we use a genetic algorithm. A Genetic Algorithm (GA) is an evolutionary-based biologically inspired technique. It tries to search global minimal from the overall solution by slowly improving over the previous solution. GA has already proved useful to solve many complex engineering problems. GA is mostly used for searching and optimization-related problems [1], where finding an exact solution is somewhere difficult. This is a heuristic approach that yields better solutions. The application of the GA is immense and it is widely used in the industries. One such application of GA is the installment of pipe in three-dimensional space. There are many more related problems like finding the shortest path in a network and the design of a circuit. The complexity of each of the problems is similar to the traveling salesman problem. Which we already know is an np-hard problem that means it's hard to find an exact solution and requires a heuristic algorithm to get a reasonable solution.

Keywords Genetic algorithm · np-hard · Route optimization · Maze algorithm · Optimization algorithm

1 Introduction

Nowadays, wireless networks play a significant role in data communication where data can be transfer from one place to another place. Wireless technologies are using widely for various applications such as precision agriculture, military surveillance,

V. Maan (✉) · A. Malik
Dr. B.R Ambedkar National Institute of Technology, Jalandhar, Punjab, India
e-mail: malika@nitj.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_11

healthcare systems, automation in industries, etc., due to the advancement in computing devices and their capacities [2]. Thus, in wireless networks, there is a need to get/collected data effectively. Some routing strategies help in collecting data in an effective manner such as distance vector routing, link-state routing, 3D pipeline [3].

The 3D pipe installment problem involves many inlets and outlet pairs (I1, [O1, O2]), (I2,[O3]), ... We have to install pipes optimally in all the inlets and outlets by satisfying all the constraints. The pipes can have several branches in them with different sizes of pipe diameters. The 3D space can be full of obstacles.

Traditionally, the path routing algorithm has various applications from traffic optimization to circuit design. Finding a way to optimize the overall path respecting all the constraints helps industries to reduce cost and efforts. There have been many heuristics approaches been used to solve the problem. One such approach is the GA.

A GA is an evolutionary-based biologically inspired technique. It tries to search global minimal from the overall solution by slow improvement over the previous solution. GA has been already proved useful to solve many complex engineering problems. GA is mostly used for searching and optimization-related problems, where finding an exact solution is somewhere difficult to find.

The application of GAs is immense and its widely used in the industries. There is much research has been done on the application of the GA. For the goal of this paper, we discuss path routing through GA with the help of papers revolving around its application. There is many more related problem like finding the shortest path in network and design of a circuit. The complexity of each of the problems is similar to the traveling salesman problem. Which we already know is an np-hard problem that means it's hard to find an exact solution and requires a heuristic algorithm to get a reasonable solution.

From the past, routing problems is widely studied because of their various application in shipping industries, from aero design to large-scale expensive circuit design. The application to this problem is many but due to the nature of the problem and the sheer number of possibility to design a layout which is both easy to implement and optimal is a very challenging and time-consuming job for even an experienced person. Therefore, it is important to have an automatic algorithmic-driven routing layout method.

This review paper aims to talk about the application of the GA in routing. We focus on several ways in which a GA is been used to find the route for the given inlet and outlet. There are several variations like single inlet and outlet, multiple inlet and outlet, branching in routing, and so on. In this paper, we try to understand what different techniques are used to solve the problem (Fig. 1).

The rest of the paper is organized as follows: Sect. 2 contains operators and parameters used for evaluation of the GA, Sect. 3 explains the GA, Sect. 4 is the literature review of what has been till now, and finally the conclusion of the paper.

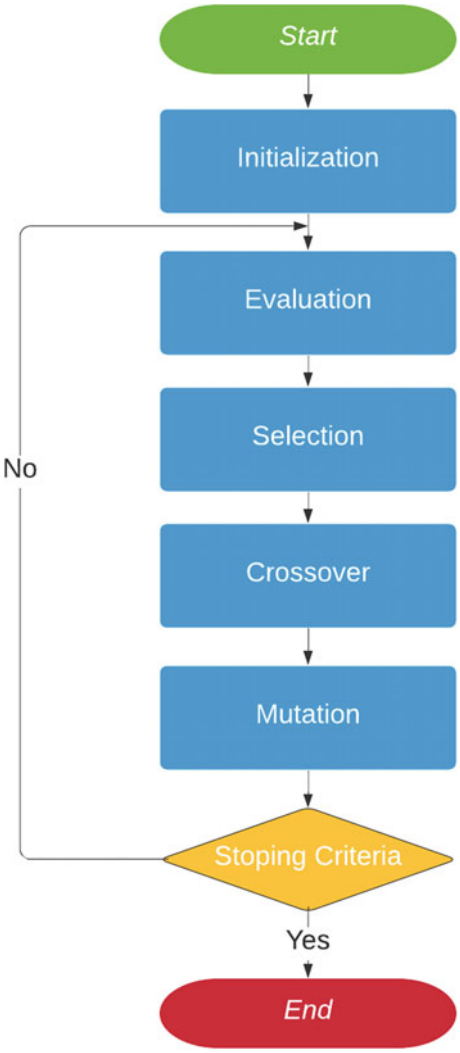


Fig. 1 Flow diagram for genetic algorithm

2 Definition and Parameters Used

1. *Selection Operator:* [4] In the evolutionary algorithm, we use the survival of the fittest theory. We give preference to the fitter individual to pass their genes in the successive generation.
2. *Crossover operator:* [4] The crossover represent mating between parents. We choose two individuals completely at random from the pool and the crossover point is chosen. This crossover point can be single or multiple. Then the sites are exchanged and completely new offspring is produced.

3. *Mutation Operator:* To avoid premature convergence and introduce diversity to the population, we randomly change a certain part of the individual. It is an operation that makes sure there is a variation in genetic encoding from one generation of the population to the next generation of the population. The mutation alters a few positions of the chromosome in every generation to maintain slight variation in children from the parent solution.
4. *Fitness Function:* It is defined as the function which takes the solution as an input and output a quantitative analytical value of the input which represents how good the solution is.
5. *Pareto Front:* The Pareto front is a set of non-dominated solutions where no solution is better than the other, i.e., no one parameter can be improved without at least sacrificing one or more parameters.

3 Generalized Genetic Algorithm

The GA as a core is divided into five stages of work:

1. Initial population: As the first step of an algorithm, we need to generate the first set of solutions. Each individual represents the solution to the original problem. These solutions then become the building block of our algorithm. We iterate over this solution by slowly progressing and improving the solution little by little at the time. Now the key point here is the solution has to be diverse and should represent the overall solution. The more the diverse initial population is the better a chance we reach the global optimal solution.
2. Fitness function: This function takes population as input and tells the fitness score of each of the chromosomes. Now the fitness score is dependent on what we need our solution to be. Based on all the constraints it to map, the fitness score to each chromosome which is the identifier of how good the chromosome solution is. Based on the fitness score the probability of getting a chance to reproduce also increases.
3. Selection: The idea between doing the selection is choosing two-parent and pass their genes to the next generation. Now based on all the fitness scores, we get there are several ways of selecting the parent. In general, we give more chances to the fitter chromosome this thing is known as survival of the fittest. We assign a probability to each chromosome based on the fitness score and then based on probability we randomly choose parents and do crossover to produce their offsprings. This method works under the assumption that the best solution is around the fitter solution.
4. Crossover: It is the most vital part of the algorithm. The crossover point is chosen at random from the selected parents and the genes are exchanged from them so to create their offsprings which contain genes of both the parents. Now there children are added to next generation.

5. Mutation: Mutation is slight variation in one or two genes of children population to maintain the diversity in the population and not letting solution converging prematurely and ending up with local minima/maxima instead of global minima/maxima. The mutation is done with a lower probability. This process is repeated until the desired solution is achieved.

4 Literature Survey

The algorithms to find the shortest path between two points are not new ones; it has been studied for decades, now the [5] Dijkstra's algorithm, to find the shortest path between points. A new improved version of Dijkstra A^* has been created by [6] Hart et al. The most common pathfinding maze algorithm was developed by lee [7]. The algorithm is good for small problems but it is not very effective for large solutions. It does not deal with multiple constraints. And does not perform well when there are multiple starting and ending points. Later, there has been a lot of work and improvement on it from [8, 9], but they were not very efficient and could not guarantee an optimal solution. In the recent studies, the GS approach has been taken place in [10] but it was not done for multi-branch. Then in [11], the multi-branch approach has been taking place with one point crossover with a maze algorithm for generating the initial population. But in a practical scenario, it lacks efficiency. The solution became constant and no extreme possibility are been checked. It cannot deal with constraints like serviceability which is just an important practical problem.

Many solutions had been proposed for the path routing problem using various algorithms. Here, we discuss some of the GA-based approaches for pipe routing problems to optimize the design of the pipe routes in an environment using certain constraints.

In 1996 [12], GA approach is used to minimize the length of pipe where interconnections are pre-specified by taking obstacles into account. The results of this paper suggest that GA is superior to either simulated annealing or hill climbing on these kinds of problems.

In 1999 [13], a non-deterministic approach was proposed based on a GA to generate sets of pipe routing with good searching efficiency. Here, .STL files are used to represent the obstacles and to show distinctive advantages from them. This approach is useful to optimize the total length and number of bends of connecting pipes while ignoring the obstacles.

In 2005 [14], GA approach is used for the minimization of network cost in a practical environment. The motive is to find optimal pipe design with insufficient time and limited resources. Here, gray coding is combined with the improved GA with the help of elitist strategy, which means based on fitness degree of offspring. The findings of this paper show that the proposed method is better than the enumeration techniques concerning solution cost and speed.

In 2006 [15], 3D multi-pipe route optimization is done with the help of a generalized GA. The pipe route has been coded into a string using multiple variables

and constrain. The algorithm has been claimed to work better than the traditional pathfinding algorithm for most of the cases. They haven't talked about the multi-branch pipe.

In 2009 [16], they have used a weighted cost-based system. The layers are been weighted to show the cost. The efficiency of the GA has been compared with the other two, the existing routes and least cost routes. The GA has been proven to be 20% cost-effective. Results showed that least-cost routes and GAs were producing the same results for most parts.

In 2011 [17], a GA approach is used to design pipe path routes in a coal-fired boiler. Here, virtual prohibited cells are introduced in search space and a GA-based searching method is used to find multiple paths by avoiding these cells. Virtual prohibited cells are generated with the help of GA to prevent these cells from the random allocation, a sharing method [18, 19] is used in the fitness function to generate these cells. The comparative result shows that the number of patterns with sharing method is 2.5 times more than without sharing method.

In 2015 [20], a method-based GA was proposed where the main constraint is distance. The motive is to find a suitable path for pipes that is collision-free. This path is either optimal or near to optimal concerning length, the number of bends, and layout near or away from something. For the distance constraint of the problem, an effective evaluation method is developed for the proposed GA. To evaluate the distance constraint, two evaluation functions are added to the evaluation.

In 2016, [21] GA is used with the combination of NSGA II (Non-Dominated Sorting algorithm II) and CCNSGA (cooperative Co-evolutionary non-dominated sorting genetic algorithm II) along with rat-maze pathfinding algorithm to solve the 3D multi-pipe route optimization. The paper has used fixed-length encoding along with adaptive region strategy.

In 2020 [22], 3D pipe routing algorithm was proposed by combining adaptive A^* algorithm with GA. They showed that it improves the quality of the solution. Also with simulation showed that Adaptive A* with GA works better than Adaptive A^* and A^* algorithm (Table 1).

5 Future Scope

Future work can be on to diversify the population using different kinds of pathfinding algorithms. By doing so, we take benefits of all the pathfinding algorithms which help to maintain diversity in the population even further. This diversified population increases the probability to get the optimal solution.

Table 1 Comparative study of previous paper

S. no.	Author	Year	Paper title	Proposed work	Limitation
1	Kim et al. [12]	1996	Industrial plant pipe-route optimization with genetic algorithms	GA is used to minimize the length of pipe. Results show that GA is superior then simulated annealing or hill climbing in such problems	The paper focuses on an array of multiple single inlet and outlet pairs. It does not talk about the single inlet and multiple outlets or vice versa
2	Sandurkar et al. [13]	1999	GAPRUS-genetic algorithms-based pipe routing using tessellated objects	Non-deterministic approach was proposed based on a GA to generate sets of pipe routing with good searching efficiency	The paper does not consider real-life constraints such as the variable diameter of the pipe
3	Shau et al. [14]	2005	Genetic algorithms for the design of pipe network systems	Gray coding is combined with the improved genetic algorithm with the help of elitist strategy, which means based on the fitness degree of offspring	It does not talk about multi-pipe and optimization
4	Wang et al. [15]	2006	Three-dimensional multi-pipe route optimization based on genetic algorithms	Generalized genetic algorithm is used for 3D multi-pipe route optimization. The pipe route has been coded into the string using multiple variables and constrain	The paper does not focus on the complexity of multi-pipe or multi-constrain. It only focuses on path optimization
5	Ebrahimipo et al. [16]	2009	Routing of water pipeline using GIS and genetic algorithm	A weighted cost-based system is used. The efficiency of GA is compared with the other two existing routes. GA has been proven 20% cost-effective	It is difficult to have the GIS (Geographic Information System)

(continued)

Table 1 (continued)

S. no.	Author	Year	Paper title	Proposed work	Limitation
6	Kusumi et al. [17]	2011	Genetic Algorithm-Based Searching Method for Piping Path Routing in Coal-Fired Boiler Buildings	Virtual prohibited cells are introduced in search space and a GA-based searching method is used to find multiple paths by avoiding these cells	The order in which the path has been laid becomes an important factor
7	Wang et al. [20]	2015	A method based on genetic algorithm for pipe routing design	For distance constraint of problem, an effective evaluation method is developed for proposed GA	They have not considered branching in pipes
8	Niu et al. [21]	2016	Ship pipe routing design using NSGA-II and coevolutionary algorithm	To solve 3D pipe route optimization GA is used with a combination of NSGA II and CC-NSGA along with rat-maze pathfinding algorithm	The use of adaptive region may lead to a longer path
9	Lv et al. [22]	2020	Pipe routing of reactor based on adaptive A* algorithm combined with genetic algorithm	3D pipe routing algorithm was proposed by combining adaptive A* algorithm with GA	Using an A* path algorithm reduces the diversity from the initial population

6 Conclusion

In this paper, we provided a detailed overview of 3D pipe routing techniques based on a genetic algorithm. We see that how with the help of GA we can make the pipe routing modal time-efficient, cost-efficient with limited resources. With the help of GA, we achieved near to optimal solutions by considering different parameters. Here, we discuss the basic functionality of the genetic algorithm and see its workflow. With the help of GA, we can fully automate the pipe routing process which is manually very time-consuming, it's not possible to explore all routes manually present in a given domain, but with GA, we can search all solutions present for pipe routes and pick the best suited for the given situation. Modification in GA can be done by merging different algorithms with it which can help to yield a better solution.

References

1. Y. Kumar, P. Singh, Improved cat swarm optimization algorithm for solving global optimization problems and its application to clustering. Appl. Intell. **48**, 2681–2697 (2017)
2. P.K. Singh, Y. Singh, M.H. Kolekar, A.K. Kar, J.K. Chhabra, A. Sen, *Recent Innovations in Computing, Proceedings of ICRIC 2020*. Lecture Notes in Electrical Engineering Book Series (LNEE), vol. 701 (Springer, 2020)
3. P.K. Singh, A.K., Kar, Y. Singh, M.H. Kolekar, S. Tanwar, *Recent Innovations in Computing, Proceedings of ICRIC 2029*. Lecture Notes in Electrical Engineering Book Series (LNEE), vol. 597 (Springer, 2019)
4. M. Mitchell, *An Introduction to Genetic Algorithms* (MIT Press, 1998)
5. E.W. Dijkstra, A note on two problems in connexion with graphs. Numer. Math. **1**(1), 269–271 (1959)
6. P.E. Hart, N.J. Nilsson, B. Raphael, A formal basis for the heuristic determination of minimum cost paths. IEEE Trans. Syst. Sci. Cybern. **4**(2), 100–107 (1968)
7. C.Y. Lee, An algorithm for path connections and its applications. IRE Trans. Electron. Comput. **10**(3), 346–365 (1961)
8. D.W. Hightower, A solution to line routing problems on the continuous plane," Papers on Twenty-five years of electronic design automation - 25 years of DAC (1988)
9. Y. Kobayashi, K. Yoshida, T. Yoshinaga, A knowledge-based approach to verification and improvement of industrial plant layout design. Expert Syst. Eng. Appl. 179–189 (1993)
10. B. Wanamaker, T. Cascino, V. McLaughlin, H. Oral, R. Latchamsetty, K.C. Siontis, Atrial arrhythmias in pulmonary hypertension: pathogenesis, prognosis, and management. Arrhythmia Electrophysiol. Rev. **7**(1), 43 (2018)
11. W. Niu, H. Sui, Y. Niu, K. Cai, W. Gao, Ship pipe routing design using NSGA-II and coevolutionary algorithm. Math. Probl. Eng. **2016**, 1–21 (2016)
12. D.G. Kim, D. Corne, P. Ross, Industrial plant pipe-route optimisation with genetic algorithms, in *International Conference on Parallel Problem Solving from Nature* (Springer, Berlin, 1996), pp. 1012–1021
13. S. Sandurkar, W. Chen, GAPRUS-genetic algorithms based pipe routing using tessellated objects. Comput. Ind. **38**(3), 209–223 (1999)
14. H-M. Shau, B.-L. Lin, W.-C. Huang, Genetic algorithms for the design of pipe network systems. J. Marine Sci. Technol. **13**(2), 116–124 (2005)
15. H. Wang, C. Zhao, W. Yan, X. Feng, Three-dimensional multi-pipe route optimization based on genetic algorithms, in *International Conference on Programming Languages for Manufacturing* (Springer, Boston, 2006), pp. 177–183
16. A.R. Ebrahimipo, A. Alimohamad, A.A. Alesheikh, H. Aghighi, Routing of water pipeline using GIS and genetic algorithm. J. Appl. Sci. **9**(23), 4137–4145 (2009)
17. N. Kusumi, D.E. Goldberg, N. Ichinose, Genetic algorithm-based searching method for piping path routing in coal-fired boiler buildings, in *ASME Power Conference*, vol. 44601 (2011), pp. 281–286
18. D. E. Goldberg, J. Richardson, Genetic algorithms with sharing for multimodal function optimization, in *Genetic algorithms and their applications: Proceedings of the Second International Conference on Genetic Algorithms* (Lawrence Erlbaum, Hillsdale, 1987), pp. 41–49
19. K. Deb, D.E. Goldberg, An investigation of niche and species formation in genetic function optimization, in *Proceedings of the Third International Conference on Genetic Algorithms* (1989), pp. 42–50
20. C. Wang, X. Sun, T. Yuan, A method based genetic algorithm for pipe routing design, in *2015 International Conference on Advanced Engineering Materials and Technology* (Atlantis Press, 2015), pp. 826–830

21. W. Niu, H. Sui, Y. Niu, K. Cai, W. Gao, Ship pipe routing design using NSGA-II and coevolutionary algorithm. Math. Probl. Eng. (2016)
22. W. Lv, N. Qin, X. Zhao, P. Yuan, J. Huang, Pipe routing of reactor based on adaptive A* algorithm combined with genetic algorithm, in *2020 Chinese Automation Congress (CAC)* (IEEE, 2020), pp. 5567–5572

Directed Undersampling Using Active Learning for Particle Identification

Zakarya Farou, Sofiane Ouaari, Balint Domian, and Tomáš Horváth

Abstract Accurate particle identification is an ongoing task in the European organization for nuclear research, known as *CERN* where the challenge remains that targeted particles/events represent tiny minorities in front of the overwhelming presence of common particles such as protons. This paper presents a directed undersampling using an active learning method named DUAL to handle the high imbalance problem present in the particle identification dataset. The proposed approach was used to reduce the training set size while maintaining classifiers' performance. Compared against various imbalance learning approaches, the experimental results show that using DUAL as a data reduction technique with a random forest classifier enhances classification performance in terms of *Macro-*F_1 score and decreases the training time needed to train the models, which is very relevant while dealing with large-scale datasets. Despite being experimented only with particle identification dataset, we believe that DUAL could be adopted as a generic method for multi-class imbalanced classification problems with big data scale difficulties.

Keywords Active learning · Directed undersampling · Imbalance learning · Multi-class classification · Particle identification

Z. Farou (✉) · S. Ouaari · B. Domian · T. Horváth
Department of Data Science and Engineering, ELTE Eötvös Loránd University,
Pázmány Péter Sétány 1/C, 1117 Budapest, Hungary
e-mail: zakaryafarou@inf.elte.hu
URL: http://t-labs.elte.hu/

S. Ouaari
e-mail: dfmyzk@inf.elte.hu

B. Domian
e-mail: bu1t3e@inf.elte.hu

T. Horváth
e-mail: tomas.horvath@inf.elte.hu

T. Horváth
Institute of Computer Science, Pavol Jozef Šafárik University,
Jesenná 5, 040 01 Košice, Slovakia

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_12

1 Introduction

CERN with its different experimental labs (LHC, CMS, LHCb, …), is the core of the research made to learn more about minor constituents of matter, their interactions and the origin and evolution of the universe. The large hadron collider (LHC) [11] is the world's largest and most powerful particle accelerator. High-energy particle beams travel at the speed of light, and their collisions produce tons of generated particle showers. Ninety petabytes of data is generated every year by LHC, and this showed the synergies between the respective domains of experimental high-energy physics and data science ecosystem.

The key challenge in this area is overcoming the high imbalance ratio present since targeted particles, or decay types are rare events in front of the overwhelming presence of ordinary particles like *pion* and *electrons* [17], which high-energy physicists and researchers usually describe as background events [1]. However, most of the researches tackled binary classification type where the studied particle or event (Higgs boson particle decay to two leptons) is labelled as a positive signal, and the background as a negative signal [19]. Also, as stated earlier, the amount of particle shower, hence the generated data, is tremendously immense, leading to computationally expensive model training.

This study aims to overcome the data imbalance present in the multi-particles classification task (identification of four-particle types i.e. *pion*, *proton*, *kaon* and *positron*) by proposing a directed undersampling method based on active learning called DUAL. DUAL is considered as an internal approach (pre-processing method) applied before starting the classification process. We use it to undersample and filter a large dataset while only keeping the most critical and significant data point, directly impacting the model's performance and achieving faster analysis and training time. The motivation behind this active learning approach comes from [25] where the authors investigated an innovative active learning approach utilised for binary imbalanced datasets with a modest number of instances ($D_{size} \leq 20.000$). Thus, DUAL aims to generalise the usage of active learning for multi-class imbalanced problems and large-scale datasets. Note that diverse strategies for imbalanced classification based on active learning have been proposed [5]. However, these approaches are concentrated on SVM learning systems and are based on the fact that the most informative examples are the ones closest to the hyperplane for this type of learners. To our knowledge, the use of AL with a Random Forest learner as a directed undersampling method concerning particles identification has not yet been performed.

The paper is structured as follows: Sect. 2 explains various concepts related to imbalance learning, active learning and highlights previous works related to particle identification. In Sect. 3, we describe the proposed DUAL method. The detailed description regarding the used data set, performed experiments, results and discussion are incorporated in Sect. 4. Lastly, Sect. 5 outlines the conclusion about the conducted research and the potential research direction further to improve learning and classification with class imbalance problems.

2 Background

2.1 Particle Identification

Various machine learning (ML) approaches are highly exploited in particle identification problems. In [19], they used multiple neural network architectures to detect the decay of the Higgs boson. In [8], deep neural networks (DNN) and convolutional neural networks (CNN) were employed to distinguish among signals from electrons, photons, hadronic backgrounds and measure particle energies. They also simulated samples of individual electron, photon, charged hadron and neutral hadron images in a simple high-granularity calorimeter detector implemented with the Geant4 simulation toolkit [3]. In [23], the authors provided the implementation of a network model to detect high-momentum Higgs bosons decaying to bottom quark–antiquark pairs. The interaction model network is based on two input collections comprising Np particles, each represented by a feature vector of length P and Nv vertices, each represented by a feature vector of length S. Another interesting work [16] shows how particle identification has been done with a transition radiation detector in ALICE [2].

2.2 Imbalance Learning

Imbalance learning is a non-standard derivative data science problem [13]. Derivative because it is an extension of core problems, i.e. classification problems. Non-standard, since the data has an unusual distribution on the target variable.

$$D = \sum_{i=1}^{n} C_i^- + \sum_{j=1}^{m} C_j^+; L = n + m \tag{1}$$

As shown in Eq. 1, in multi-class imbalanced classification a dataset D of L classes have n majority classes C_i^- and m minority classes C_j^+ [21]. In such a scenario, supervised models are biased towards the dominant, i.e. majority classes, yielding inaccurate and misleading results.

The imbalance problem can arise due to the nature of the application at hand. As mentioned earlier, in high-energy physics and astroparticle domains, *pion*, *proton* and *electron* particles are available in bulk compared to other particle types. As a result, they belong to the set of majority classes C_i^- while the remaining ones (*kaon* and *positron* in our case) are considered as minority classes C_j^+. Besides the high-energy physics domain, the imbalance problem is highly present in medical and biomedical domains [18, 24] as the data collection process is time-consuming, requires expert knowledge for data annotation and faces many privacy constraints. Diverse techniques have been proposed to overcome the problem mentioned above [12]. These techniques aim to enhance supervised models sensitivity towards data by inferring

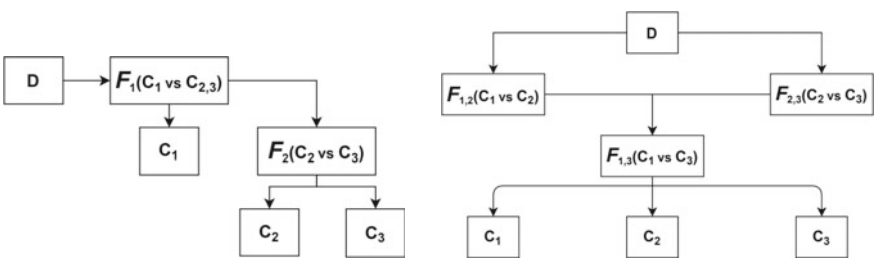
rules that form a good decision boundary that segregates well negative classes C_i^- from positive ones C_j^+. We can divide the techniques into two subsets which are as the following:

2.2.1 Decomposition-Based Approaches

The original problem is reduced to a set of two-class subsets, then it can be directly solved by one of existing techniques for the binary imbalanced scenarios. We distinguish two types of decomposition techniques, one-versus-all (OVA) and one-versus-one (OVO). As illustrated in Fig. 1a, OVA [13] tries to set $F-1$ classifiers for L classes. Initially, we elect a positive class C^+ from the available L classes, whereas the remaining $L-1$ classes form a superset (considered as the set of negative classes C^-). Then, we train the F_ith classifier with both positive and negative sets. After that, we omit C^+ and substitute it with another class from the constructed superset C^-. We follow identical steps until both sets, i.e. positive and negative sets, have only a single class. However, the OVO [13] algorithm aims to build $F \times (F-1)/2$ binary classifiers to classify samples from L classes (Fig. 1b). It uses the voting policy, i.e. given an instance with an unknown class, the class with the most votes is identified as sample's label.

2.2.2 Resampling-Based Approaches

Resampling approaches preprocess the training data either by adding more minority observations or reducing the size of the majority class. Resampling methods can be further categorized into random sampling approaches and directed sampling approaches. Random sampling approaches are non-heuristic methods that aim to balance class distribution through the random replication of minority class examples (random oversampling, i.e. ROS) or the random dismissal of majority class instances (random oversampling, i.e. RUS). Various studies acknowledge that ROS can raise the likelihood of occurring overfitting. In contrast, the significant drawback of RUS



(a) OVA classification method (b) OVO classification method

Fig. 1 Decomposition-based approaches

is that this method can discard potentially valuable data that could be important for the induction process.

Compared to random sampling methods (ROS and RUS), in directed sampling, the choice of samples to add/eliminate is decided based on some specific criteria rather than randomly chosen.

2.3 Active Learning

Active learning (AL) [7, 10] is a semi-supervised approach in which the learning algorithm can interactively gather information from the user. Mainly applied with unlabelled data, we could also utilise AL when all class labels are known. In this case, the active learning strategy affords the ability to actively selecting the best, i.e. the most informative samples.

3 Directed Undersampling Using Active Learning

Most ML models such as DNN and CNN require a considerable amount of labelled data to train. This requirement injects a burden considering the labelling cost and training time.

The first step of the process is generating a random subset from the original training data, which can be manually labelled. Next, we train an active learner on the generated data, which will allow the active learner to predict the labels of the remaining unlabelled data element based; this process is referred to as the *data pool*. The predicted labels are ranked based on a selected query score where the process usually prioritizes the data points to which the active learner gives an uncertain prediction. The least confidence or uncertainty sampling (See Eq. 2) is one of the most frequent query strategies.

For each element of the unlabelled dataset, the active learner gives a class probability prediction (probabilities for each class). Out of those classes, only the one with the highest probability is selected (referred to as $\hat{y}$). Where $\hat{y} = \text{argmax}_y P_\theta(y|x)$ or the class label with the highest posterior probability under the model. This probability is then subtracted from 1 to return the highest uncertainty. Finally, the algorithm picks that data point from all x elements where this value is the highest, meaning x_{LC} contains the (index of the) x element where the model is the most uncertain, i.e. The data points reaching the maximum uncertainty probability values. The algorithm operates for a fixed number of epochs, considered in this case as a hyper-parameter.

$$x_{LC} = \text{argmax}_x \; 1 - P_\theta(\hat{y}|x) \tag{2}$$

For example, in a classification problem with three different possible outputs, an element with prediction probabilities [0.4, 0.4, 0.2] is preferred over an element with

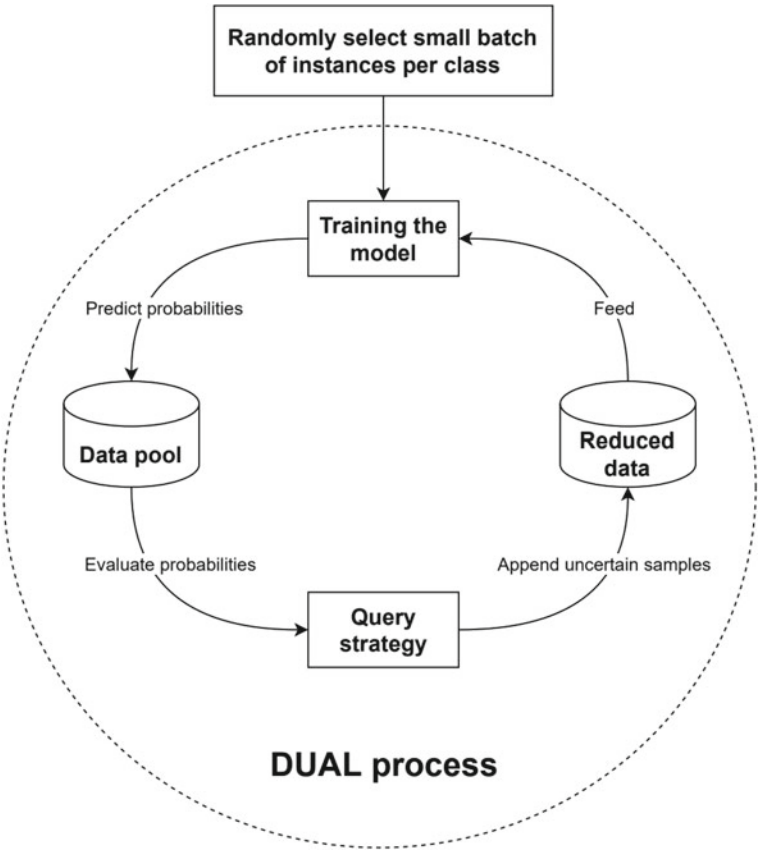


Fig. 2 DUAL process

prediction scores of [0.8, 0.1, 0.1], because the first has 0.6 (1 – 0.4) uncertainty and the other has 0.2 (1 – 0.8). After acquiring the desired number of top-scoring elements from the *data pool*, an oracle labels them, and we retain the active learner with the extended data.

Preprocessing the data using an undersampling method can notably decrease the models' training time as the training set has fewer instances than initially; however, randomly doing it degrades the model's performance in most cases. We recommend using active learning as a directed undersampling method, proven by our results that it can converge to the same or even better performance as the training the model on the complete data. Figure 2 is the schema describing the general process of the proposed DUAL.

Using an active learner with a massive, labelled dataset follows the same procedure as explained before. However, as we are aware of each element's labels, we initialize the starting subset where each class has the same amount of randomly selected

elements. After identifying the elements by the query strategy, we also modify the oracle labelling process since the *data pool* contain genuine labels. Meaning that the oracle, in this case, is a simple query for the correct targets from the pool, and there is no need to involve domain experts to label the data.

4 Experiments, Results and Discussion

4.1 Dataset Description

Geant4 [3], is the world-standard software toolkit for high-energy physics detector simulation, and it intends to reproduce the results of a possibly observable process by modelling its component and behaviour. Our data[1] comes from a simplified dataset of a GEANT based simulation of electron-proton inelastic scattering measured by a particle detector system [6]. The dataset features are the values of six detector responses being recorded (momentum p, θ, β, number of photoelectrons $nphe$, inner ein and outer energy $eout$). p is a vector quantity possessing a magnitude and a direction and defined as the product of the mass and velocity of an object. **c** being the speed of light, and Giga-electron Volt (GeV) is a unit used for measuring the energy of subatomic particles. θ is the particle scattering angle in rad, while β is roughly the width of the beam squared divided by the emittance. If β is low (resp. high), the beam is narrower and squeezed (resp. wide and straight). Another feature concerns the number of the photoelectric effect, a phenomenon in which electrically charged particles are released from or within a material when it absorbs electromagnetic radiation. The effect is often defined as the ejection of electrons from a metal plate when light falls on it. The last two features are ein and $eout$ given by them GeV value.

The dataset D have four classes C_{pion}, C_{proton}, C_{kaon}, $C_{positron}$, each one representing a specific particle type. Before starting the classification process, it is essential to pre-process the data. As a result, we have scaled the data using *Standard Scaling* defined by Eq. 3 where μ and σ represent the mean and standard deviation respectively.

$$x_{norm} = \frac{x - \mu}{\sigma} \tag{3}$$

Further pre-processing steps involve the stratified splitting of the data into three disjoint subsets, i.e. training, validation, and testing sets with the following proportions, respectively 75, 5 and 20%. We use the training set to train various models, the validation set to tune the hyper-parameters and select the best model, and the test set to evaluate the performance of the chosen model based on some evaluation metrics.

As the used dataset is highly imbalanced (see Table 1), i.e. D have an imbalanced classes distribution, we use imbalance ratio (IR) to give more details about the pro-

[1] https://www.kaggle.com/naharrison/particle-identification-from-detector-responses.

Table 1 Data distribution before and after using the proposed DUAL

C_{pion}	C_{proton}	C_{kaon}	$C_{positron}$	C_{total}
Original data distribution				
2105124	1459387	174353	11135	3749999
Data distribution after using DUAL				
45557	39081	23321	2791	110750

portion of the number of instances in C^- to the number of instances in C^+. For our case, we propose to replace C^- (resp. C^+) by a superset C^-_{super} (resp. C^+_{super}) constructed from the union of the majority classes (resp. the minority classes). Therefore, we compute IR using Eq. 4. In our case, $IR = 19.22$.

$$IR = \frac{|C^-_{super}|}{|C^+_{super}|} = \frac{\sum_{i=1}^{n} |C^-_i|}{\sum_{j=1}^{m} |C^+_j|} \tag{4}$$

4.2 Methods for Comparison

For the experimental part, we have used many variants of tree-based ML models ranging from *Decision Trees* (DT) to ensemble ones utilizing Random Forest (RF) as *Bagging* method and *Boosting*[2] techniques. We used the models' implementation provided by *sklearn* library.[3] Furthermore, we include four categories of methods that include nine methods used as baselines than the data reduction using directed sampling, i.e. our proposed method DUAL. For instance, we used the output of the selected supervised methods on the original training data (OD) with its full size, ROS and RUS explained in Sect. 2.2.2 as random sampling methods. Additionally, we used synthetic minority oversampling technique (SMOTE) [9] and partially guided oversampling (PGO) [20] as directed oversampling methods, instance hardness threshold (IHT) [22] and edited nearest neighbours (ENN) [4] as directed undersampling methods. The last category is related to decomposition-based approaches (OVA and OVO explained in Sect. 2.2.1).

For selecting the active learner, we took the following steps, after evaluating the different ML models on the original dataset (See the first part of Table 3), we looked at the two best performing models. Since RF gave the second-best results and had half the training time of the XGB, we selected RF as the active learner. Our choice for the random forest is also related to the *bagging* ensemble nature of the model as it tends

[2] Extreme Gradient Boosting (XGB), Ada Boost(ADB), Gradient Boosting (GB).

[3] https://scikit-learn.org/stable/supervised_learning.html.

to reduce variance, i.e. selecting data points for the reduced dataset without falling into overfitting. We first initialize the batch by taking 500 instances for each class, then for 250 epochs, we query 750 instances from the remaining batch with uncertainty sampling, also known as $least\ confidence$ (See Eq. 2), saving the best training sets. The active learning process took 31 min to execute; the proposed DUAL reduced the dataset size by 98% compared to the original dataset. Consequently, the original IR was reduced from 19.22 to 3.24. Table 1 provides the new data distribution after using DUAL is provided. Note that $C_{total} = C_{pion} + C_{proton} + C_{kaon} + C_{positron}$.

4.3 Evaluation Metric and Hardware Specifications

We utilise the *Macro-*F_1 score as an evaluation metric [14] to measure the performance of all the methods. *Macro-*F_1 is a conventional metric applied to evaluate classification decisions; compared to *Micro-*F_1 which gives equal importance to all instances in the averaging process, *Macro-*F_1 assigns equal weight to each class-label. We compute *Macro-*F_1 using Eq. 7.

$$P_l = \frac{TP_l}{TP_l + FP_l} \tag{5}$$

$$R_l = \frac{TP_l}{TP_l + FN_l} \tag{6}$$

$$Macro\text{-}F_1 = \frac{1}{|L|} \sum_{l \in L} \frac{2P_l R_l}{R_l + R_l} \tag{7}$$

TP_l, FP_l, FN_l denote the true positives, false positives and false negatives for the class-label $l \in L$, while P_l and R_l denotes the $precision$ and $recall$ for the same class-label l.

Table 2 reports the hardware and run-time environment (RTE) used for the conducted experiments. Furthermore, we used parallel computing with a fixed number of workers set to 14 to estimate the training time achieved by each model.

Table 2 Hardware and software environment specifications

Specification	Description
Processor	AMD Ryzen 7 3700X 8-Core Processor 3.60 GHz
Memory	G.Skill RipjawsV DDR4 32GB (2x16 GB) 3600 MHz
Storage	Samsung 970 EVO Plus SSD, 500GB, NVMe, M.2
Graphics card	EVGA GeForce RTX 3070 XC3 ultra gaming 8GB GDDR6
Operating system	Windows 10 education N 20H2
RTE	Conda python 3.8

4.4 Classification Results

We present two sets of results. The first set of results reported in Table 3 compares the performance of various supervised ML algorithms before and after using DUAL without conducting any hyper-parameters tuning. The second set of results, i.e. Table 5 presents a pairwise comparison between the picked classifiers, i.e. RF and GB, where others baselines are appended to the comparison previously done. For instance, ROS and RUS, which are random sampling-based approaches, SMOTE, PGO, IHT (where the IHT strategy was *majority* as the *not minority* strategy gave even worse results), and ENN as directed over/undersampling approaches, and OVA, OVO decomposition-based approaches. Hyper-parameters tuning for the selected classifiers is performed using grid-search[4] and the chosen hyper-parameter values are given in Table 4. Additionally to the overall performance and the training time expressed by $seconds$, we pay attention to the performances for each class $l \in L$ to ensure that the models are performing well with all dataset classes and are not biased toward the majority classes.

From the results displayed in Table 3, we observe that despite the DT classifier is the fastest prior and posterior data reduction. However, it gives a lower $Macro\text{-}F_1$ score compared to other classifiers. Nevertheless, the training time got diminished drastically for all the used models, which is evident as the use of DUAL on the original training set induces a drop in the training set size (Table 1).

We might think that we could observe such a scenario with any supervised learning methods. However, by investigating a neural network-based architecture with three hidden layers of sizes 256, 128 and 32 respectively and training it only for 20 epochs using the same hardware and software specifications outlined in Table 2, and despite giving reliable performances in term of $Macro\text{-}F_1$ score (≈ 86.6), its training time still very large opposed to other supervised models even after reducing the training set size using DUAL. Initially, it has a training time of 24×10^2 s, which decreased to 8.2×10^2 s.

Additionally, we observe an increase in the performance after reducing the training set size for both RF and GB classifiers. That means that our proposed directed undersampling method (DUAL) is doing its job correctly with these classifiers, i.e. the active learner maintained the performance with fewer training samples and improved the quality of the decision boundaries (more samples are correctly identified for all particle types). These results are clearly visible in Table 5.

After selecting the two qualified classifiers, we performed a grid-search algorithm to tune them hyper-parameters and improve performance. We argue that creating the reduced dataset made it feasible to finish a grid-search inconceivable time (2.25 h) which would have taken probably days on the whole training set. Grid-search results are shown in Table 4.

Table 5 summarizes the results of pairwise comparison between RF and GB. For an informative comparison and to further validate our results, we also incorporate other baseline methods described previously in Sect. 4.2. While we remain to monitor

[4] https://scikit-learn.org/stable/modules/grid_search.html.

Table 3 Comparison of different supervised ML methods prior and posterior data reduction

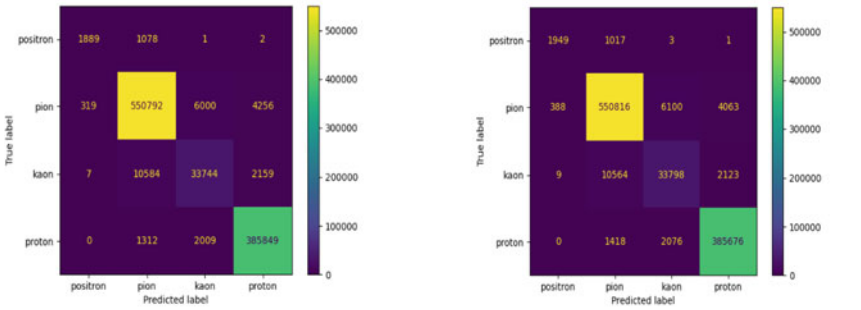
ML model	Macro-$\mathbf{F_1}$	Training time	Macro-$\mathbf{F_1}$	Training time
	OD		DUAL	
ADB	63.14	2.15×10^2	45.98	0.03×10^2
DT	81.43	$\mathbf{0.613 \times 10^2}$	70.44	$\mathbf{0,008 \times 10^2}$
GB	83.67	27.8×10^2	85.45	0.65×10^2
RF	86.35	1.95×10^2	**86.46**	0.024×10^2
XGB	**86.71**	4.07×10^2	86.41	0.11×10^2

Table 4 Results of grid-search for tuning RF and GB hyper-parameters

ML model	Max_depth	Min_samples_split	N_estimators
RF	25	10	300
GB	3	5	175

Table 5 Summary of RF and GB results based on *Macro-*F_1 score and training time

Method	$\mathbf{C_{pion}}$	$\mathbf{C_{kaon}}$	$\mathbf{C_{positron}}$	$\mathbf{C_{proton}}$	Macro-$\mathbf{F_1}$	Training time
RF						
OD	97.90	76.45	72.56	98.75	86.41	4.68×10^2
ROS	97.11	73.18	66.29	98.71	83.82	10.2×10^2
SMOTE	96.40	70.03	51.40	98.66	79.12	13.2×10^2
PGO	96.73	71.04	62.10	98.68	82.13	13.2×10^2
RUS	95.87	65.56	64.64	98.50	81.14	0.39×10^2
IHT	94.71	62.27	57.64	98.64	78.31	3.96×10^2
ENN	97.75	73.85	**74.25**	98.65	86.13	3.9×10^2
OVO	97.90	76.37	72.92	98.76	86.49	11.4×10^2
OVA	97.91	76.46	73.18	98.73	86.56	17×10^2
DUAL	**97.93**	**76.53**	73.34	**98.78**	**86.65**	$\mathbf{0.06 \times 10^2}$
GB						
OD	97.74	75.46	65.08	98.67	84.24	47.4×10^2
ROS	94.77	65.52	27.95	98.43	71.67	103.8×10^2
SMOTE	95.20	67.32	29.84	98.51	72.46	150×10^2
PGO	95.14	67.03	29.77	98.50	72.61	150×10^2
RUS	95.78	65.33	59.08	98.41	79.65	5.88×10^2
IHT	95.09	63.75	55.56	98.58	78.24	39.96×10^2
ENN	97.66	73.40	68.60	98.60	84.56	42.6×10^2
OVO	97.80	75.75	69.56	98.66	84.94	12.6×10^2
OVA	97.77	75.19	64.52	98.65	84.03	12.9×10^2
DUAL	**97.86**	**76.09**	**71.24**	**98.73**	**85.98**	$\mathbf{1.07 \times 10^2}$



(a) Confusion matrix of RF trained with the original training set

(b) Confusion matrix of RF trained with reduced data issued from DUAL

Fig. 3 RF's confusion matrix for particle identification

the training time for all the methods included in the comparison, we pay particular attention to the performances achieved in each class that represent a particular particle type as we know that our dataset is originally imbalanced and considering only the *Macro-*F_1 average is not enough to judge the performance of the models. We see that both classifiers achieved slightly better performance and considerably faster training time compared to the baselines. However, RF is slightly better than GB both in terms of *Macro-*F_1 score and training time.

To further examine the performance of RF on the test set, we adopted the confusion matrix [15]. A *confusion matrix* is a particular table that enables visualization of the performance of a given supervised learning method.

As observed in Fig. 3, expect C_{kaon} where we have a decrease in FP value, all the remaining particle types have an increase in FP values. Another significant remark is related to the minority class $C_{positron}$, where its FP increased by 3.18 % after reducing the training set using the proposed DUAL. Again, this attests that DUAL is an auspicious directed undersampling method that could be successfully used with the multi-class imbalanced problem.

5 Conclusion and Future Works

In this era, a tremendous amount of data is being produced and stored constantly and exponentially. Whether in research or industry, one of the main challenges is that the under analysis classes represent the main centre of interest (minority classes) are known to be in a way smaller proportion than the ordinary events (majority classes). Thus, here we have two issues to handle: data size complexity and high multi-class imbalance ratio. This paper has proposed a directed undersampling approach based on active learning on a classification task that involves both issues mentioned above, this task being particle identification in high-energy physics and astroparticles. We

have shown that by using DUAL, we could reduce the dataset by 98%, allowing domain experts to focus on the most significant instances having a direct impact. In addition, the imbalance ratio was reduced by 83%, while improving, for some models, the performance.

In future works, a potential direction to further investigate is the generic framework of using the directed active learning for undersampling, which could be done using DUAL with other datasets and further comparing it with other sampling-based approaches or data augmentation techniques. We can explore convolutional neural networks (CNN) as active learner or recurrent neural network (RNN) architectures when handling data of sequential nature if we are dealing with image-based data. In addition, we can manipulate and adjust the query strategy to choose the uncertain data points according to the domain's needs hence discover more sophisticated query criteria.

References

1. M. Abbas, A. Khan, A.S. Qureshi, M.W., Khan, Extracting signals of higgs boson from background noise using deep neural networks. arXiv preprint arXiv:2010.08201 (2020)
2. B. Abelev, J. Adam, D. Adamová, M. Aggarwal, G.A. Rinella, M. Agnello, A. Agostinelli, N. Agrawal, Z. Ahammed, N. Ahmad et al., Alice collaboration. Nucl. Phys. A **931**, 1211–1221 (2014)
3. S. Agostinelli, J. Allison, K.A. Amako, J. Apostolakis, H. Araujo, P. Arce, M. Asai, D. Axen, S. Banerjee, G. Barrand, et al., Geant4-a simulation toolkit. Nuclear instruments and methods in physics research section A. Acceler., Spectrom., Detect. Assoc. Equipm. **506**(3), 250–303 (2003)
4. R. Alejo, J.M. Sotoca, R.M. Valdovinos, P. Toribio, Edited nearest neighbor rule for improving neural networks classifications, in *International Symposium on Neural Networks* (Springer, 2010), pp. 303–310
5. P. Branco, L. Torgo, R.P. Ribeiro, A survey of predictive modeling on imbalanced domains. ACM Comput. Surv. (CSUR) **49**(2), 1–50 (2016)
6. R. Brun, L. Urban, F. Carminati, S. Giani, M. Maire, A. McPherson, F. Bruyant, G. Patrick, Geant: detector description and simulation tool. Technical report, CERN (1993)
7. B.A. Bucklin, N.L. Asdigian, J.L. Hawkins, U. Klein, Making it stick: use of active learning strategies in continuing medical education. BMC Med. Educ. **21**(1), 1–9 (2021)
8. F. Carminati, G. Khattak, M. Pierini, S. Vallecorsafa, A. Farbin, B. Hooberman, W. Wei, M. Zhang, B. Pacela, M.S. Vitorial, et al., Calorimetry with deep learning: particle classification, energy regression, and simulation for high-energy physics, in *Workshop on Deep Learning for Physical Sciences (DLPS 2017), NIPS* (2017)
9. N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority oversampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
10. S. Dasgupta, J. Langford, A tutorial on active learning, in *Proceedings of ICML* (2009)
11. L. Evans, The large hadron collider. New J. Phys. **9**(9), 335 (2007)
12. Z. Farou, N. Mouhoub, T. Horváth, Data generation using gene expression generator, in *International Conference on Intelligent Data Engineering and Automated Learning* (Springer, 2020), pp. 54–65
13. A. Fernández, S. García, M. Galar, R.C. Prati, B. Krawczyk, F. Herrera, *Learning From Imbalanced Data Sets*, vol. 11 (Springer, 2018)

14. S. Gopal, Y. Yang, Recursive regularization for large-scale classification with hierarchical and graphical dependencies, in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2013), pp. 257–265
15. S. Haghighi, M. Jasemi, S. Hessabi, A. Zolanvari, Pycm: multiclass confusion matrix library in python. J. Open Sour. Softw. **3**(25), 729 (2018)
16. M. Heide, A. Wilk, Particle identification with the transition radiation detector in alice. Verhandlungen der Deutschen Physikalischen Gesellschaft (2010)
17. D.H. Perkins, D.H., Perkins, *Introduction to High Energy Physics* (Cambridge University Press, Cambridge, 2000)
18. M.M. Rahman, D.N. Davis, Addressing the class imbalance problem in medical datasets. Int. J. Mach. Learn. Comput. **3**(2), 224 (2013)
19. P.J. Sadowski, D. Whiteson, P. Baldi, Searching for higgs boson decay modes with deep learning. Adv. Neural. Inf. Process. Syst. **27**, 2393–2401 (2014)
20. T. Sandhan, J.Y. Choi, Handling imbalanced datasets by partially guided hybrid sampling for pattern recognition, in *2014 22nd International Conference on Pattern Recognition* (IEEE, 2014), pp. 1449–1453
21. W.C. Sleeman IV., B. Krawczyk, Multi-class imbalanced big data classification on spark. Knowl.-Based Syst. **212**, 106598 (2021)
22. N.A. Verdikha, T.B. Adji, A.E. Permanasari, Study of undersampling method: instance hardness threshold with various estimators for hate speech classification. IJITEE (Int. J. Inf. Technol. Electr. Eng.) **2**(2), 39–44 (2018)
23. C.G. Viljoen, Machine learning for particle identification and deep generative models towards fast simulations for the Alice Transition Radiation Detector at CERN. Master's thesis, Faculty of Science (2019)
24. P. Vuttipittayamongkol, E. Elyan, Overlap-based undersampling method for classification of imbalanced medical datasets, in *IFIP International Conference on Artificial Intelligence Applications and Innovations* (Springer, 2020), pp. 358–369
25. X. Wang, B. Liu, S. Cao, L. Jing, J. Yu, Important sampling based active learning for imbalance classification. Sci. China Inf. Sci. **63**(8), 1–14 (2020)

Smart Agriculture Using Internet of Things: An Empirical Study

Mohit Kumar Saini and Rakesh Kumar Saini

Abstract IOT is the life-changing technology nowadays. That represents the attribute of computing & communication. Nowadays, IOT is using everywhere like home automation, smart health, smart cities, air smog monitoring, water distribution system, etc. The part of IOT is very huge and can be implemented everywhere. IOT also play a very important role in the area of agriculture. In the traditional agriculture system, farmers are using the oldest system of the farming. There is no technique of soil monitoring, rain water monitoring system etc. In this paper, I surveyed typical agriculture methods, which is used by the farmers in and what are the problems farmers are facing? I personally meet the farmer and visited their fields for collecting more information about the new technology which can be implementing in field nowadays. The goal of this study is to show how to use the Internet of Things to monitor humidity, soil condition, temperature, and give water to the field, as well as climate conditions. This report's IoT-based Smart Farming System is integrated with several sensors and a Wi-Fi module, resulting in a live data feed that can be accessed online.

Keywords Smart farming · IOT · Sensors · Wi-Fi · Agriculture

1 Introduction

The Internet of Things (IoT) has establish its application in various areas like smart industry, smart energy, smart city, home automation, smart car, connected building and campus smart agriculture, logistics, health care and every domain, also IoT.

The goal is to connect the physical and virtual worlds by using the Internet as the standard for exchanging information and communicating. The IoT is a network

M. K. Saini (✉)
School of Computing,, DIT University, Dehradun, Uttarakhand, India

Doon Business School-Dehradun, Dehradun, Uttarakhand, India

R. K. Saini
School of Computing, DIT University, Dehradun, Uttarakhand, India
e-mail: rk.saini@dituniversity.edu.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_13

of unified mechanical, computing, and digital machines, as well as animals, objects, and people, with unique identifiers and the ability to transfer data without the need for human-to-computer or human-to-human interaction (Fig. 1).

Smart agriculture is a hi-tech system and capital-intensive of rising sustainable and food cleanly for the sufficient. The goal of modern ICT (Information and Communication Technologies) in agriculture is to achieve this. Sensors are used to monitor the harvest field and automate the irrigation system in an IoT-based smart agriculture system. Farmers may see the state of their fields from any location. When compared to the traditional method, IoT-based smart farming is quite beneficial (Fig. 2).

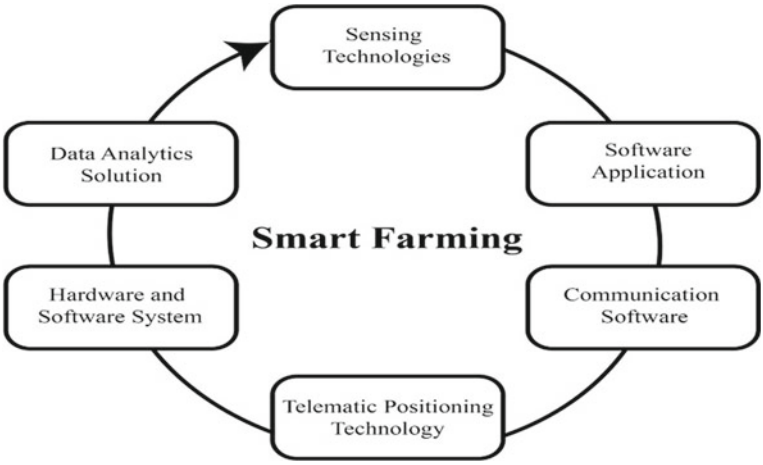


Fig. 1 Basic architecture of IOT based on H/S

Fig. 2 Computerize laser-land leveler using IOT device

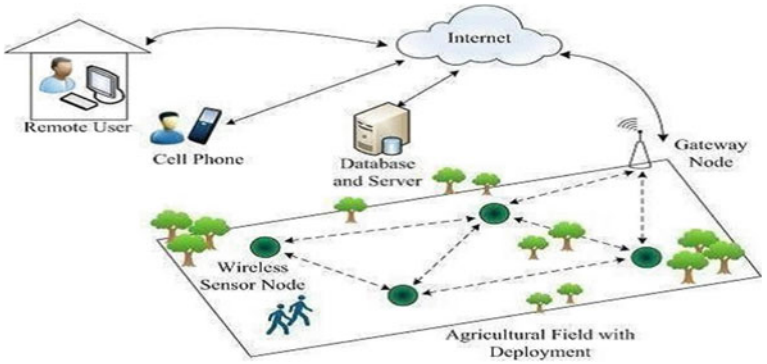


Fig. 3 Basic architecture of IoT in agriculture

The benefits of IoT based on agriculture are not only in terms of routine planning, large-scale farming activities but can also be realistic solutions to promote new or emerging agricultural practices such as family farming, organic farming, and improving more transparent farming. Using an IoT-based system may necessitate less energy, which can be provided by smaller solar panels. This might mean that an IoT-based system can profit whether electricity is supplied to a farm or a village. IoT makes use of a large number of sensors linked to the host, which provide data to a server, which saves it and sends it to an application that can process it (Fig. 3).

2 Application of IOT in Smart Agriculture

IoT applications in agriculture have enabled farmers to monitor water tank levels in real time, making irrigation more efficient. The advancement of IoT technology in agriculture operations has resulted in the usage of sensors in each phase of the farming process, such as how long it takes a seed to mature into a fully grown vegetable and how much it costs.

The following are the benefits of adopting new technology—Internet of Things in Agriculture.

1. **Climate Conditions**

   Climate change has a significant impact on agriculture, as it lowers the value and quality of crop production. However, IoT systems provide real-time weather information. Agricultural sensors can be found both within and outside the fields. They gather information from the natural world. There is never a yes. All of these components are detected by the sensors, which are designed to work with your smart agricultural equipment. These sensors keep an eye on the health of the plants as well as the weather. If unfavorable weather is identified, a warning is delivered.

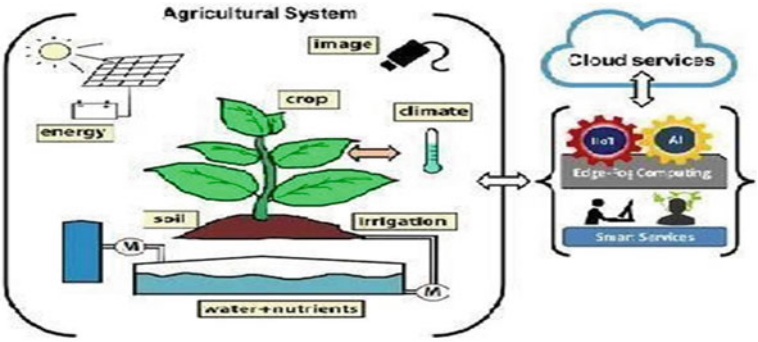


Fig. 4 Smart precision farming using IOT device

2. **Precision Farming**

P Precision agriculture is the most major application of the IoT in agriculture. It detects good farming applications such as field inspections, animal monitoring, vehicle and record keeping, and it prepares the practice of farming more correctly and prevents it. The goal of precise farming is to look into the specifics of the sensors and respond accordingly. Farmers utilize precision farming to create data with the help of sensors and analyses that data in order to make rapid and informed decisions (Fig. 4).

3. **Smart Greenhouse**

To make smart greenhouses, IoT has enabled weather forecasting to automatically correct climate conditions based on a precise set of data. Acceptance of the IoT in greenhouses has reduced human participation, making the overall process profitable while also increasing accuracy. Current and low-cost greenhouses are made with the use of solar-powered IoT sensors. These sensors capture data and transfer it to real-time data, which might be beneficial for monitoring greenhouse conditions in real-time (Fig. 5).

4. **Data Analysis**

The data store space available in the predicted data base system is insufficient to accommodate the data gathering provided by IoT sensors. In the smart farming system, the Cloud data base storage space Platform plays an important function. The sensors are an important part in gathering data on such a large scale. Using analytics technologies, the data is processed and analyzed into useful information. Data analytics tools aid in the extraction of information on weather, crop conditions, and livestock conditions. Using a data analytics technique for bid data (Fig. 6).

Fig. 5 Smart greenhouse using the IOT

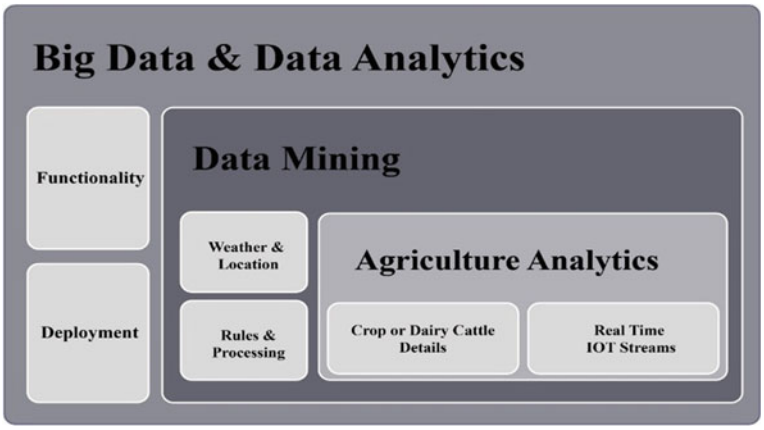


Fig. 6 Data Analytics process in agriculture using IOTs

5. Agricultural Drones

Every day technology is growing too improved over time and agricultural drones are the best example. Agriculture is the key industries to integrate drones. Drones are using in agriculture sector to improve a range of agricultural practice. The ways on the ground and aerial drones are used in agriculture are crop health evaluation, crop monitoring, irrigation, soil and planting, and field analysis. Once the crops are going on to growing, sensors point out their vegetation index health, and calculate them. Eventually, smart drone's rover cum the environmental impact factor (Fig. 7).

Fig. 7 Agriculture Drone

3 Gaps in Existing Work

After study of the existing work, I found some gaps that are as follows:

1. In the existing system, farmers have very less information; even sometime, they don't have any information regarding the weather forecast.
2. In the existing agriculture system, there is no sufficient sales and distribution information about the crop.
3. Farmers have the less knowledge about the ICT infrastructure and ICT illiteracy.
4. Farmers have the very less knowledge about the expenses and profit of ICT in the existing agriculture system.
5. Lack about the marketing research skills and research center.
6. Farmers do not know about the weather forecasting in advance and there are the radical changes in the climatic conditions every time.
7. Lack of awareness in agriculture profession among young and educated professionals.
8. High-cost machineries for work which are using by the farmers nowadays without any technology.
9. Farmers are using more manual work.
10. Maintenance cost is very high of tracking the record manually.

4 Motivation for Smart Agriculture Using IoT

Agriculture plays very vital and energetic role in the growth of farming motherland. In India, around 80% of people be determined by upon the agriculture and farming andone third of the country capital comes from farming industry. The difficulty about farming system has been already hampering to the development of the nation. The only key solution of this crisis is smart agriculture by change the existing established system of agriculture. IOT techniques in agriculture are pervasive and motivate me for the following reasons:

1. It helps to the farmers through prediction of rain and whether forecasting at early stage and suggest good remedies.
2. Provide the full wireless connectivity to the farmers for their farm.
3. To increase the production and ease the distribution of agricultural products.
4. It increases productivity, reduce manual work, reduce time, and makes farming more efficient.
5. Crop sales will be increased in global market. Farmer can easily connected to the global market without restriction of any geographical area.

5 Literature Review

Ashifuddin et al. [1]. The purposed research work presented a smart farming fields monitoring technology which monitors the temperature and soil humidity. Dealing out the sensed data takes necessary action depend upon the standards without human involvement. Here, moisture and temperature of the soil are analyses and these sensed principles are stored in cloud storage for future database depend at above-defined system setup, different types of level of temperature value and soil moisture are sensed and depend on existing values of soil moisture and temperature.

Dagar et al. [2] The proposed research effort is a relatively simple structure of IoT sensing devices that gather data and transfer it to a network server through a Wi-Fi network, where the data base server can perform actions based on the received data. With the use of IOT devices, these agricultural practises can be made more capable and excellent. IoT sensors can be used in a variety of agricultural applications.

Elijah et al. [3]. Authors presented IoT network and how the blend of IoT and data analytics is smart agriculture. We also offer the future opportunity and trends. We can categorize it into application scenarios, technological innovations, and agriculture business. One of the major areas that have drawn lot of research consideration is the utilization of LoWPan Wide Area communication technology for farming uses.

Kamyod et al. [4]. OPNET is used to evaluate the back-to-back consistency qualities of popular IoT communication architectures utilizing main reliability network metrics. When the number of sensor nodes grows, the reliability issues are studied and replicated. The simulation results show that the envisaged IoT communication network design is extremely reliable. Using end-to-end reliability things, we present

back-to-back reliability of two primary IoT communication network architectures at the time of network reliability parameters..

Sumathi et al. [5]. We have to recommend a platform of data analysis catering to desires of the stakes in agricultural division keeping in the mind the smart farming decision support system. Were presents a model of data analytical tool for intelligent and smart agriculture to ensuring the stakes drawn in agricultural industry?

Tran et al. [6]. We propose a fresh definition of the plant model that can be examined remotely and autonomously. The agricultural estate model is well discussed. The Central Processing Unit (CPU) is then responsible for the full system's aftercare and irrigation. To operate the ecological farm, a large number of sensors and actuators have been installed. Using this method, farming becomes less difficult and more appealing. Based on IoT, a worker can regulate the cultivated area in real-time monitoring. This is a dependable and quick system that aids farmers in field monitoring.

Mekala et al. [16]. In this paper, we look at some of the most common Cloud-based Agriculture IoT Sensor Monitoring Network applications. The first is water supply and energy, and an IoT farming addressed to the wireless network can be useful in several fields of agriculture. Due to a flawed irrigation model, inefficient field application, and the planting of water-stressed crops in the incorrect growing environment.

Rajeswari et al. [17]. IoT is employed in this article to sense and gather farming data, which is subsequently saved to a Cloud database server. Big data analysis based on cloud data is performed using a big data algorithm. Fertilizer requirements, crop analysis, stock supplies, and crop market. This model also makes it easier to estimate total production and crop yields based on region.

SauravVerma et al. [7]**.** In this research paper, the steps worried for the agriculture are discussed and primary focus of using IoT technology in agriculture industry,

i.e., in projected model which define the growth of agriculture.

Grimblatt et al. [8]**.** Regarding nutrients (NPK), soil moisture, soil properties, soil pH, temperature, light, and weather, among other things. To handle all relevant information as well as the complexity of the vegetation growth system, rely on IoT-based technology that can analyse, measure, and act as needed. The IoT makes it simple for technology to assess environmental variables such as soil and climate conditions. IoT has the potential to liberate a huge agriculture business by providing basic farming techniques and assisting in the cultivation of smart farming.

Puranik et al. [9]**.** In this model, we purpose to automate the Crop Monitoring, Control of pesticide, and Water distribution system. In this paper, we try to resolve difficulty using IoT-based system and automation system, in which, we can directly perform the most of the agricultural job and we can manipulate which one crops to cultivate as per the market before expenses most of the time in maintenance and crop production. It can also provide help to the farmers for more time to their personal life.

Ahmad et al. [10]**.** In this research, two different types of technologies, one being the IoT, and the other being a wireless sensor network, are merged in the current

technique for smart crop monitoring. Sensor-based nodes are used in agriculture fields to collect data on various factors.

Salam et al. [11]**.** In this study, an IoT technology is employed to research and develop the precision agriculture (PA) roadmap. The various useful trends and problems that exist have been highlighted. Precision agriculture difficulties can be addressed with IoT-based sensing and communications technologies. Precision agriculture practices and management solutions will benefit significantly from the deployment of cutting-edge technology, sensing, and communications..

Ananthi et al. [12]. The aim of this paper is give the embedded system for irrigation and soil monitoring to decrease the physical monitoring of the agriculture field and obtain the data using the mobile APPs. This model is planned to help the farming system to increase the production of the crops. Various sensors can be used for soil testing. The farmers use the cultivation the appropriate crop that suits the soil based on the results.

Hebel et al. [13]. Today different developing countries in the world are also using traditional methods and older mechanism for the agriculture segment. The little scientific development is established here that has bigger efficiency of production radically. To boost up the productivity, a narrative design approach is to design in this research paper. Smart agriculture with the use of IoT has been planned.

Jaiganesh et al. [14]**.** This work would be advance several study in the agriculture using of IoT in the farming system. Today farming is well established with thrust benefit, the sensors that can empower to the express. Information technology gives the benefit in to the cloud storage of agricultural. Agricultural cloud storage and IT advantage provide an expertise to agriculturists with growth of manures, yields, evaluate, illnesses strategy for healt o utilized Scientist chipping away at agriculture.

Jha et al. [15]. This study is more focuses on filed monitoring using IoT model that would be provide the humidity of the soil moisture and temperature of the field. This study is used to take a deterrent calculate for beating of crop and also boost up the productivity Filed.

6 Proposed Model

The primary goal of this model is to extend the central monitoring system and manage agricultural land. This may be controlled wirelessly from anywhere with a mobile phone. Users of the system can control the basic functions of collecting irrigation, environmental, soil moisture, and fertilization data without the need for human intervention. This information can be utilized to assess crop performance, create crop projections, and make individualized harvest recommendations for any farm that uses the app. Temperature, Soil Moisture, Humidity, and Light are among the variables that the sensor network will collect based on the meteorological conditions of the farmland. We will make decisions about the operations on the agricultural field with the help of this model. We can have multiple crops segregated into fields on a single field.

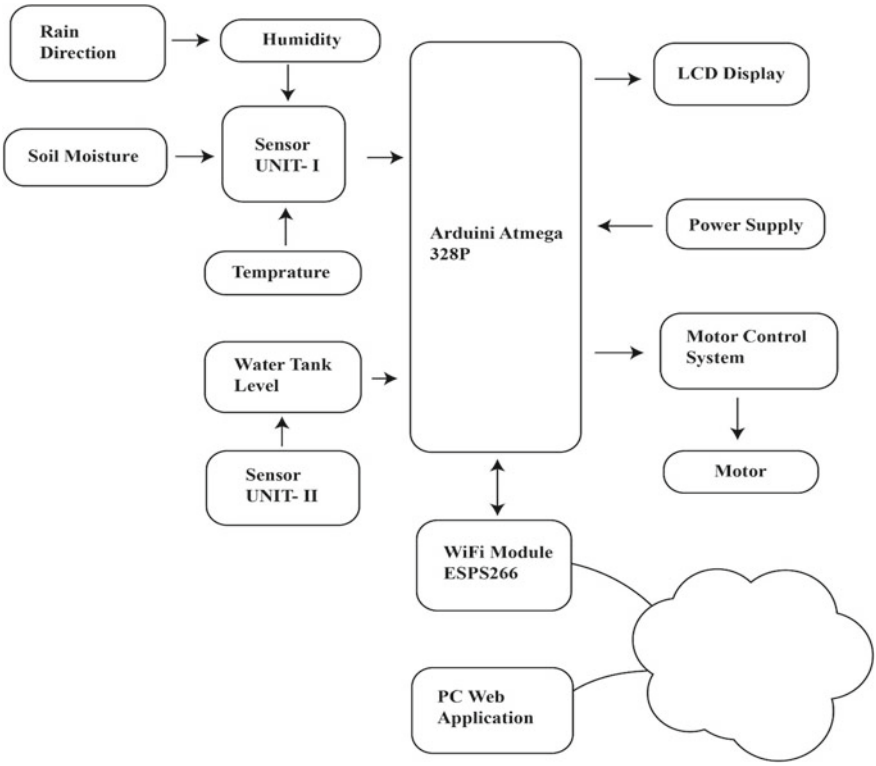


Fig. 8 Proposed Model

The data from all of these nodes is gathered and sent to a cloud storage facility. We're using the cloud service as a storage database for the data we've gathered. The information can be transferred to the cloud and stored in a database there. Farmers can access their accounts to check their history and current data for each node by logging in (Fig. 8).

7 Working of Model

1. **Sensor data attainment**: The sensor is used in tandem with the Adriano Uno board, and includes the DHT11 temperature sensor, rain detection sensor, humidity sensor, and soil moisture sensor.
2. **Wireless data communication**: The data will be collected from the numerous sensor nodes and remotely transferred to the computer server.

Table 1 Technique evolved

Techniques	Description
Big data	Data collecting and decision-making assistance
Evapotranspirati on	Any crop's water need can be calculated using the scenario. Crop coefficient and evapotranspiration
Wireless sensor network	It givesIt gives a cost-effective way to monitor and control the environment atmospheric pressure, temperature, humidity, and soil pH are all factors to consider
Cyber Physical System	System which integrates computational and physical components and interacts between them to sense the occurrence of change

3. **Data analysis and Decision-making**: Data analysis is the process of comparing the data acquired by various sensors from various agricultural fields to predetermined values. The motor will automatically turn on if the soil moisture falls below the threshold and vice versa. With the use of smartphone apps, the farmer may even start the engine.
4. **Irrigation automated system**: Once the control has just been enabled, the irrigation automation system will be controlled via a web or mobile application. Data processing is performed to pass access from the web application to the electrical switches via the Arduino Board.
5. **Web application**: The online application will be developed to allow anyone in the world to watch the field and crops via the internet. The web application can be interconnected using the Arduino processing IDE to control the Arduino.
6. **Mobile Application**: The Android operating system will be used to create the mobile apps. The phone application aids in the monitoring and control of agricultural fields from any location.

8 Discussion

Prepare the system as though it were a newly discovered individual. Data mining technology can be used to do this. Now that the IoT has arrived, it is easy to create an intelligent thinking system. The system must be effective, provide high-quality service, and be intelligent. The IoT will convene on a global scale. Any data can be obtained, analyzed, and used from any source. The Table 1 shows some of the ways we've implemented it to make the system smart.

9 Conclusion

IoT-enabled agriculture has aided in the implementation of cutting-edge scientific solutions. The gap between quantity, production, and quality has been bridged by this paper. Data is gathered via collecting and importing data from a variety of sensors for real-time use or cloud storage in a database, ensuring quick action, and little damage to the farmer's crops. Producers get their products processed faster and reach supermarkets in the shortest time feasible thanks to seamless end-to-end intellectual operations and better business process execution.

References

1. M. Ashifuddin, Z. Rehena, Iot based intelligent agriculture field monitoring system. in *2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, (IEEE, 2018), pp. 625–629
2. R. Dagar, S. Subhranil, S.K. Khatri, Smart Farming–IoT in Agriculture. in *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*, (IEEE, 2018) pp. 1052–1056
3. O. Elijah, T.A. Rahman, I. Orikumhi, C.Y. Leow, M.H.D. NourHindia, An overview of Internet of Things (IoT) and data analytics in agriculture: benefits and challenges. IEEE Int. Things J. **5**(5), 3758–3773 (2018)
4. C. Kamyod, End-to-end reliability analysis of an IoT based smart agriculture. in *2018 International Conference on Digital Arts, Media and Technology (ICDAMT)*, (IEEE, 2018), pp. 258–261
5. K. Sumathi, K. Santharam, N. Selvalakshmi, Data Analytics platform for intelligent agriculture. in *2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference on*, (IEEE, 2018), pp. 647–650
6. H.A.M. Tran, H.Q.T. Ngo, T.P. Nguyen, H. Nguyen, Design of green agriculture system using internet of things and image processing techniques. in *2018 4th International Conference on Green Technology and Sustainable Development (GTSD)*, (IEEE, 2018), pp. 28–32
7. S. Vashi, J. Ram, J. Modi, S. Verma, C. Prakash, Internet of Things (IoT): a vision, architectural elements, and security issues. in *2017international conference on I- SMAC (IoT in Social, Mobile, Analytics and Cloud) (I- SMAC)*, (IEEE, 2017), pp. 492–496
8. V. Grim blatt, F. Guillaume, F. Rivet, C. Jego, N. Vergara, Precision agriculture for small to medium size farmers—an IoT approach. in *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, (IEEE, 2019), pp. 1–5
9. V. Puranik, A. Ranjan, A. Kumari, Automation in Agriculture and IoT. in *2019 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU)*, (IEEE, 2019), pp. 1–6
10. N. Ahmad, A. Hussain, I. Ullah, B.H. Zaidi, IOT based Wireless Sensor Network for Precision Agriculture. in *2019 7th International Electrical Engineering Congress (iEECON)*, (IEEE, 2019), pp. 1–4
11. A. Salam, S. Shah, Internet of things in smart agriculture: ENABLING technologies. in *2019 IEEE 5th World Forum on Internet of Things (WF- IoT)*, (IEEE, 2019), pp. 692–695
12. N. Ananthi, J. Divya, M. Divya, V. Janani, IoT based smart soil monitoring system for agricultural production. in *2017 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, (IEEE, 2017), pp. 209–214

13. S. Heble, A. Kumar, K.V.D. Prasad, S. Samirana, P. Rajalakshmi, U.B. Desai, A low power IoT network for smart agriculture. in *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)*, (IEEE, 2018), pp. 609–614
14. S. Jaiganesh, K. Gunaseelan, V. Ellappan, IOT agriculture to improve food and farming technology. in *2017 Conference on Emerging Devices and Smart Systems (ICEDSS)*, (IEEE, 2017), pp. 260–266
15. R.K. Jha, S. Kumar, K. Joshi, R. Pandey, Field monitoring using IoT in agriculture. in *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, (IEEE, 2017), pp. 1417–1420
16. M.S. Mekala, P. Viswanathan, A survey: smart agriculture IoT with cloud computing, in *2017 international conference on microelectronic devices, circuits and systems (ICMDCS)*, (IEEE, 2017), pp. 1–7
17. S. Rajeswari, K. Suthendran, K. Rajakumar, A smart agricultural model by integrating IoT, mobile and cloud-based big data analytics. in *2017 International Conference on Intelligent Computing and Control (I2C2)*, (IEEE, 2017), pp. 1–5
18. https://r-stylelab.com/company/blog/iot/iot-agriculture-how-to-build-smart-greenhouse
19. M.K. Saini, R.K. Saini, Internet of Things (IoT) applications and security challenges: a review. Network **6**, 7
20. M. Abbasi, M.H. Yaghmaee, F. Rahnama, Internet of things in agriculture: a survey. in *2019 3rd International Conference on Internet of Things and Applications (IoT)*, (IEEE, 2019), pp. 1–12
21. S.V. Mukherji, R. Sinha, S. Basak, S.P. Kar, Smart agriculture using internet of things and MQTT protocol. in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, (IEEE, 2019), pp. 14–16
22. A. Bhattacharjee, P. Das, D. Basu, S. Roy, S. Ghosh, S. Saha, S. Pain, S. Dey, T.K. Rana, Smart farming using IOT. in *2017 8th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, (IEEE, 2017), pp. 278–280
23. W. Xue-fen, D. Xing-jing, B. Wen-qiang, L. Le-han, Z. Jian, Z. Chang, Z. Ling-xuan, Y.P. Yu-xiao, Y. Yi, Smartphone accessible agriculture IoT node based on NFC and BLE. in *2017 IEEE International Symposium on Consumer Electronics (ISCE)*, (IEEE, 2017), pp. 78–79
24. S.R. Prathibha, H. Anupama, M.P. Jyothi, IoT based monitoring system in smart agriculture. in *2017 International Conference on Recent Advances in Electronics and Communication Technology (ICRAECT)*, (IEEE, 2017), pp. 81–84
25. S. Jhansi Rani, S. MahaboobBasha, IOT agriculture system Based on LORAWAN. Int. J. Res. **6**(13), 141–161 (2019)

Intellegent Networking

A Study on the Implementation of Secure VANETs Using FPGA

Harsha Vardan Maddiboyina, V. A. Sankar Ponnapalli, and A. Naresh Kumar

Abstract Nowadays traveling becomes more passionate because of the technologies enhancing in the field of vehicular transportation. These technologies were integrated with the wireless network which develops a networking system known as the vehicular ad-hoc network (VANET) to implement a well-developed transportation system, i.e., intelligent transportation system. This paper demonstrates the role of VANETs using FPGA for a better and advanced transportation system. In this paper, different VANETs methodologies using FPGA were considered for the minimization of the live traffic congestions, vehicular calamities, etc., and enhancing the level of security for the communication network. This paper also focused on the internal architecture of the FPGA.

Keywords VANETs · FPGA · Intelligent transportation system · Communication

1 Introduction

Nowadays the usage of vehicles is abruptly elevating day by day for transportation. At present, the developed technologies facilitate transportation accessible [1]. These technologies make the transportation system very intelligent to form an intelligent transportation system (ITS) [2]. An ITS majorly set out a better solution for emergency utilities. This ITS can be implemented by using vehicular ad-hoc networks (VANETs) [3–6, 11]. The essential of VANETs is fabricated by using the ad-hoc network protocol known as Mobile ad-hoc network (MANETs) [12–14]. An inter-vehicle communication (IVC) gets initiated by using the VANETs mechanism.

H. V. Maddiboyina
R&D Department Sigma Embedded Systems Pvt. Ltd., Hyderabad, India

V. A. Sankar Ponnapalli (✉)
Department of Electronics and Communication Engineering, Sreyas Institute of Engineering and Technology, Hyderabad, India

A. Naresh Kumar
Department of Electrical and Electronics Engineering, Institute of Aeronautical Engineering, Hyderabad, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_14

VANETs implement two types of communication systems such as vehicle-to-vehicle communication system (V2V) and vehicle-to-infrastructure communication system (V2I) or vice versa, i.e., I2V [15–19]. VANET follows the IEEE 802.11p standard range of communication. In the V2V communication system, the authorized vehicles communicate with each other within the existing locality range. In V2I, infrastructure stands for the roadside unit (RSU). In the V2I (or) I2V communication system, the authorized vehicles communicate and share live traffic information with the RSUs and vice versa. Here V2V and V2I or I2V accomplish for dedicated short-range (DSR) communication [20, 21]. The architecture of VANETs is shown in Fig. 1.

VANETs accommodate majorly with three physical segments such as data authentication cloud (or) server segment (DAC), on-board unit (OBU) segment, and RSU segment [22, 23]. The OBU segment is a movable unit which is integrated within the vehicle. OBU system is integrated with the components such as embedded microcontrollers, GPS, IEEE standard communication protocols and sensors, etc. This OBU system sends the vehicular/traffic status to other vehicles and RSU system. OBU receives the data from other OBUs or RSUs and processes the obtained data using OBU to give a better output to the user. The RSU segment is a physical unit

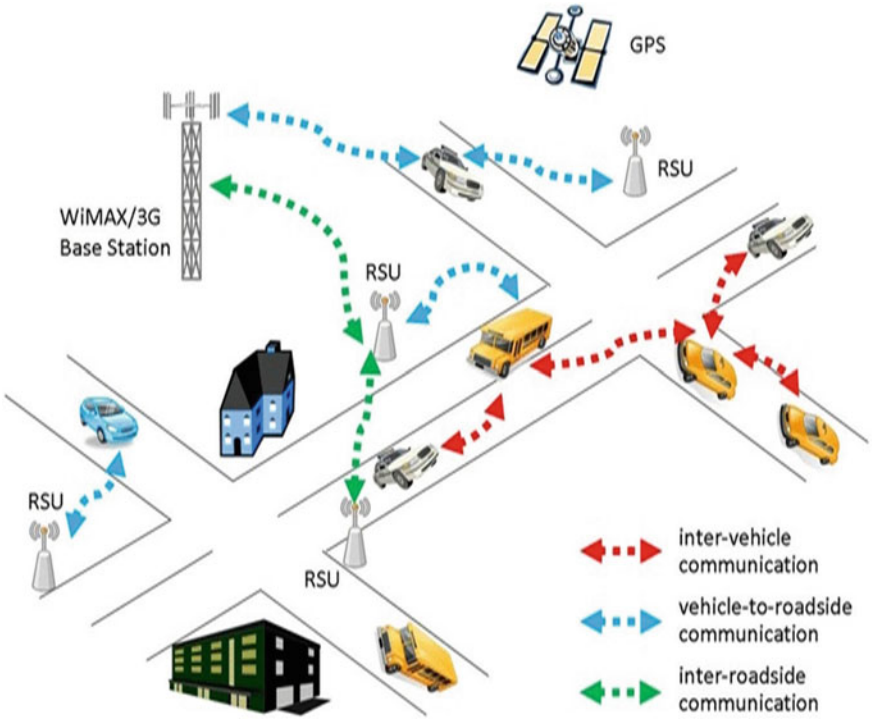


Fig. 1 Architecture of VANETs

placed along the roadsides and used to collect vehicular information such as vehicular location, velocity, etc., from the vehicles through reliable wireless communication. The data received by the RSU will be stored in the DAC segment [24, 25]. In the field of embedded systems and IoT environments [26, 27], the systems can be controlled by using different types of controllers or processors which were integrated with different integrated circuits (ICs) on their top. One such type of IC is field programmable gate array (FPGA). It has flexibility, versatility, usability, and other ideal properties by its nature. Internally FPGA is incorporated with an array of hardware blocks with customized programmable interconnects to do a particular task. In FPGA, interconnects are reprogrammable. The nature of FPGA is having a bit of similarity with PROM (programmable read only memory) and PLD (programmable logic device). The memories employed in PROM and PLD are static i.e., the data in their respected memories cannot be erasable. In FPGA, the data will be encapsulated in RAM (random access memory) and flash memory. It is reusable. The internal structure of FPGA is shown in Fig. 2.

The FPGA architecture is majorly embodied with three segments such as pads, interconnect, and logic blocks (LBs). Pads also called as I/O blocks, which are the external layer for the FPGA architecture [28–30]. This I/O blocks act as a bridge between the user and the internal FPGA architecture. LBs will process the inputs and accords the output. Interconnects set the direction between the LBs to execute the task as per the logic given by the user. FPGA is divided into three categories such as

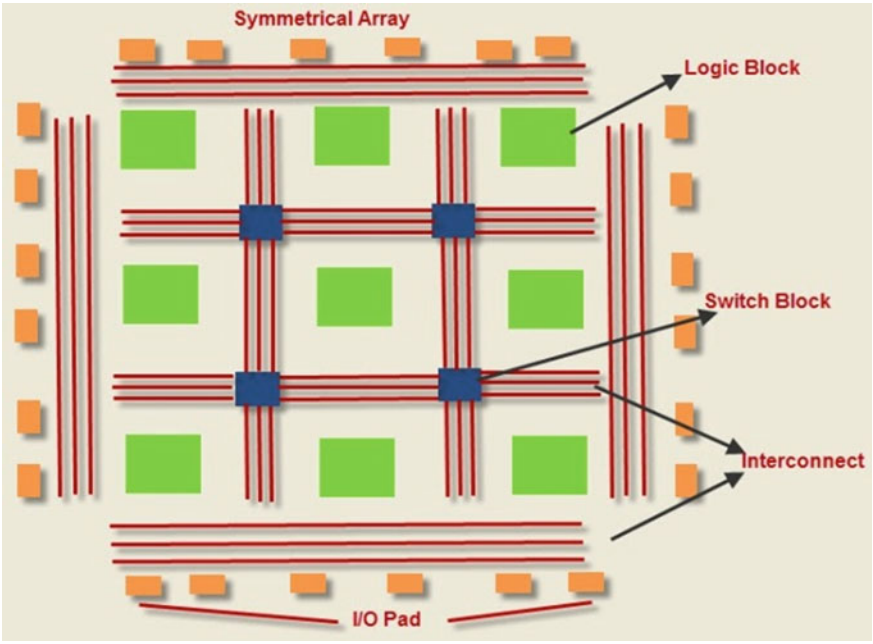


Fig. 2 Internal structure of FPGA

Low End FPGA, Middle Range FPGA, and High End FPGA. For the applications with less power consumption, less complexity, and less density these low end FPGAs are utilized. These high end FPGAs are utilized for high logic densities and gives high performance. The mid range FPGA acts as an intermediate between the low end and high end. In FPGA, the processing time is very less when compared with the other microcontroller. Microcontrollers such as LPC series, PIC, ATMega, etc., use sequential programming. FPGA uses parallel programming by its nature. Hardware Description Languages (HDLs) such as VHDL or Verilog are used to program the FPGA. Kintex, Artix, Virtex, Spartan, etc., are the different FPGA ICs used for the applications. These FPGAs are majorly used for digital signal processing, medical imaging, rf filtering, transportation, etc. Latency rate is going to be minimized in the VANETs communication with the help of FPGA because of its parallel processing nature. The rest of the paper is organized as follows: the second section of this paper deals with the various VANET-based applications with the integration of FPGA was discussed. Finally, the conclusion was drawn at the end of this review paper.

2 VANETs Integrated with the FPGA

An implementation of VANETs using FPGA-based hardware test bed approach for ITS was discussed in [31]. In this paper, a hardware model was developed for the implementation of VANETs on the desired FPGA using Verilog. A hybrid VANET-based communication was developed by integrating V2V and V2I communication systems and forms a V2V2I communication system. In V2V2I communication, master-salve communication was established. Here a transmitting vehicle will act as a master. Whereas, the receiving vehicles or RSUs will be a slave. To implement the proposed methodology, three different scenarios were considered such as V2V node, V2I node, and V2V2I node. In the V2V node, two different vehicles were considered such as 'V_a' and 'V_b'. When both the vehicles are in the same range of communication, then 'V_a' receives a positive edge of the clock to FPGA when 'V_b' is closer to 'V_a'. Now the 'V_a' gets enable and transmits an alert message to 'V_b' to avoid vehicular collision. In the V2I node, the two objects such as vehicle and infrastructure were considered as 'V_a' and 'I_a' respectively. If 'V_a' enters into 'I_a' communication range, then 'V_a' receives another clock pulse signal as discussed in the previous case. Then 'V_a' transmits the message to 'I_a' on the next rising edge of the clock pulse to publish it globally. The V2V2I node is the integration of the above two cases. Detection of surrounding vehicles using FPGA was discussed in [32]. In this paper, a wireless and portable panorama-based camera was integrated with the vehicle to detect the vehicles which are surrounded by it. The internal view of a panorama based camera system is shown in Fig. 3. The total hardware was integrated with the FPGA, to process the image and video files very quickly. To detect the surrounding vehicles, a unique sophisticated neural network methodology known as EZ-Net was implemented. A panorama camera view for vehicle detection is shown in Fig. 4. In the FPGA design, the captured image will be processed through SCCB and

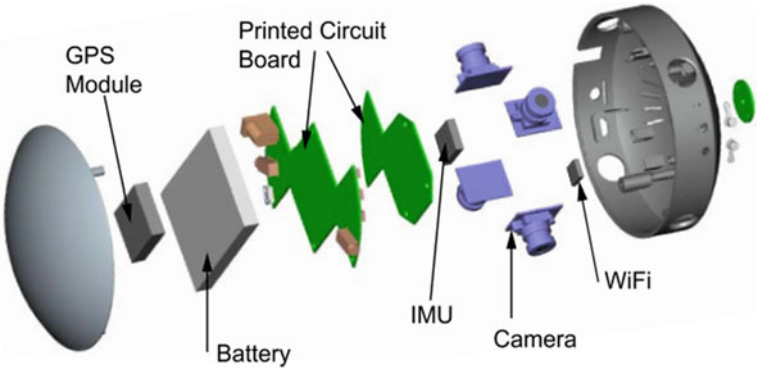


Fig. 3 Internal view of Panorama based camera system [32]

Fig. 4 A Panorama camera view for vehicle detection [32]

VC sections consecutively. SCCB stands for serial camera control bus and VC stands for video compression. Internally SCCB is processed through three different sections such as SS (simplified sample), buffer, and BAR (boundary artifacts reduction).

The architecture of FPGA based panorama camera system is shown in Fig. 5. The input image will get sampled and stored in the buffer. The BAR evaluates and reduces the dis-match artifacts from the sampled data and processed to the VC section for the compression of the video or image and generates the compressed panorama image.

FPGA-based real-time implementation of detection algorithm for automatic traffic surveillance sensor network was discussed in [33]. In this paper, an algorithm was

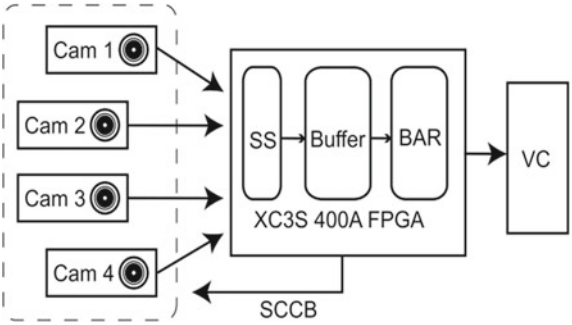


Fig. 5 Architecture of FPGA based Panorama camera system [32]

developed to detect the moving vehicles using a camera to evaluate the parameters such as no. of vehicles, their location, and speed using the sensor network cluster. The whole hardware setup was integrated with the FPGA. The block diagram of the developed algorithm using FPGA is shown in Fig. 6.

LI-FI-based smart traffic network was discussed in [35]. In this paper, LI-FI stands for light fidelity. It's a wireless-based communication mechanism used to establish communication using light. It transmits the data through the light at a very high speed of frequency 4–8 THz. The LI-FI methodology is flattering to the existing sovereign system. This LI-FI technology is integrated with the FPGA to generate and transmit an exceptional signal to traffic lights, autonomous vehicles, etc., through light pulses. When a start bit is received by the FPGA, it starts receiving the current sensed traffic data through the sensor networks until it receives the stop bit. The sensed data gets processed and generates the outputs for the respected inputs and transmits the output through the light pulses to the vehicular objects, traffic control systems, etc.

The application of RC5 for IoT devices in smart transportation system was discussed in [36]. In this paper, a smart encrypted algorithm named RC5 was developed to implement a secured ITS using IoT [37, 38]. This algorithm maintains the

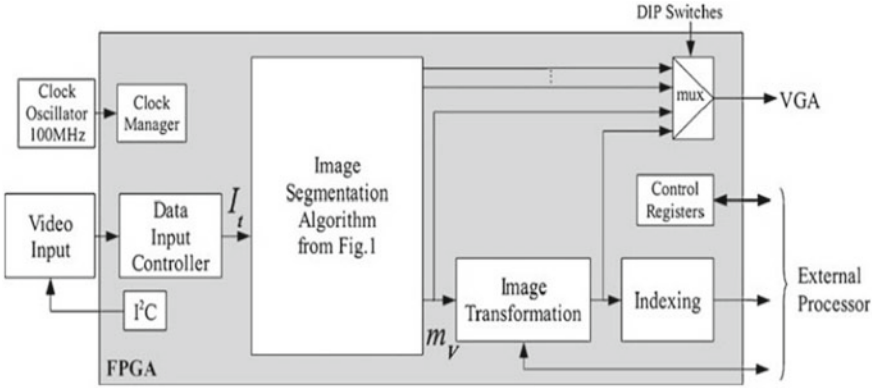


Fig. 6 Block diagram of the developed algorithm using FPGA [33]

privacy of the user data. The developed RC5 algorithm was integrated with the FPGA. The RC5 algorithm contains a secret key which was integrated with the parameters such as word size, no. of rounds, and no. of bytes as a fixed frame format. The protection of the algorithm gets strong by increasing the levels of rounds. The RC5 algorithm was divided into three segments such as key expansion (KE), encryption, and decryption. The KE segment enlarges the size of the key for the user's security. This key expands the table 'S'. The 'S' is an array that consists of random binary words. The KE segment follows 3 steps to create an array 'S'. In the initial step, the secret key was placed into the array. In the secondary step, the array 'S' was initialized into a pseudorandom bit pattern. In the final step, the generated bit pattern was integrated with the secret key. The encryption segment is used to generate 'S' table from the key expansion segment. At last, the decryption segment is obtained by using the encryption segment. In view of smart city developments, VANETs are playing a vital role in these aspects [39].

3 Conclusion

Implementation of the VANETs which are integrated with the FPGA is emphasized in this review paper. The organizing of the traffic becomes immensely smart with the enactment of the VANETs in the present days. Different methodologies for the implementation of intelligent transportation systems were discussed. This paper also focused on the importance and the uses of the FPGA in the field of transportation. In the introduction section, the working of VANETs and FPGA was explained. In the VANETs integrated with the FPGA section, six different cases were considered and demonstrated the role of FPGA in VANETs to build a well-developed ITS.

References

1. F. Abbas, P. Fan, Z. Khan, A novel low-latency V2V resource allocation scheme based on cellular V2X communication. IEEE Trans. Intell. Transp. Syst. **20**(6), 2185–2197 (2018)
2. J. Bhatia, P. Kakadia, M. Bhavsar, S. Tanwar, SDN-Enabled network coding based secure data dissemination in VANET environment. IEEE Internet Things J. **7**(7), 6078–6087 (2019)
3. H.V. Maddiboyina, V.A.S. Ponnapalli, Fuzzy logic based VANETS: a review on smart transportation system, in *2019 International Conference on Computer Communication and Informatics (ICCCI–2019), Coimbatore, Tamil Nadu, India* (2019)
4. G. Li, Q. Sun, L. Boukhatem, J. Wu, J. Yang, Intelligent vehicle-to-vehicle charging navigation for mobile electric vehicles via VANET-based communication. IEEE Access **7**, 170888–170906 (2019)
5. J.C. Mukherjee, A. Gupta, R.C. Sreenivas, Event notification in VANET with capacitated roadside units. IEEE Trans. Intell. Transp. Syst. **17**(7), 1867–1879 (2016)
6. J. Cheng, G. Yuan, M.C. Zhou, S. Gao, C. Liu, H. Duan, A fluid mechanics-based data flow model to estimate VANET capacity. IEEE Trans. Intell. Transp. Syst. **21**(6), 2603–2614 (2019)
7. W. Sun, Q. Wu, Y. Xiang, Design and implementation of the FPGA-based wireless network card. Adv. Control Eng. Inf. Sci. **15**, 2234–2238 (2011)

8. A. Appathurai, G. Manogaran, N. Chilamkurt, Trusted FPGA based transport traffic inject, impersonate (I2) attacks beaconing in the Internet of Vehicles. IET Networks **8**(3), 169–178 (2018)
9. R. Lacuesta, J. Gallardo, J. Lloret, G. Palacios, Integration of data from vehicular Ad Hoc networks using model-driven collaborative tools. Mobile Inf. Syst. 1–15 (2016)
10. N. Kshirsagar, U.S. Sutar, An intelligent traffic management and accident prevention system based on VANET. Int. J. Sci. Res. (IJSR) **4**(7), 228–230 (2015)
11. B. Marzak, K.E. Guemmat, E.H. Benlahmar, M. Talea, Clustering in vehicular Ad-Hoc network using artificial neural network. Int. Rev. Comput. Softw. **11**(6), 548–556 (2016)
12. D. Tian, K. Zheng, J. Zhou, X. Duan, Y. Wang, Z. Sheng, Q. Ni, A microbial inspired routing protocol for VANETs. IEEE Internet Things J. **4**(4), 2293–2303 (2017)
13. H.V. Maddiboyina, V.A.S. Ponnapalli, Mamdani fuzzy-based vehicular grouping at the intersection of roads for smart transportation system. Int. J. Intell. Eng. Inf. **7**(2/3), 263–285 (2019)
14. N. Lin, L. Fu, L. Zhao, G. Min, A. Al-Dubai, H. Gacanin, A novel multimodal collaborative drone-assisted VANET networking model. IEEE Trans. Wireless Commun. **19**(7), 4919–4933 (2020)
15. Z. Lu, G. Qu, Z. Liu, A survey on recent advances in vehicular network security, trust, and privacy. IEEE Trans. Intell. Transp. Syst. **20**(2), 760–776 (2019)
16. H. Li, L. Pie, D. Liao, S. Chen, M. Zhang, D. Xu, FADB: A fine-grained access control scheme for VANET data based on blockchain. IEEE Access **8**, 85190–85203 (2020)
17. M. Poongodi, M. Hamdi, A. Sharma, M. Ma, P.K. Singh, DDoS detection mechanism using trust-based evaluation system in VANET. IEEE Access **7**, 183532–183544 (2019)
18. C.M. Huang, C.F. Lai, The delay-constrained and network-situation-aware V2V2I VANET data offloading based on the multi-access edge computing (MEC) architecture. IEEE Open J. Veh. Technol. **1**, 331–347 (2020)
19. H. Tan, I. Chung, Secure authentication and key management with blockchain in VANETs. IEEE Access **8**, 2482–2498 (2019)
20. P.V. Kumar, M. Azees, A. Kannan, L.J. Deborah, Dual authentication and key management techniques for secure data transmission in vehicular Ad Hoc networks. IEEE Trans. Intell. Transp. Syst. **17**(4), 1015–1028 (2015)
21. N.W. Lo, J.L. Tsai, An efficient conditional privacy-preserving authentication scheme for vehicular sensor networks without pairings. IEEE Trans. Intell. Transp. Syst. **17**(5), 1319–1328 (2015)
22. T. Gao, X. Deng, N. Guo, X. Wang, An Anonymous authentication scheme based on PMIPv6 for VANETs. IEEE Access **6**, 14686–14698 (2018)
23. G. Sun, Y. Zhang, H. Yu, X. Du, M. Guizani, Intersection fog-based distributed routing for V2V communication in urban vehicular Ad Hoc networks. IEEE Trans. Intell. Transp. Syst. **21**(6), 2409–2426 (2019)
24. L. Wu, Q. Sun, X. Wang, J. Wang, S. Yu, Y. Zou, B. Liu, Z. Zhu, An efficient privacy-preserving mutual authentication scheme for secure V2V communication in vehicular Ad Hoc network. IEEE Access **7**, 55050–55063 (2019)
25. L. Yao, J. Wang, X. Wang, A. Chen, Y. Wang, V2X Routing in a VANET based on the hidden Markov model. IEEE Trans. Intell. Transp. Syst. **19**(3), 889–899 (2017)
26. M. Elnawawy, A. Farhan, A.A. Nabulsi, A.R.A. Ali, A. Sagahyroon, Role of FPGA in internet of things applications, in *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Ajman, United Arab Emirates* (2019)
27. T.H. Szymanski, Securing the industrail-tactile internet of things with deterministic silicon photonics switches. IEEE Access **4**, 8236–8249 (2016)
28. X. Han, L. Xing, Mobile robot based on FPGA, in *2018 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), Xiamen, China* (2018)
29. P.S. Huynh, D. Vincent, L. Patnaik, S.S. Williamson, FPGA-based PWM implementation of matrix converter in inductive wireless power transfer systems, in *2018 IEEE PELS Workshop on Emerging Technologies: Wireless Power Transfer (Wow), Montreal, QC, Canada* (2018)

30. L. Kang, C. Wang, The design and implementation of multi-precision floating point arithmetic unit based on FPGA, in *2018 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), Xiamen, China* (2018)
31. S.W. Alam, B.M. Khan, Implementation of VANETs using FPGA-based hardware Test-Bed approach for intelligent transportation system. Bahria Univ. J. Inf. Commun. Technol. **9**, 30–34 (2016)
32. L. Chen, Q. Zou, Z. Pan, D. Lia, L. Zhu, Z. Hou, J. Wang, D. Cao, Surrounding vehicle detection using an FPGA Panoramic camera and deep CNNs. IEEE Trans. Intell. Transp. Syst. **21**(12), 5110–5122 (2019)
33. M. Wojcikowski, R. Zaglewski, B. Pankiewicz, FPGA-based Real_Time implementation of detection algorithm for automatic traffic surveillance sensor network. J. Signal Process. Syst. **68**(1), 1–18 (2012)
34. W.X. Guo, J. H. Luo, X.M. Hu, Application of microcontroller FPGA technology in designing intelligent transportation system internet of things. Appl. Mech. Mater. **540**, 381–384 (2014)
35. A. Correa, A. Hamid, E. Sparks, LI-FI based smart traffic network, in *2018 IEEE Transportation Electrification Conference and Expo (ITEC), Long Beach, CA, USA* (2018)
36. N. Alsaffar, W. Elmedany, H. Ali, Application of RC5 for IoT devices in smart transportation system, in *2019 8th International Conference on Modeling Simulation and Applied Optimization (ICMSAO), Manama, Bahrain* (2019)
37. S. Shinde, S. Kumar, P. Johri, A review: eye tracking interface with embedded system & IOT, in *IEEE International Conference on Computing, Power and Communication Technology (GUCON), Greater Noida, India* (2018)
38. S. Kumar, P. Raja, G. Bhargavi, A comparative study on modern smart irrigation system and monitoring the field by using IOT, in *IEEE International Conference on Computing, Power and Communication Technology, (GUCON). Greater Noida, India* (2018)
39. D.S.S. Satyanarayana, K.M.V.V. Prasad, Multilayered antenna design for smart city applications, in *2nd Smart Cities Symposium (SCS 2019), Bahrain, Bahrain* (2019)

Adoption of Microstrip Antenna to Multiple Input Multiple Output Microstrip Antenna for Wireless Applications: A Review

Nitasha Bisht and Praveen Kumar Malik

Abstract Over the past decade, the production of various applications involving wireless communication has been increasing rapidly. All such wireless system's efficiency depends upon the antenna's design and its proper functioning. Today for most of the wireless applications, MIMO antenna designed by using a combination of microstrip antenna is being desired. Several distinguished researchers have made various efforts to develop the antennas as per the requirements of industrial needs. Further, to grasp the idea for designing of multiple input multiple output (MIMO) microstrip antenna, one must have an understanding of conventional microstrip antenna. Taking this point into consideration, this paper provides an extensive review on the journey of a microstrip antenna from conventional microstrip antenna to MIMO microstrip antenna for wireless applications. This extensive review is fractioned into two sections; the first section includes the simulated and fabricated studies conducted on return loss, bandwidth, and gain enhancement strategies that rely on the use of insertion of slots and slits, vias, fractal geometry with the use of different substrates in microstrip antenna and the second section presents the studies based on various diversity performance parameters, bandwidth, and isolation enhancement strategies for MIMO microstrip antenna that depends on the use of defective ground structure (DGS), decoupling structure (DS), geometrical modification of patch and insertion of slots and slits. Further, for better understanding of scientific developments various findings are then delineated, which shows that future work in this area will offer further growth.

Keywords Wireless · Microstrip patch antenna · MIMO antenna

N. Bisht (✉) · P. K. Malik
School of Electronics and Communication, Lovely Professional University, Phagwara, Punjab, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_15

1 Introduction

The wireless technology basically uses radio waves to transmit and receive information without cables and wires [1]. With the help of wireless technology, billions of people are exchanging information with each other by using many electronic gadgets like laptops, tablets, mobile phones, pager, etc. This rapid progress is made in wireless system to make voice and data service available at any place and time. Although wireless systems have been used from 1970s, but with the increasing demand of wireless networking technology, nowadays researchers are interested to develop the fastest wireless communication system which will provide high data rate and also will increase the capacity of the system [2]. The main technology for achieving elevated data rate for present and future wireless communication service requirement is the MIMO scheme. It is a wireless technology that uses multiple transmitter and receiver to simultaneously transfer more information as shown in Fig. 1a [3]. There have been various benefits of MIMO system like it can significantly improve the data rate, coverage area, reliability, and efficiency of the wireless system. But few concerns are also related with it such as capacities in real environment are much lower and physical dimension of equipment and the price of equipment grow with number of antenna's used since extra processing power is required. Designing a MIMO antenna for modern wireless gadgets, viz. laptops, phones, and tablets requires a small size antenna whose compact sizes make very little space available. That is why light weight, low profile, economical, ease of mass production, and easy installation microstrip antennas have become appealing to use in MIMO antenna design. Microstrip patch antenna structure has three layers, viz. ground plane, dielectric substrate, and radiating patch as presented in Fig. 1b [4]. A copper patch is generally placed as the top layer, dielectric substrate layer is sandwich at the center, and conducting ground plane is at the bottom [5, 6].

Further, the combination of several microstrip antennas to build a MIMO antenna is the best solution for fulfilling the fastest wireless communication requirement. This paper critically analyzed different conventional microstrip antenna and MIMO antenna to meet the essential demands of wireless applications with usage of various techniques that will be of benefit for many upcoming studies. The present work

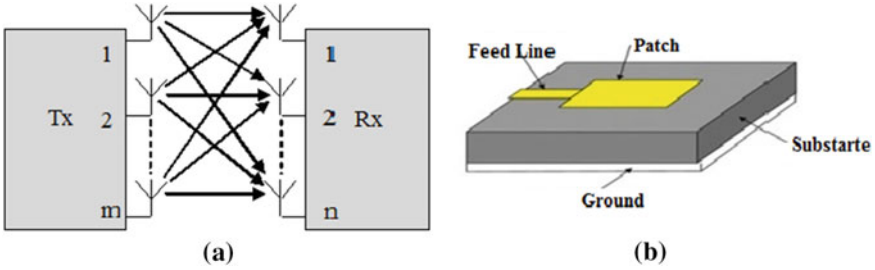


Fig. 1 **a** Structure of MIMO [3]. **b** Microstrip patch antenna [4]

covers three objectives as 1. The purpose here is limited to provide a description of various techniques described in numerous simulation and fabricated studies, for the designing of conventional microstrip patch antenna. 2. The present study encapsulates various analyses of MIMO antenna using different techniques like defective ground structure, decoupling structure, insertion of slits and slots, etc. The influence of these techniques has been comprehensively presented. 3. The analysis of technologies based on different antenna parameters such as gain, bandwidth, multiband, return loss, isolation, diversity performance, etc. gives a deeper insight about both the said antennas.

Keeping this in view, the thorough analysis will improve reader's comprehension of exploring the various techniques which are used for the designing of microstrip antenna and MIMO antenna, respectively, to achieve the desired parameter of wireless applications. These studies are presented in Sect. 2 and Sect. 3. Out of these two sections, Sect. 2 will cover the studies based on microstrip patch antenna. This is followed by Sect. 3, where MIMO antenna investigations are discussed. For better understanding of scientific developments, a comparison delineated in discussion section.

2 Microstrip Patch Antenna

This section covers the studies based on microstrip patch antenna. Various researchers have performed a number of investigations for the enhancement of gain, bandwidth, return loss, and multiband behavior of antenna. Comparison of reported microstrip antenna studies is also shown in Table 1. Verma and Srivastva [7] presented a design of triple band rectangular patch microstrip antenna for various applications of wireless technologies. The proposed antenna patch has been built by successively loading rectangular patches of different sizes (length and width). The configuration of the antenna is presented as Fig. 2a. Panda and Mishra [8] designed a bi-circular patch antenna. Firstly, two circular arcs are placed at a certain distance between each other. Then to enhance the bandwidth, an optimized circular slot was implemented at the center of arc. This antenna is presented in Fig. 2b.

Sharaf et al. [9] designed an electromagnetically coupled two rectangular patch antennas. The dimensions of first and second patch are set in such a way so that it operates at 38 and 60 GHz. The geometry of the antenna can be seen as in Fig. 2c. Devesh et al. [10] proposed a square microstrip antenna loaded with hexagonal gasket fractal structure. Four iterations of fractal geometry have been made for design of the said antenna. By using the fractal technique 68.4% area of patch is reduced. This antenna is presented in Fig. 2d. Liu et al. [11] presented the idea to design a broadband circular patch shape antenna with artificial structure to enhance the gain. Here, a comparison between an antenna with one ring and circular shape patch and an antenna with one ring, circular shape patch, and loaded by three layers of I-shaped structure is done. The configuration is shown in Fig. 3a.

Table 1 Comparison of reported microstrip patch antenna (MPA) studies

Reference	Antenna size (mm^3)	Analyzed parameter	Technique used
Verma and Srivastava [7]	39.04 × 47.64 × 1.6	BW = 28/188/813 MHz, Gain = 3.287/3.067/3.480 dB, Return Loss = −18.8/−12.78/−34.52 dB	Notches and slot: multiple bands, BW increases, gain increases
Panda and Mishra [8]	60 × 60 × 1.6	Return Loss = −20.35/−21.08/−19.52/−19.11 dB, Gain = 3 dB	Slot: multiple bands
Sharaf et al. [9]	15 × 25 × 0.25	BW = 2000/3200 MHz, Gain = 4.36/3.35 dB, Return Loss = −42/−47 dB	Slit and electro-magnetic coupling: BW increases, gain increases
Devesh et al. [10]	50 × 50 × 1.6	BW = 75/151/192/358/137/206 MHz Gain = 6/8.37/9.65/9/7.84/9.34 dB, Return loss = −19.59/−25.10/−24.04/−29.83/−20.82/−24.90 dB	Fractal Geometry: multiple bands, gain Increases
Liu et al. [11]	–	BW = 1070/1270 MHz, Gain = 1.8 dB Return loss = −35/−28 dB	ISS structure, slots, and shorting pins: BW increases, gain increases
Siddiqui et al. [12]	80 × 90 × 1.6	Return loss = −19/−27.37/−20.16/−20.74/−35.52 dB	Koch fractal structure provide: multiple bands
Aoad et al. [13]	28 × 28 × 1.5	Gain = 6.73/5.22/5.08/4.91 dB, BW = 98.022/134.9/157.51/138.27 MHz, Return loss = −13.8/−23.2/−24.4/−16.1 dB	Spiral antenna along with slot multiple bands, increase gain
Devesh et al. [14]	50 × 50 × 1.6	Return loss = −20/−25/−22/−28 dB, Gain = 6.81/7.91/7.32/8.89 dB	Gasket fractal geometry: multiple bands, increase gain
El-Khamy et al [15]	90 × 110 × 1.6	BW = 5300/2000/1500/3/3700 MHz, Gain = 9 dB (max value)	Fractal geometry: multiple bands, increase gain, increase BW

(continued)

Table 1 (continued)

Reference	Antenna size (mm^3)	Analyzed parameter	Technique used
Raviteja [16]	25.4 × 19.4 × 1.6	Return loss = −17.248, BW = 120 MHz, Gain = 9.67 dB	Trapezoidal slot: increase gain
Gupta et al. [17]	32 × 36 × 1.25	Return loss = −30/−21/−22/−20/−16 dB, Gain = 1/1.62/2.96/3.02/1.96 dB	Wheel shape fractal structure: multiple bands, increase gain
Singh et al. [18]	55 × 46 × 1.58	Gain = 9.19/3.04/5.19/5.39 dB, BW = 120/200/330/690 MHz, Return Loss = −12/−34/−20/−17 dB	Slot on patch: multiple bands, increase gain

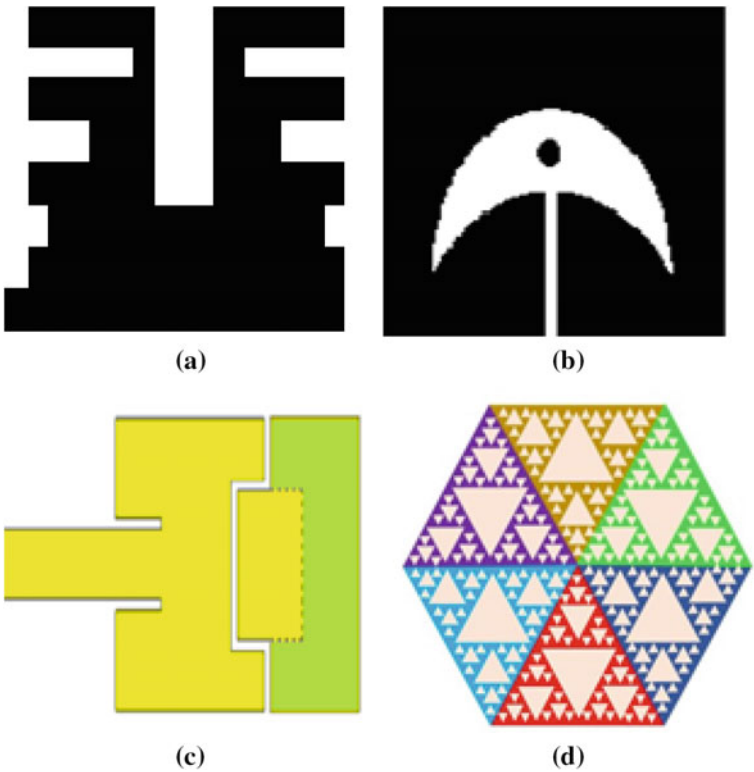


Fig. 2 **a** Rectangular patch with slots [7]. **b** Bi-circular patch antenna [8]. **c** Dual band microstrip patch antenna [9]. **d** Hexagonal gasket fractal antenna [10]

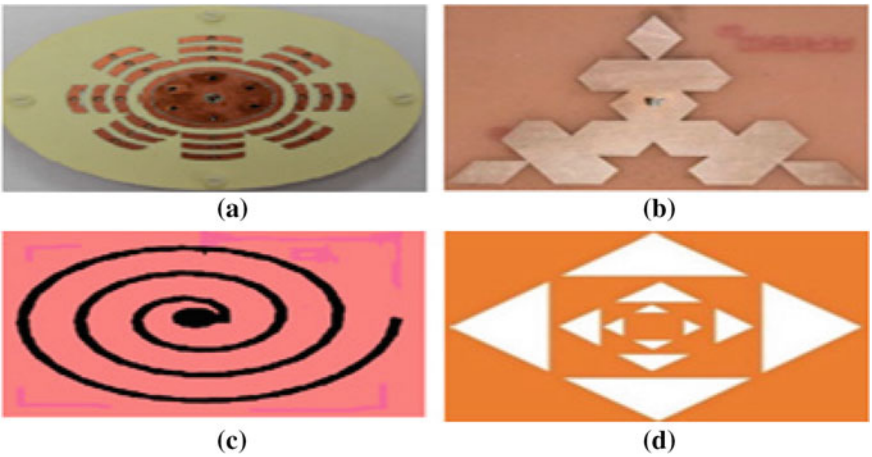


Fig. 3 **a** Antenna with I-shaped structure, slots, rings, and shorting pin [11]. **b** A-shaped triangle patch antenna [12]. **c** Circular microstrip patch antenna with spiral [13]. **d** Square Sierpinski triangular fractal microstrip antenna [14]

Siddiqui et al. [12] designed an A-shaped triangle microstrip antenna based on Koch's snowflake concept. Each side of the triangle is divided into four lines. Fractal geometry reduces the overall size of the proposed antenna. The geometry of the antenna can be seen as in Fig. 3b. Aoad et al. [13] presented a circular microstrip antenna in combination with spiral. The configuration is shown in Fig. 3c. During designing of the proposed antenna careful selection of parameters like dielectric substrate, width of spiral, number of turns in spiral antenna, and radius of circular patch has been done. Devesh et al. [14] designed a square Sierpinski triangular fractal microstrip antenna. Basically, triangular slots are loaded in square patch during each iteration (up to third iteration). The geometry of the proposed antenna is shown in Fig. 3d. El-Khamy et al. [15] designed a multiband fractal antenna. Circular patches of different scaling factors are connected to each other with sub-feedline to get better impedance matching. The configuration of the antenna is presented as Fig. 4a. Raviteja [16] presented a rectangular MPA with two trapezoidal slots for wireless applications. A comparison between the results of proposed antenna and conventional rectangular antenna is done, and there is a lot of improvement in the parameters of the proposed antenna as compared to conventional antenna. The geometry of the antenna is shown in Fig. 4b.

Gupta et al. [17] designed a wheel-shaped fractal microstrip patch antenna using iteration technique. A combination of circular patch and square patch is used for designing of wheel-shaped antenna. The configuration of the antenna is presented as in Fig. 4c. Singh et al. [18] presents the idea of a microstrip circular patch antenna with defective ground plane and rectangular slot. Different dimensions of rectangular slots are cut from circular patch using iteration technique to increase the gain of antenna [19]. The configuration of the antenna can be seen in Fig. 4d.

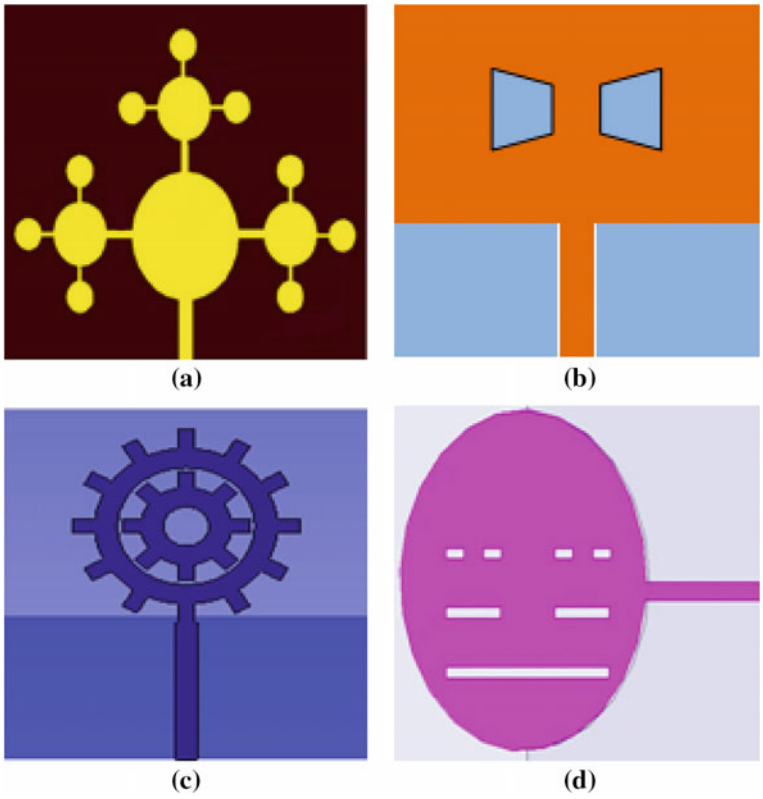


Fig. 4 **a** Circular antenna with fractal geometry [15]. **b** Rectangular MPA with two trapezoidal slots [16]. **c** Wheel-shaped microstrip patch antenna [17]. **d** Circular MPA with rectangular slot geometry [18]

3 Multiple Input Multiple Output Antenna

Numerous investigations for the enhancement of isolation, gain, and various diversity performance parameters like envelop correlation coefficient (ECC), diversity gain (DG), etc. are critically analyzed in this portion. Comparison of reported MIMO microstrip antenna studies are also shown in Table 2. The values of isolation parameter are obtained from s-parameter graph of each research paper. Gurjar et al. [20] presented a two-element modified rectangular fractal MIMO antenna with square-shaped funnel like stub in ground plane to achieve wideband and miniaturization performances. The fractal rectangular patch MIMO antenna geometry is presented in Fig. 5a. Ullah et al. [21] designed a semi-circular-shaped radiator MIMO antenna with tapered microstrip feeding. T-shaped parasitic element is placed in the ground plane to improve isolation. The configuration of the antenna is presented in Fig. 5b. Kumar et al. [22] presented a study on UWB-MIMO antenna having dual band

Table 2 Comparison of reported MIMO microstrip patch antenna studies

Reference	Antenna size (mm^3)	Analyzed parameter	Technique
Gurjar et al. [20]	24 × 30 × 0.8	BW = 9.6 GHz, Isolation = < −16.3 dB, Diversity Gain = > 9.95 dB, ECC = < 0.05	DGS and fractal geometry: improved isolation and BW
Ullah et al. [21]	40 × 47 × 1.5	Isolation < −20 dB, Diversity Gain = 10 dB, ECC = < 0.02	Tapered microstrip feedline and DS: improved BW and isolation
Kumar et al. [22]	19 × 30 × 0.8	BW = 7.5 GHz, Isolation < −18 dB, Diversity Gain > 9.7 dB, ECC = < 0.13	DGS, patch slot, and feedline slits: improved isolation and BW
Niu et al. [23]	48.5 × 60.6 × 1	Isolation = −34.2 and −36.3 dB, ECC = < 0.01, Diversity Gain = 8.51 dB	Decoupling structure and DGS: improved isolation
Ghannad et al. [24]	33 × 22 × 1	Isolation = −46.5 dB, ECC = < 0.15, Diversity Gain = 9.7 dB	Feedline stubs, strip and DGS: improved BW
Dkiouak et al. [25]	27 × 21 × 1.6	Isolation = < −23 and −22.5 dB, ECC = 0.13 and 0.002, Diversity Gain = 9.7 and 9.97 dB	DGS and patch slot: improved BW
Abdullah et al. [27]	30 × 55 × 1.524	Isolation = −20 dB, Diversity Gain = 5.80 dB, ECC = < 0.01	Shorting pins: improved isolation
Khattak et al. [28]	31 × 7 × 0.508	BW = 1.2 GHz and 1 GHz,SAR = 1.19/1.16/1.37 W/kg	Patch elliptical slot: improved BW
Marzouk et al. [29]	55 × 110 × 0.508	Isolation = −28.32 dB/−26.27 dB, BW = 1.0683/1.4306 GHz, Diversity Gain = 7.95/8.27 dB, ECC = < 0.0005	DGS: improved isolation
Firmansyah et al. [30]	180 × 180 × 1.6	Bandwidth = 2.322 MHz, Isolation = −26.18 dB	Beveled half-cut patch: improved isolation
Babu and Anuradha [31]	60 × 40	Isolation = −54/−50/−40/−44 dB, ECC = 0.03/0.00473/0.0075/0.0495, Diversity Gain = 9.87/9.983/9.97/9.803 dB	Patch slots: improved isolation
Jilani et al. [32]	12 × 12 × 0.8	Bandwidth = 12.4 GHz, Isolation = < −22 dB, ECC = < 0.1	DGS: improved BW and isolation

(continued)

Table 2 (continued)

Reference	Antenna size (mm^3)	Analyzed parameter	Technique
Shoaib et al. [33]	31.2 × 31.2 × 1.57	Diversity Gain = 15.8 dB, ECC = < 0.03, Bandwidth = 5.68 GHz, Isolation = < − 15 dB	Electromagnetic band gap ground plane: increase gain and isolation
Saif et al. [34]	13 × 25 × 0.254	BW = 10 GHz, Diversity Gain = 9.8 dB, ECC = 0.009, Isolation = < −20 dB	DGS: increase isolation
Lee et al. [35]	10.1 × 8.5 × 0.14	Bandwidth = 10 GHz	PSSP: increase BW

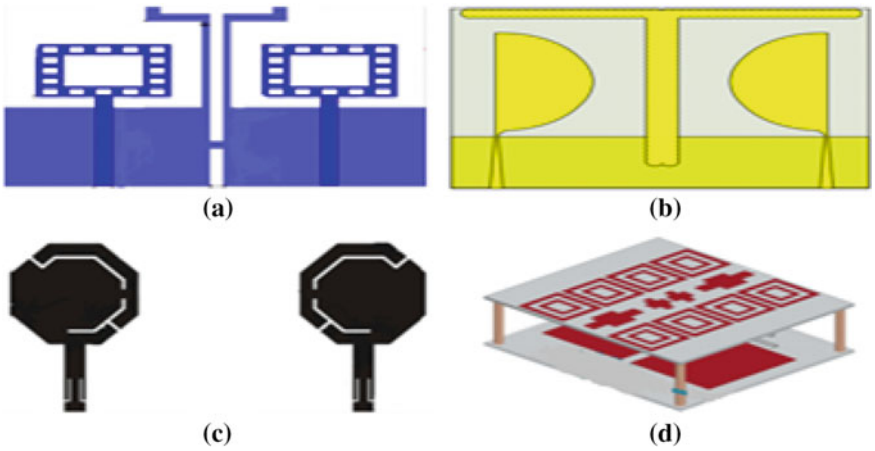


Fig. 5 **a** Fractal rectangular patch MIMO antenna [20]. **b** Semi-circular-shaped MIMO antenna [21]. **c** Octagonal patch MIMO antenna [22] **d** MIMO antenna with decoupling structure [23]

notched characteristic. Different shapes of stubs and slots are placed on patch and ground plane of two-element MIMO antenna to increase its bandwidth. The octagonal patch MIMO antenna geometry is presented in Fig. 5c. Niu et al. [23] presented a closely coupled MIMO antenna with decoupling structures (H-shaped defect ground structure (HDGS) and modified array antenna decoupling surface (MADS)), which will provide better isolation. The MIMO antenna geometry is presented in Fig. 5d. Ghannad et al. [24] proposed a modified two-element MIMO antenna, to improve isolation low profile feeding structure. Finally, a comparison is done between the proposed antenna and the other antennas, and it is found that it has better reduction in mutual coupling. The proposed antenna geometry is presented in Fig. 6a.

Dkiouak et al. [25] presents a dual band MIMO antenna which provides high isolation. The proposed antenna is designed by placement of two symmetrical rectangular patches with slot on the top side and two symmetrical slots on ground plane

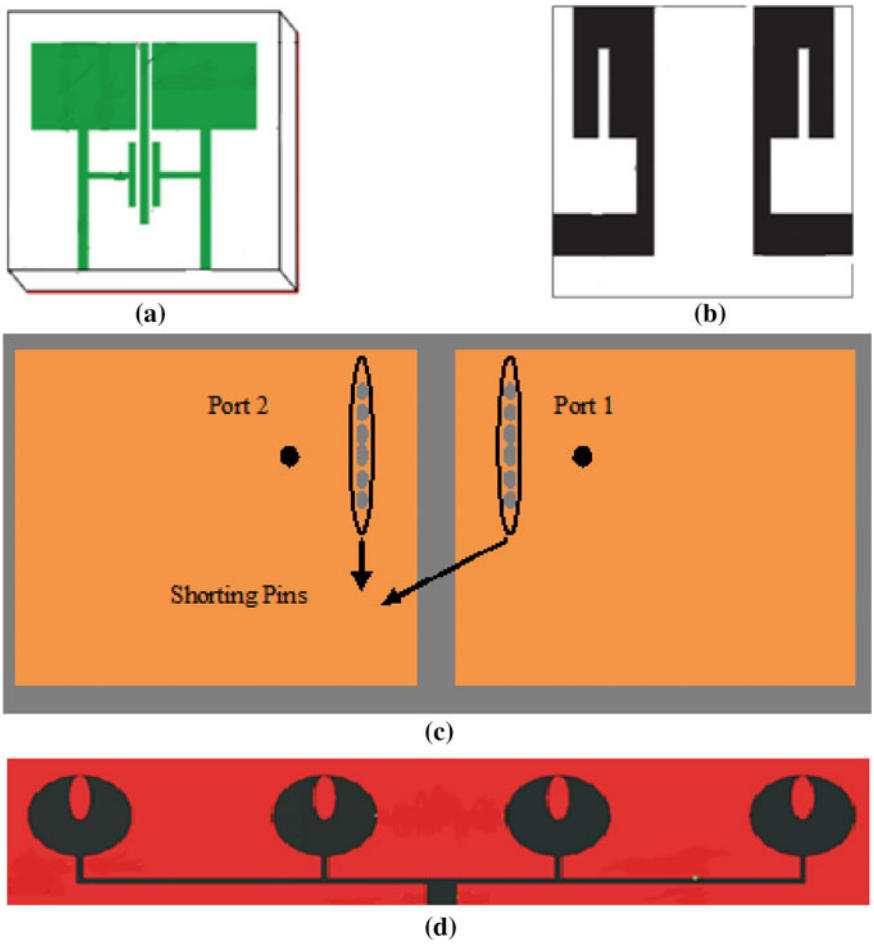


Fig. 6 **a** Rectangular patch MIMO antenna [24]. **b** Nine-shaped patch MIMO Antenna [25]. **c** Square patch MIMO Antenna [27]. **d** Geometry of Circular Patch Microstrip Antenna with Elliptical Slot [28]

[26]. The geometry of the proposed antenna is shown in Fig. 6b. Abdullah et al. [27] proposed a new technique of loading six metallic pins near the edges of the two square radiating patches for the enhancement of isolation. The square patch MIMO antenna geometry is presented in Fig. 6c. Khattak et al. [28] studied the design of a circular-shaped patch antenna with elliptical-shaped slot which provides dual band behavior. The comparison between a unit cell and array configured antenna is carried out. The efficiency, gain, and directivity increases in array configuration as compared to unit cell, but bandwidth decreases. The geometry of the antenna can be seen in Fig. 6d. Marzouk et al. [29] presented dual band MIMO antennas. Antenna is designed by using four element MIMO antenna with one slot of inverted I-shaped on the patch and

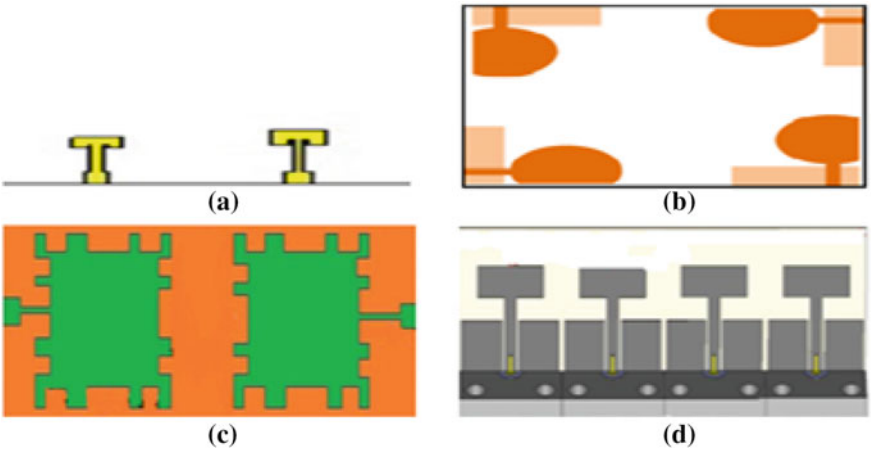


Fig. 7 **a** Rectangular patch MIMO antenna with slots [29]. **b** Beveled-shaped MIMO antenna [30]. **c** Minkowski-shaped MIMO antenna [31] **d** T-shaped rectangular patch MIMO antenna [32]

one slot on the ground plane which formed defective ground structure. The MIMO antenna geometry is presented in Fig. 7a. Firmansyah et al. [30] designed a 2 × 2 MIMO circular patch microstrip antenna with beveled half-cut structure.

Beveled means cutting one side of the antenna patch. By doing so, size of the antenna reduces [36]. The MIMO antenna geometry is presented in Fig. 7b. Babu and Anuradha [31] designed a good performance Minkowski shape MIMO antenna. Design of patch for the proposed antenna is obtained by cutting rectangular slots of different length sizes from the square patch which also provides miniaturization of mutual coupling. The geometry of the antenna can be seen in Fig. 7c. Jilani et al. [32] proposed a T-shaped rectangular patch millimeter wave MIMO antenna with five-split ring slots in the ground plane [37]. After the designing of single element antenna using iteration technique, four element MIMO antenna was configured. The T-shaped rectangular patch MIMO antenna geometry is presented in Fig. 7d.

Shoaib et al. [33] presented the design of H-shaped MIMO antenna with eight elements. For the enhancement of efficiency and gain, EM band gap structure is used for the designing of ground plane. Finally, a comparison is done with the other design, and it is found that the proposed antenna is compact in size. The H-shaped MIMO antenna geometry is presented in Fig. 8a. Saif et al. [34] presented the idea to design and fabricate a novel UWB 2 × 2 MIMO antenna for lower 5G band of frequency. The antenna is designed by using two "F"-shaped structures and with defective ground. The geometry of the antenna is shown in Fig. 8b.

Lee et al. [35] discussed a three-layer microstrip patch design with four parasitic surrounding stacked patches (PSSP) for 60 GHz. Parasitic patches were added to enhance the bandwidth of an antenna [38, 39] as shown in Fig. 8c. To further increase the antenna gain, 2 × 2 MIMO with PSSPs was proposed in this paper.

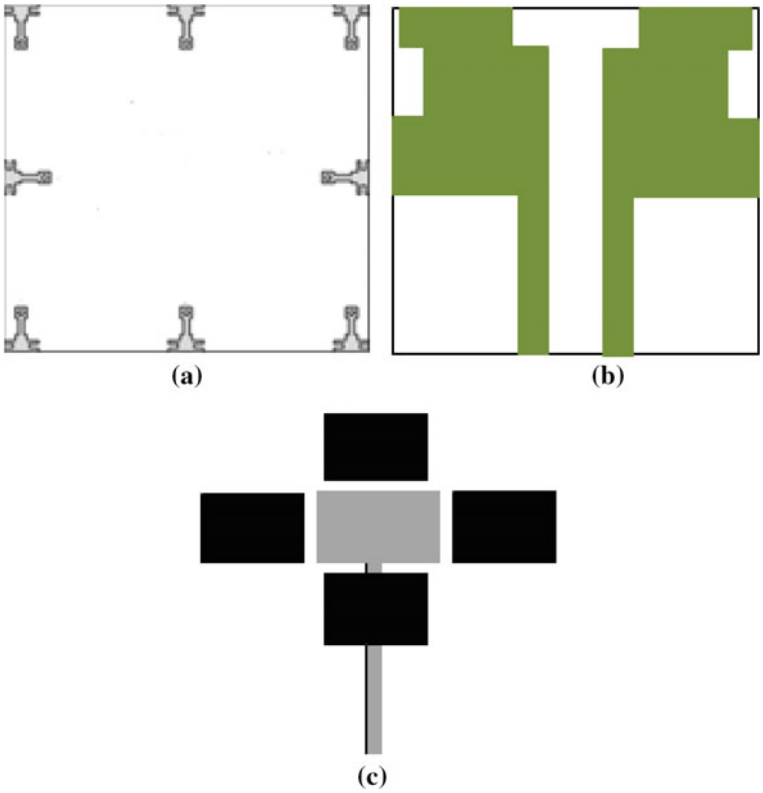


Fig. 8 **a** H-shaped MIMO Antenna [33]. **b** Geometry of Microstrip Antenna with F Type Slot [34]. **c** Microstrip patch antenna array with PSSP [35]

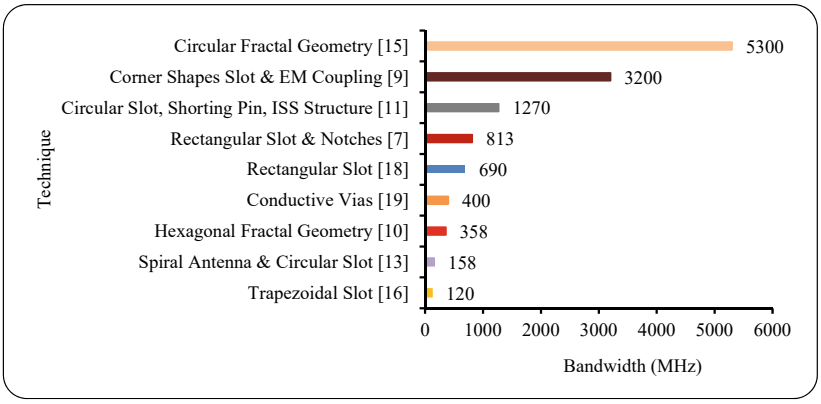
4 Discussion

Within the above parts, various techniques utilized in both antennas (microstrip patch and MIMO) for enhancing their performance were discussed. This segment offers a comparison among the various investigations of respective antennas considering different technical parameters, viz. bandwidth, gain and return loss of microstrip antenna, ECC, isolation and diversity gain of MIMO antenna. Figure 9 is prepared to show the difference among studies of microstrip patch antenna.

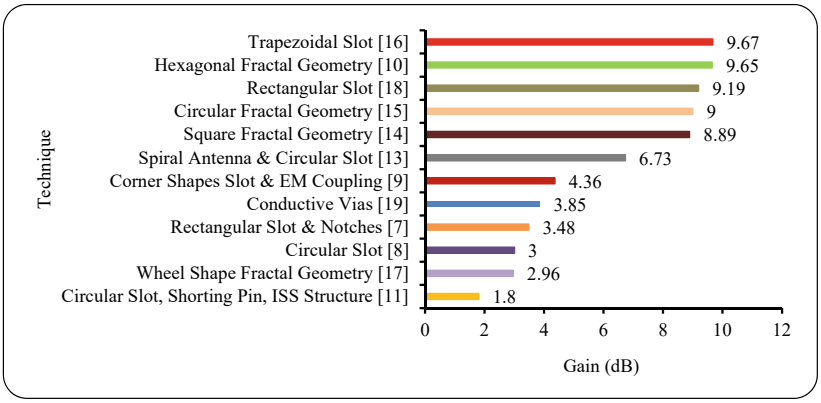
The comparison charts shown in Fig. 9 reveals the following:

(i) Of all the techniques, the usage of circular fractal geometry leads to enhance the bandwidth as compare to other techniques. The maximum bandwidth of 5.3 GHz is obtained in the investigation carried out by El-Khamy et al. [15].

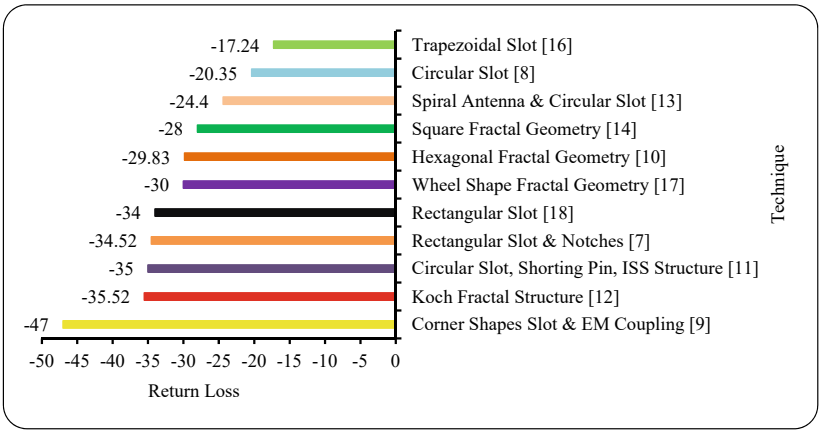
(ii) Among the various techniques, trapezoidal slot on patch contributes to maximum gain. The highest gain equals to 9.67 dB has been reported in [16].



(a)



(b)



(c)

Fig. 9 Comparison of investigations using different techniques in MSP antenna **a** Bandwidth. **b** Gain. **c** Return loss

(iii) Return loss is also improved by using corner shape slot on the patch [9].

These results show that using fractal structure geometry and insertion of slots on patch result in better bandwidth, gain and return loss in case of microstrip patch antenna. It may be due to the fact that in fractal geometry structure, changing the iteration numbers increases the electrical path length of an antenna which further enhances the impedance of microstrip patch antenna [40, 41]. This property of fractal structure provides enhanced bandwidth and the insertions of slot in the patch adjusts the current distribution so that at input point, impedance and current path length changes, and high impedance is achieved which provides better gain and return loss [42]. In addition to above, high data rate transmission and enhanced channel capability are also required in fastest wireless communication [43, 44]. These requirements are fulfilled with the help of MIMO antenna. Figures 10a, b and c present a comparison between the three parameters named as ECC, isolation, and diversity gain, respectively.

Following conclusions can be drawn from comparison charts shown in Figs. 10a, b and c:

(i) Isolation and ECC are often improved by providing a rectangular slot on the ground plane (DGS). The maximum isolation and ECC of –46.5 dB and 0.0002 are obtained in the investigation carried out by Ghannad et al. [24] and Dkiouak et al. [25], respectively.
(ii) The maximum value of diversity gain achieved is 15.8 dB by using electromagnetic band gap technique [33].

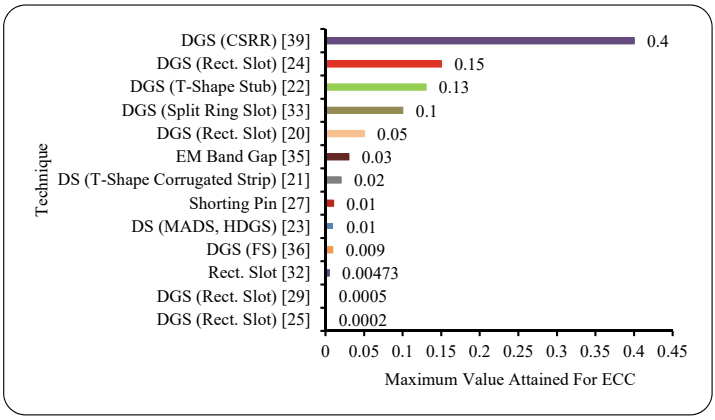
Defective ground structure disturbs the surface current, which further affects the impedance of transmission line. Moreover, by decreasing the current in ground plane, coupling between the adjacent element of MIMO antenna gets suppressed due to which isolation, ECC and diversity gain result in better values [45–47].

5 Conclusion and Future Scope

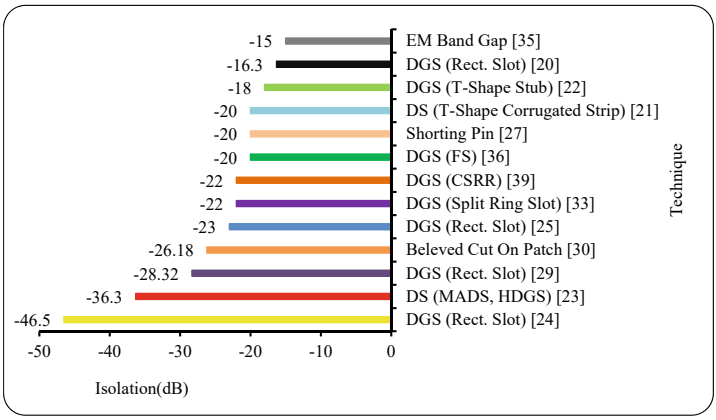
Adoption of microstrip to MIMO microstrip antenna is the primary focus of this review paper. Numerous simulated- and experimental-based investigations are presented and compared. It has been concluded that circular fractal geometry, insertion of trapezoidal slot, and corner shape slot on patch may be the effective techniques for achieving the good bandwidth, gain, and return loss, respectively, for microstrip patch antenna. Further, the reviewed literature suggests that DGS (rectangular slot on the ground plane) technique provides better isolation and ECC. Also, electromagnetic band gap technique generates high diversity gain for MIMO antenna.

In addition, the comprehensive literature review offered a thorough description into the research work to be carried out to improve the performance parameters of covered antennas. The following suggestions are given for potential improvements:

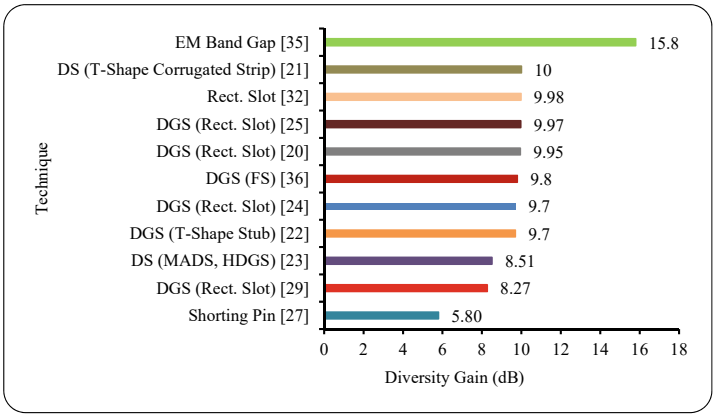
(i) Modification in the geometry of fractal structure may further be explored to achieve better results.



(a)



(b)



(c)

Fig. 10 Comparison of investigations using different techniques in MIMO antenna **a** ECC. **b** Isolation. **c** Diversity Gain

(ii) Different techniques, such as replacing conventional substrate with air substrate, by using reflecting layer, parasitic patches, etc., can be a great benefit in achieving the enhanced gain and bandwidth of microstrip antenna.
(iii) Some new techniques like loading of neutralization lines, placing of antenna elements periodically to each other, stepped ground lane, planar decoupling network, etc. may be used for designing a MIMO antenna with low mutual coupling, low ECC, and better gain.

References

1. S. Haykin, *Digital Communication.* (John Wiley & Sons, New York, 1988)
2. TRAI, GOI A white paper on enabling 5G in India, https://www.main.trai.gov.in. Accessed 22 Feb 2019
3. Study paper on multiple-input multiple-output (MIMO) technology, https://www.tec.gov.in. Accessed 10 Sep 2019
4. C.A. Balanis, *Antenna Theory, Analysis and Design.* (John Wiley & Sons, New York, 1997)
5. A. Kaur, P.K. Malik, Multiband elliptical patch fractal and defective ground structure microstrip patch antenna for wireless applications. Prog. Electromag. Res. B **91**, 157–173 (2021)
6. A. Rahim, P.K. Malik, V.A.S. Ponnapalli, Fractal antenna design for overtaking on highways in 5G vehicular communication Ad-hoc network environment. Int. J. Eng. Adv. Technol. 9(IS6), (2019)
7. R.K. Verma, D.K. Srivastava, Design and analysis of triple band rectangular microstrip antenna loaded with notches and slots for wireless applications. Wireless Pers. Commun. **114**, 1847–1864 (2020)
8. R.A. Panda, D. Mishra, Efficient design of bi circular patch antenna for 5G communication with mathematical calculations for resonant frequencies. Wireless Pers. Commun. 112, 717–727 (2020)
9. M.H. Sharaf, A.I. Zaki, R.K. Hamad, M.M.M. Omar, A Novel dual band (38/60 GHz) patch antenna for 5g mobile handsets. Sensor **20**(9), 2541 (2020)
10. D. Tiwari, J.A. Ansari, A. Saroj, M. Kumar, Analysis of a miniaturized hexagonal Sierpinski gasket fractal microstrip antenna for modern wireless communication. Int. J. Electron. Commun. 123, (2020)
11. A. Liu, L. Huang, L. Lu, Wideband circular patch antenna with i-shaped structure for horizontal omni directional gain enhancement. IET Microwaves Antennas Propag. **13**, 608–614 (2019)
12. M.G. Siddiqui, M.G. Saroj, A.K. Devesh, J.A. Ansari, Multi band fractaled triangular microstrip antenna for wireless application. Prog. Electromag. Res. 65, 51–60 (2018)
13. A. Aoad, M.S.U. Tureli, Design, simulation, and fabrication of a small size of a new spiral shaped of circular microstrip patch antenna. Microwave Optic. Technol. Lett. **60**, 2912–2918 (2018)
14. T. Devesh, J.A. Ansari, M.G. Siddiqui, A.K. Saroj, Analysis of modified square Sierpinski gasket fractal microstrip antenna for wireless communication. Int. J. Electron. Commun. 94, 377–385 (2018)
15. S.E. El-Khamy, A. Zaki, S. Hamdy, A. El- Khouly, A new fractal like tree structure of circular patch antennas for UWB and 5G multiband applications. Microwave Optic. Technol. Lett. 59, 2168–2174 (2017)
16. G.V. Raviteja, Design and analysis of a novel dual trapezoidal slot based rectangular microstrip antenna for wide area network using wimax application. Microwave Optic. Technol. Lett. **60**, 1057–1060 (2017)

17. M. Gupta, V. Mathur, Wheel shaped modified fractal antenna realization for wire-less communications. Int. J. Electron. Commun. **79**, 277–266 (2017)
18. S.J. Singh, G. Singh, G. Bharti, Circular microstrip antenna with fractal slots for multiband application. J. Instit. Eng. Ser. B–Springer, (2017)
19. N. Bisht, P. Kumar, A dual band fractal circular microstrip patch antenna for C-band application, in *Proceeding of Progress in Electromagnetics Research*, (Suzhou China, 2011), pp. 852–855
20. R. Gurjar, D.K. Upadhayay, B.K. Kanaujia, A. Kumar, A compact modified Sierpinski carpet fractal UWB MIMO antenna with square-shaped funnel-like ground stub. Int. J. Electron. Commun. 117, 153126 (2020)
21. H. Ullah, S.U. Rahman, Q. Cao, I. Khan, H. Ullah, Design of SWB MIMO antenna with extremely wideband isolation. Electronics **9**(2), 258 (2020)
22. A. Kumar, A.Q. Ansari, B.K. Kanaujia, J. Kishor, S. Kumar, An ultra-compact two-port UWB-MIMO antenna with dual band-notched characteristics. Int. J. Electron. Commun. 114, 152997 (2019)
23. Z. Niu, H. Zhang, Q. Chenand, T. Zhing, Isolation enhancement in closely coupled dual-band MIMO patch antennas. IEEE Antennas Wireless Propag. Lett. **18**, 1686–1690 (2019)
24. A.A. Ghannad, M. Khalily, P. Xiao, R. Tafazolli, A.A. Kishk, Enhanced matching and vialess decoupling of nearby patch antennas for MIMO system. IEEE Antennas Wireless Propag. Lett. **18**, 1066–1070 (2019)
25. A. Dkiouak, A. Zakriti, M.E. Quahabi, Design of a compact dual-band MIMO antenna with high isolation for WLAN and X -band satellite by using orthogonal polarization. J. Electromag. Waves Appl. 1–14 (2019)
26. P.K. Malik, D.S. Wadhwa, J.S. Khinda, A survey of device to device and coopera-tive correction for the future cellular network. Int. J. Wireless Inf. Netw. **27**, 411–432 (2020)
27. M. Abdullah, Q. Li, W. Xue, G. Peng, Y. He, X. Chen, Isolation enhancement of MIMO antennas using shorting pin. J. Electromagn. Waves Appl. **33**, 1249–1263 (2019)
28. M.I. Khattak, A. Sohail, U. Khan, Z. Barki, G. Witjaksono, Elliptical slot circular patch antenna array with dual band behavior for future 5G mobile communication networks. Prog. Electromagn. Res. C **89**, 133–147 (2019)
29. H.M. Marzouk, M.I. Ahmed, A.E.H. Shaalan, Novel dual-band 28/35 GHz MIMO antennas for 5G mobile application. Prog. Electromagn. Res. C **93**, 103–117 (2019)
30. T. Firmansyah, S. Praptodiyono, D. Aribowo, S. Alam, D.W. Astuti, M. Yunus, Bandwidth enhancement and miniaturization of circular-shaped microstrip antenna based on beleved half-cut structure for MIMO 2x2 application. Int. J. Electric. Comp. Eng. 9, 1110–1121 (2019)
31. K.V. Babu, P.B. Anuradha, Design of multi-band MIMO antenna to reduce mutual coupling. J. King Saud Univ. Eng. Sci. **1018–3639**(18), 30074–30076 (2018)
32. S. Jilani, A. Alomainy, Millimeter-wave T-shaped MIMO antenna with defected ground structures for 5G cellular networks. IET Microw Antenna Propag **12**, 672–677 (2017)
33. N. Shoaib, S. Shoaib, R.Y. Khattak, I. Shoaib, X. Chen, A. Perwaiz, MIMO antennas for smart 5G devices. IEEE Access **6**, 77014–77021 (2018)
34. H. Al-Saif, M. Usman, M.T. Chughtai, J. Nasir, Compact ultra-wide band MIMO antenna system for lower 5G bands. Wireless Commun. Mob. Comput. 1–6 (2018)
35. J. Lee, E.S. Li, H. Jin, C-Y. Li, K-S. Chin, 60 GHz wideband LTCC microstrip patch antenna array with parasitic surrounding stacked patches. IET Microw. Antennas Propag. 13, 35–41(2018)
36. N. Shaik, P.K. Malik, A retrospection of channel estimation techniques for 5G wire-less communications: opportunity and challenges. Int. J. Adv. Sci. Technol. **29**(5), 8469–8479 (2020)
37. S.V. Akram, P.K. Malik, R. Singh, G. Anita, S. Tanwar, Adoption of block-chain technology in various realms: opportunity and challenges. Sec. Priv. 3(5), (2020)
38. M. Abdullah, Y-L. Ban, K. Kang, M.O.Y. Li, M. Amin, Eight-element antenna array at 3.5 GHz for MIMO wireless application. Prog. Electromagn. Res. C 78, 209–216 (2017)
39. A. Ramachandran, S.V. Pushpakaran, M. Pezholil, V. Kesavath, A four port MIMO antenna using concentric square ring patches loaded with CSRR for high isolation. IEEE Antennas Wirel. Propag. Lett. **15**, 1196–1199 (2015)

40. A. Ramachandran, A.S. Mathew, V.P. Viswanathan, M. Pezholil, V. Kesavath, Diversity-based four port multiple input multiple output antenna loaded with inter digital structure for high isolation. IET Microw. Antenna Propag. **10**, 1633–1642 (2016)
41. R. Ghatak, D.R. Poddar, R.K. Mishra, Design of Sierpinski gasket fractal microstrip antenna using real coded genetic algorithm. IET Microw. Antenna Propag. 1133–1140 (2009)
42. B.K. Shukla, N. Kashyap, R.K. Baghal, Circular slotted elliptical patch antenna with elliptical notch in ground. Prog. Electromagn. Res. C **74**, 181–189 (2017)
43. E. Thakur, N. Jaglan, S.D. Gupta, Design of compact UWB MIMO antenna with enhance bandwidth. Prog. Electromagn. Res. C **97**, 83–94 (2019)
44. N.N. Lan, Gain enhancement in MIMO using defective ground structure technique. Prog. Electromagn. Res. M **87**, 127–136 (2019)
45. L. Xi, H. Zhai, L. Li, A low profile antenna system with compact new structure for reducing mutual coupling. J. Electromagn. Waves Appl. **33**, 71–83 (2018)
46. X. Chen, S. Zhang, Q. Li, A review of mutual coupling in MIMO systems. IEEE Access **6**, 24706–24719 (2018)
47. T.H. Jang, H.Y. Kim, D.M. Kang, S.H. Kim, C.S. Park, 60-GHz low profile, wideband dual polarized U-slot coupled patch antenna with high isolation. IEEE Trans. Antenna Propag. 16, 4453–4462 (2019)

Massive MIMO System—Overview, Challenges, and Course of Future Research

Shailender, Shelej Khera, Sajjan Singh, and Jyoti

Abstract In wireless communication industry, the shortage of worldwide bandwidth has inspired the research and development of technology for wireless access named as massive MIMO. It is a basic and fundamental technology of 5G network. Massive MIMO is a multi-antenna technique that deploys arrays of antennas at transmitter and receiver to support several users at the same time. As massive MIMO improves the spectrum efficiency and channel capacity of communication system, it also efficiently improves the link reliability and data transmission rate. A comprehensive review on massive MIMO system is presented in this paper outlining various challenges with their mitigation techniques and future research direction. In the end, we have comprehensively investigated the potential of machine learning and deep learning techniques in mitigating the issues of massive MIMO systems.

Keywords 5G · Precoding · Channel estimation · Pilot contamination · Massive MIMO · Machine learning · Spectral efficiency · Signal detection · Deep learning

1 Introduction

Massive MIMO is defined as a multi-user technology that develops wireless network terminals in high-mobility environments. The main concept behind the MIMO is to develop the capacity of the base stations by installing arrays of antennas. Massive MIMO is a high-end solution of multi-antenna evolution. It helps in serving several

Shailender (✉) · S. Khera
Lovely Professional University, Phagwara, Punjab, India
e-mail: Shailender.41900027@lpu.in

S. Khera
e-mail: shelej.22390@lpu.co.in

S. Singh
Chandigarh Group of Colleges, Jhanjeri, Mohali, India
e-mail: hodecejhanjeri@cgc.ac.in

Jyoti
Delhi Global Institute of Technology, Jhajjar, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_16

terminals simultaneously that are having similar time–frequency resources. So with massive MIMO, we can do both diversity gain and multiplexing gain. The terminology "massive" determines the number of antennas that are installed with MIMO communication and not their physical size [1].

The remaining part of the paper is arranged as follows: Sect. 2 offers insights into comprehensive overview of massive MIMO system. Section 3 explores the problems of massive MIMO systems and addresses several up-to-date methods of mitigation. The application of deep learning and machine learning is presented in Sect. 4 for massive MIMO system. Section 5 explains the dynamic topic of research on massive MIMO system for networks for future generations and Sect. 6 concludes the paper.

2 Massive MIMO

For 5G and the age of wireless communication beyond, massive MIMO is the most attractive technology. Massive MIMO is the development of modern MIMO systems for existing cellular networks that put collectively hundreds and perhaps thousands of base station's antennas and support tens of users at the same time [2, 3]. Massive MIMO allows antenna array to target a user with a narrow beam which increases spectral efficiency and throughput. For massive MIMO device, as the quantity of antenna increases, radiated beams turn narrower and concentrate on the consumer in space. Such spatially oriented antenna beams enhance the target user's throughput and decrease the neighboring user's interference [4]. Antenna array also focused narrow beam in small specific section which improves energy efficiency.

Uplink Transmission:

It uses the uplink channel to relay user terminal data and pilot signals to the base station during uplink transmission, as illustrated in Fig. 1a [5]. Here we are considering a massive MIMO uplink system fitted with M base station's antennas and interacting concurrently with K ($M \gg K$) users having a single antenna. If $x \in \mathbb{C}^K$ is the user-transmitted signal or the known pilot signal for estimating the channel then signal obtained during uplink at the base station is defined as

$$\mathrm{y} = \mathrm{H}_{\mathrm{x}} + n_{uplink}$$

where signal obtained at the base station is $y \in \mathbb{C}^M$. The channel vector matrix among the terminal of the user and the base station is represented by H, and element of $H \in \mathbb{C}^{M \times K}$ are independent and distributed identically (IIDs) with mean zero and variance of unity, i.e., $H \sim \mathcal{CN}(0, 1)$. The extra term nuplink $\in \mathbb{C}^M$ indicates the inclusion of interference from multiple transmissions and the noise of receiver. The added interference can be reliant on channel H but it is independent of signal x transmitted by user.

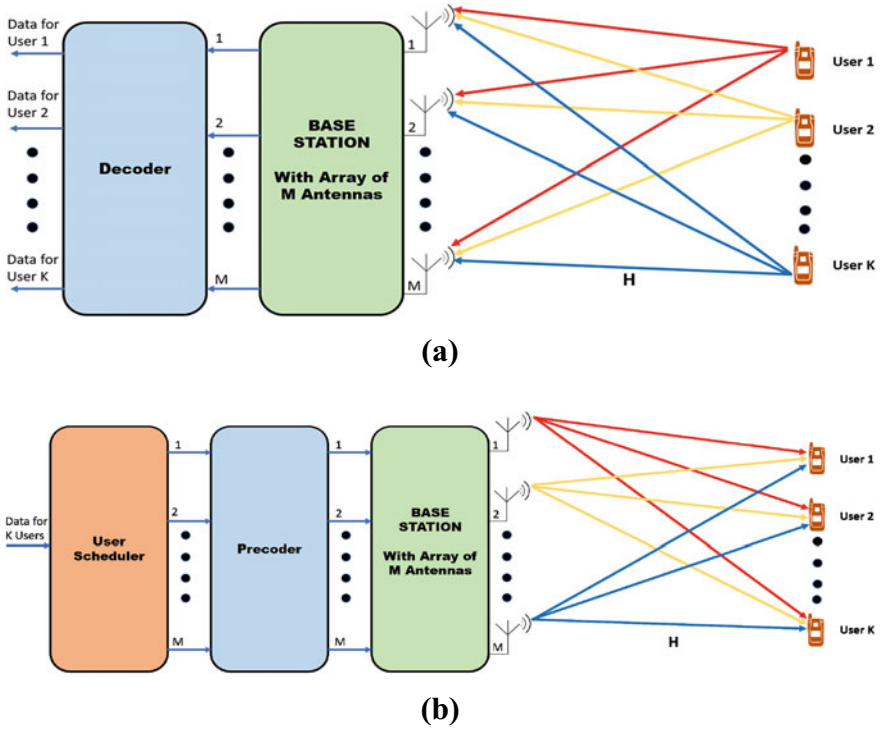


Fig. 1 **a** Operation of uplink in massive MIMO [5] and **b** Operation of downlink in massive MIMO [5]

$$n_{uplink} = n_{uplink-interference} + n_{noise}$$

Downlink Transmission:

For transmitting data or estimating the channel between the base station and user, downlink channel is used. Training pilots are used by base station for estimating the channel. Figure 1b [5] indicates a downlink transmission between terminals of users and the base station. Here we are considering a massive MIMO downlink system having base station fitted with M antennas, communicating simultaneously with K single antenna users. Independent information is sent by the base station to multiples users simultaneously. The received signal $y_k \in \mathbb{C}^{M\times 1}$ at the k_{th} user is

$$y_k = h_k X_k + n_{downlink}$$

Elements of channel vector h_k between base station and k_{th} user are independent and distributed identically (IIDs) with mean zero and variance of unity, i.e., $H \sim \mathcal{CN}(0, 1)$. Signal sent by base station for k_{th} user is $x_k \in \mathbb{C}^M$ and $n_{downlink}$ is the extra noise that is made up of the noise from the receiver $n_{noise} \sim \mathcal{C}(0,\sigma^2 I)$ and the

interference caused by simultaneous transmission to other users during downlink $n_{downlink-interference}$ is given as [5]

$$n_{downlink} = n_{downlink-interference} + n_{noise}$$

3 Massive MIMO Challenges and Techniques for Mitigation

Massive MIMO technology is more than a MIMO technology expansion, and still there are so many challenges and problems that has to be solved in order to make it a reality. For massive MIMO systems various basic challenges are displayed in Fig. 2 [5].

3.1 Pilot Contamination

For the system of massive MIMO, to get the channel's prediction, the base station requires the user terminal's channel response. When pilot signals are orthogonal in the home cell and adjacent cells, base station obtained the exact estimate of the channel.

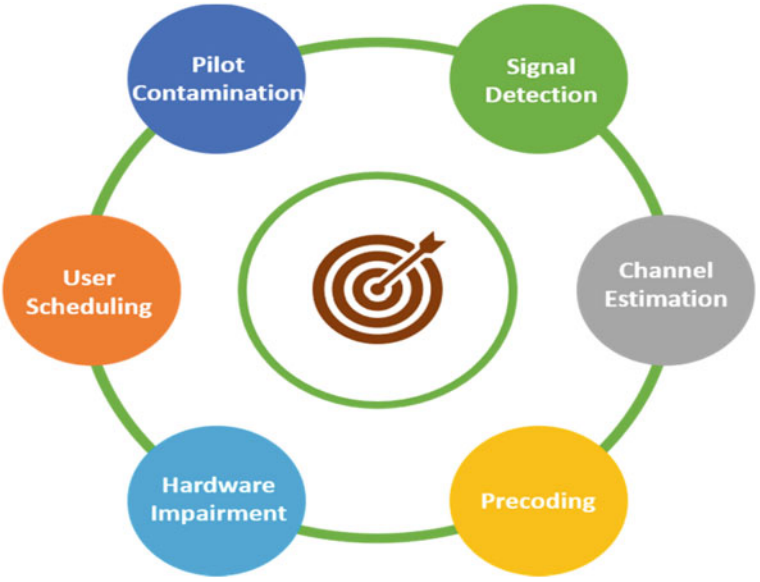


Fig. 2 Implementation issues for massive MIMO [5]

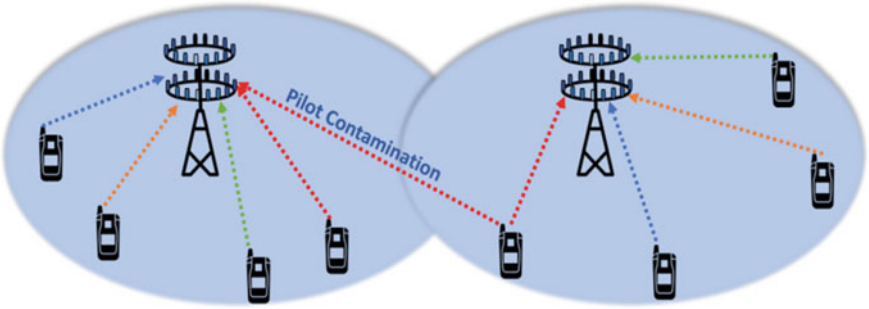


Fig. 3 Massive MIMO pilot contamination effect [7]

However, the orthogonal pilot number signals are limited in the specified bandwidth and duration, which forces the orthogonal pilots to reuse in adjacent cells [6]. Thus, using this same pilot sequence by two adjacent cells results in inter-user interference in the estimation of the channel, which is known as pilot contamination (PC). A linear combination of home cell's channel response and channel response of the adjacent cells will be received by base station due to interference caused by reusing orthogonal pilot signal in home cell and adjacent cell. This whole phenomenon is called as pilot contamination and, as shown in Fig. 3 [7], it limits the attainable throughput. The influence of the pilot contamination was studied in [8, 9] on system performance.

3.2 Channel Estimation

For massive MIMO system it is essential to have precise channel state information (CSI) for signal detection, decoding, beam forming data, and resource allocation. In wireless communication, Channel State Information (CSI) refers to the communication link's known channel properties. CSI explains how a signal propagates from the transmitter to the receiver, reflecting the cumulative effect of, for example, fading, scattering, and power delay with distance. CSI enables transmission to be adapted to the current channel condition, which is important for the multi- antenna system to achieve reliable communication with high data rates. It helps in enhancing the system performance by reducing the rate of bit error [10]. It includes the use of LTE cell-based references also known as pilot symbols so that there is a better estimation of channel characteristics.

Figure 4a indicates the FDD and TDD process in wireless communication, and the standard transmission of pilots and CSI feedback system in FDD and TDD mode is demonstrated in Fig. 4b [5].

Downlink data will be beam formed toward the user terminal by the base station using the calculated CSI. Since there is a small orthogonal number, the pilot contamination issue exists and is a major challenge during estimation of the channel of

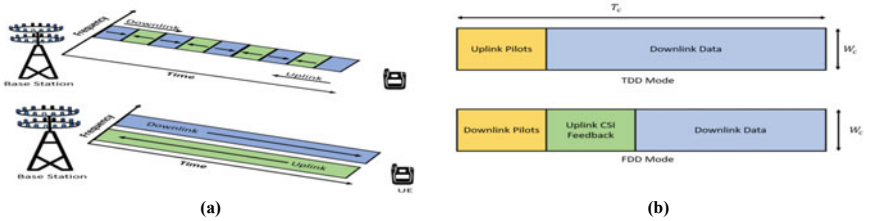


Fig. 4 **a** Duplexing of Frequency Division (FDD) and Duplexing of Time Division (TDD) mode: In TDD mode Massive operates best. **b** Normal transmission of pilot and Mechanism of CSI feedback for TDD and FDD mode [5]

massive MIMO system by pilots which can be reused from cell to cell. Increased complexity of computations and hardware are other problems due to more antennas. Therefore, for massive MIMO systems, channel estimation algorithms having low computational complexity and small overhead are extremely desirable [11]. While massive MIMOs are intended to use TDD, there has been a great deal of research using operations of FDD in massive MIMO systems.

3.3 Precoding

In multi-antenna systems, precoding is a beamforming concept that enables the transmission of multi-streams. Precoding refers to the multiple beam superposition of many data streams for spatial multiplexing. It is a preprocessing technique that performs transmit diversity by weighting the data stream. In order to gain channel pre-knowledge, the encoded information is transmitted by the transmitter to the receiver. In massive MIMO systems precoding plays an imperative function as it can mitigate the impact of interference and loss of paths and increase the throughput. The base station estimates the channel state information (CSI) in massive MIMO systems with the assistance of pilot signals during uplink or by the user terminal feedback. Because of the wireless channel's multiple environmental variables, the CSI obtained at the base station is not optimal and not uncontrollable [12]. Although a perfect CSI is not obtained by the base station, the downlink output at the base station is still mainly dependent on the CSI estimated.

In order to minimize interference and achieve spectral efficiency gains, the base station, thus, utilizes the approximate CSI and technique of precoding. Massive MIMO downlink performance depends on the precoding technique used and the precise measurement of CSI. While the precoding method offers enormous advantages for massive MIMO systems, by incorporating additional computations, overall system's complexity also increases. Therefore, powerful precoders with low complexity are more realistic to be used for system of massive MIMO. For a massive MIMO system, Fig. 5 [5] demonstrates the precoding with N-users and base station having M antennas.

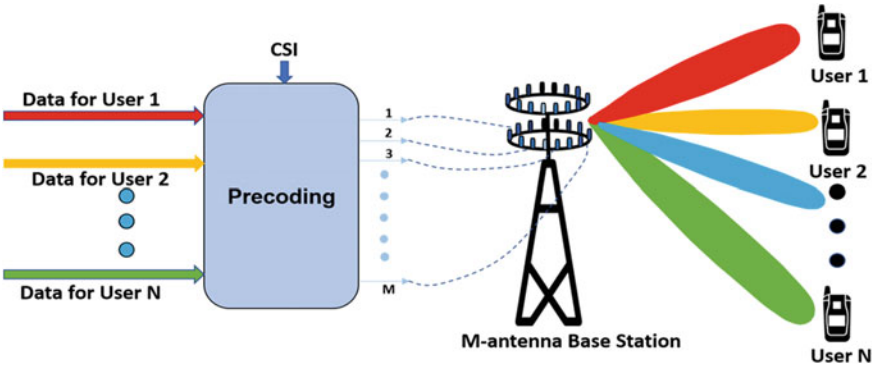


Fig. 5 Precoding for massive MIMO [5]

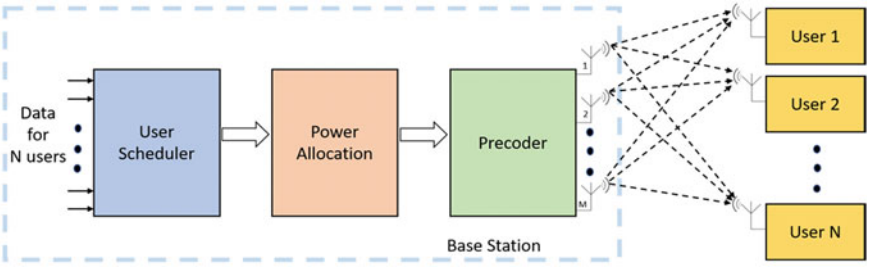


Fig. 6 Scheduling of users for massive MIMO [5]

3.4 User Scheduling

Massive MIMO having huge amount of base station antennas can interact simultaneously with several users. Concurrent contact with many users generates multi-user interference and the efficiency of the throughput degrades. During the downlink, precoding methods are applied to minimize the impact of interference from multi-users, as seen in Fig. 6 [5]. Because at the base station, the number of massive MIMO system antennas is limited, when the number of users exceeds the number of antennas, the correct scheme for user scheduling is implemented to attain higher throughput and sum rate efficiency before precoding.

3.5 Hardware Impairments

To decrease the impact of fading, noise, and interference, the massive MIMO system relies on a large number of antennas. In massive MIMO, a higher quantity of antennas increase the size of the device and increase the cost of hardware. It should be

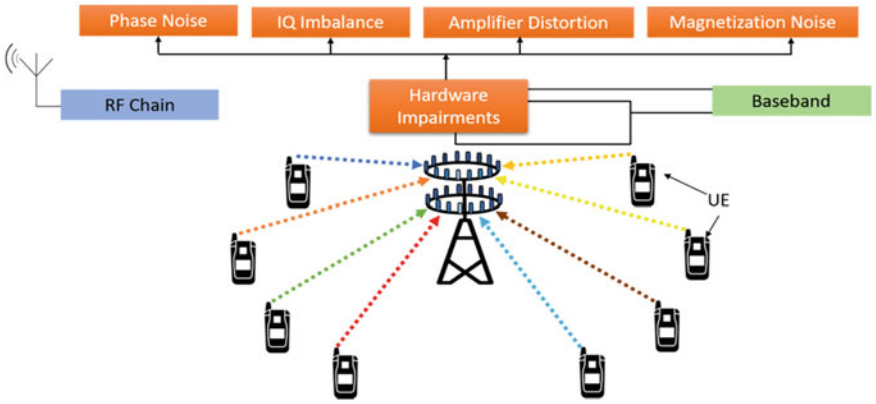


Fig. 7 Hardware impairments for Massive MIMO [5]

designed with small components and low costs to deploy massive MIMO, to minimize device complexity and the size of hardware. "The use of a low-cost component will increase the hardware imperfections such as phase noise, magnetization noise, amplifier distortion and IQ imbalance" [13]. These imperfections have a direct influence on the overall performance of the system. Mutual coupling is there among the antenna components due to a higher number of antennas, which affects the load impedance and induces distortions [14]. While the massive MIMO aims to minimize the 100 times more radiated power than traditional MIMO systems, and as the number of antennas increases, there is a linear increase in the power consumption of data converters and baseband hardware. Figure 7 demonstrates the impairment of hardware at base station of massive MIMO [5].

3.6 Signal Detection

Detection of uplink signal for massive MIMO system becomes computationally complicated and decreases the attainable throughput because of huge number of antennas. In addition, all user-transmitted signals are superimposed on the base station to produce interference, which further leads to a decrease in spectral efficiency and throughput. A system of massive MIMO with base station having M antennas and N user terminal is shown in Fig. 8 [5]. At base station superposition of all signals, coming from N user terminal through various wireless route makes signal detection difficult and inefficient at base station. For massive MIMO system comprehensive research has been carried out to find the optimum method of signal detection which can deliver higher throughput efficiency with less complication of computation (Table 1).

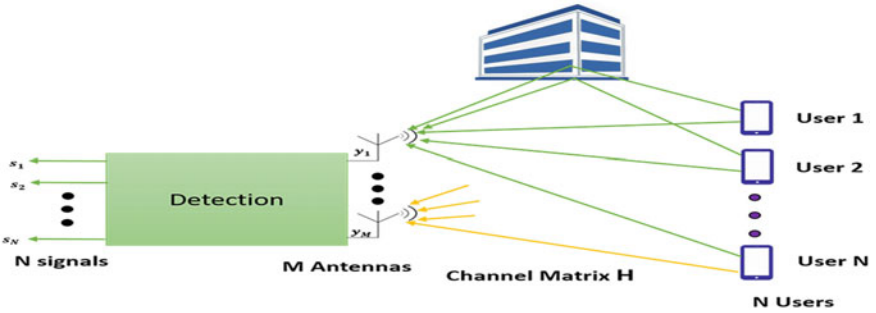


Fig. 8 Uplink system for massive MIMO [5]

Table 1 Mitigation techniques for various challenges in massive MIMO system

Challenges	Mitigation techniques
Pilot contamination	Estimation based on Pilot [15, 16], Reuse pilot [17], Partial sound resource [18], Precoding scheme for pilot contamination [19], Decontamination by a blind pilot [20, 21], Smart scheme for pilot allocation [22]
Channel estimation	LS [23], MMSE [24, 25], Enhanced MMSE [26, 27], DNN-based estimation[28], Blind channel estimation [29, 30], CSI acquisition and feedback through DL [31], MICED [32], Untrained DNN [33], Compressive sensing based adaptive channel estimation [34, 35], Blind convolutional denoising [36], Channel estimation with pilot reuse [37], VAMP [38], Sparse channel estimation using deep learning [39], Estimation based on CNN [40], Estimation based on machine learning [41, 42], Estimation with deep learning [43, 44]
Precoding	Dirty paper precoding [45], TH [46, 47], VP [48], ZF [49, 50], WF [51], Non-iterative hybrid precoding [52], Complex regularized ZF precoding [53], DLQP hybrid precoder [39]
User scheduling	ZF [54], MMSE [55], User scheduling and joint antenna selection [56], Greedy [57], Multi-user scheduling [58], Joint user scheduling [59], Machine learning supported user scheduling [60]
Hardware impairments	Phase noise compensation using Kalman tracking [61], Minimization of phase noise errors [62], Energy efficient scheme with hardware impairments [63]
Signal detection	Sphere decoder [64], SIC [65], MMSE [66], Conjugate gradient [67], NSA [68], Richardson method [69], SOR [23], Jacobi [70], Gauss Siedel [71], LS regression selection [72], Huber ADMM [73], AMP [74], Adaptive scheme based on compressive sensing [35], CNN [75], Gauss Siedel refinement [76], Semi-supervised and supervised learning based detection [77, 78]

4 Machine Learning and Deep Learning for Massive MIMO Systems

Machine learning is an artificial intelligence subset, regarded to be a strong instrument for problems of prediction and classification. Deep learning is sub-set of machine learning, using more specialized methods that are able to construct general functions and universal classifiers. In fields such as network security, processing natural languages and automated processes, these new ideas have been used widely. Machine learning as well as deep learning are actually very important technologies for 5G and 6G network design. Massive MIMO needs highly complicated optimizations, and the conventional algorithm like game theory and geometry of stochastic is very sophisticated and requires immense computational resources. In this complex analysis, the dynamic behavior of algorithms for deep learning and machine learning can be instrumental. and could save a large amount of computing power [79]. For massive MIMO beam forming, load balancing, signal detection, channel estimation, and optimization of the spectrum available, these algorithms of deep learning and machine learning are extremely useful [80, 81].

The precise channel prediction through machine learning greatly improves the performance of the massive MIMO system. Figure 9 illustrates the application of deep learning or machine learning for the estimate of channel in massive MIMO [5]. For channel estimation, the convolutional neural network (CNN) technique was used by the authors of reference [40], but optimal efficiency was not attained. Reference [41] has explored the application of machine learning for estimation of channel in conditions of the complex channel model. Compared to traditional channel estimation algorithms, more precise channels were predicted by deep learning-based channel estimation [82]. A channel in massive MIMO was considered as a picture by the authors of Reference [43] and denoising technique and image superposition based

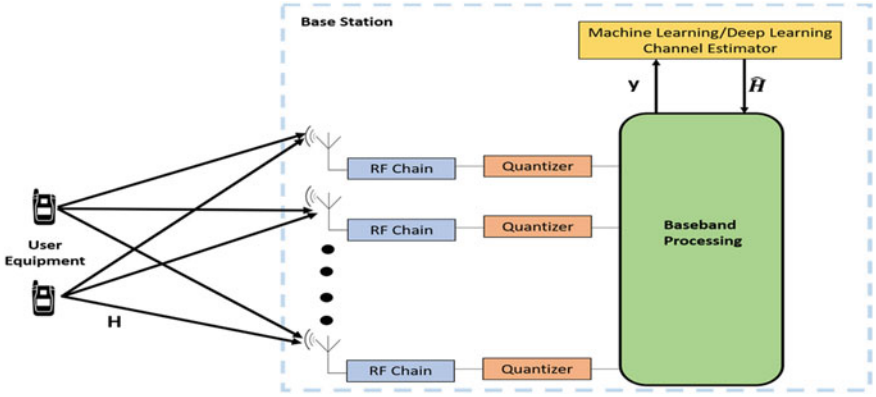


Fig. 9 Deep learning and machine learning based channel estimation for Massive MIMO system [5]

on deep learning was applied. Several additional researches have been done in order to create a deep neural network (DNN) end-to-end architecture to change the base station and UE modules [83]. For different scenarios, estimation of channel based on deep learning has been provided in the reference [44], and the findings were similar to those of the optimum algorithm for MMSE. In massive MIMO systems, during CSI estimation, algorithms for machine learning can minimize overhead in channel estimation. Sparse channel estimation approaches based on deep learning and their benefits over conventional methods of estimation have been discussed in reference [39]. Considering the property of channel ageing of massive MIMO, the problem of CSI estimation can be regarded as a time-series learning issue. An important method for solving this time-series learning problem is the recurrent neural network (RNN). In wireless communication for predicting distant data basic RNN tools are less effective because CSI estimation has distant data. "Therefore, various architectures have recently been proposed in massive MIMO, such as long short-term memory (LSTM) and non-linear autoregressive network with exogenous inputs (NARX), that deal with this problem of distant data" [84, 85]. In reference [42], prediction of a channel based on machine learning has been studied in massive MIMO having properties of aging of the Channel.

CNN was analyzed in combination with the autoregressive network (ARN) and the RNN in reference [42]. For massive MIMO systems a less complex scheduling scheme is provided by the machine learning supported user scheduling approach introduced in reference [60]. In massive MIMO reference author [86] proposed deep learning based new frequency and space channel mapping. In massive MIMO many techniques of deep learning and machine learning are too helpful for detection of uplink signal. For large-sized antenna systems such as massive MIMO, traditional methods for detection of signal are highly complicated and ineffective in computational terms. "Several approaches to semi-supervised learning (SSL) and supervised learning (SL) has been suggested, producing more rigorous results" [77, 78].

5 Area of Research for Massive MIMO Used in 5G and Beyond Networks

While massive MIMO have enormous advantages, there are still numerous issues, for example, signal detection, precoding, estimation of channel, user scheduling, pilot contamination, energy efficiency, and hardware impairments. In a real-world environment, these challenges need to be tackled and checked before we accomplish the promised benefits. These implementation problems in massive MIMO have prompted both industry and academia to concentrate on this system of massive MIMO. In addition, emerging innovation such as millimeter waves, massive MIMO, visible-light communication, ultra-massive MIMO, and terahertz waves need a great deal of study before these are introduced in our existing wireless system. In massive MIMO, some of the potential research subjects for 5G and beyond networks are

- To minimize the effect of noise, interference, and fading, the massive MIMO systems rely on huge number of antennas. In massive MIMO, huge number of antennas raises the device complexity and boosts the price of hardware. It must be designed with components which are small and having low cost, to decrease the size of hardware and complexity of computation to deploy massive MIMO. Imperfections of hardware such as noise of magnetization, phase noise, imbalance of IQ, and amplifier distortion would be increased by low-cost facilities. While hardware impairment cannot be fully eliminated, with the proper use of compensation algorithms, its effect can be mitigated. The designing of these algorithms for compensation in massive MIMO is a strong research area.
- As the number of orthogonal pilots available at any given time is limited, then due to this pilot contamination is one of the major obstacles in the deployment of massive MIMO. Interference is increased by pilot contamination, and it also restricts the attainable throughput. To reduce the effects of pilot contamination, numerous studies have been performed. Nevertheless, there is a need for an optimized process that mitigates its impact [15–22]. Therefore, an important area to investigate is efficient methods for minimizing the impact of pilot contamination.
- Although precoding methods enhance throughput and decrease interference, but it also raises the overall system's complexity of computation by incorporating additional computation. This complication of computation grows with a considerable number of antennas. The use of powerful precoders with low complexity in massive MIMO is, thus, more realistic. An important field of research is finding effective precoding methods for massive MIMO.
- The number of massive MIMO system antennas is limited, when number of users exceed the number of antennas, the correct scheme for user scheduling is implemented to attain higher throughput and sum rate efficiency before precoding. In order to discover a more effective design of a fair algorithm of scheduling which can deliver more data rate and ensure user equity, further research should be carried out.
- All user-transmitted signals are superimposed on the base station producing interference, which further leads to decrease in efficiency of spectral and throughput. Near-optimal performance has been achieved by a recent experiment, but the realization of massive MIMO [23, 35, 64–74] requires more powerful algorithms. Finding a more effective and low-complex algorithm for uplink signal detection is one of the key areas of study.
- In massive MIMO, accurate CSI is required for detection of user signal, data beamforming, and allocation of resources [87]. Signals from huge number of base station antennas has to be estimated by user terminals. In addition, the overhead of the pilot is also rising significantly. Therefore, an effective channel estimation method with low overhead of pilot, particularly for the FDD system, is an interesting field for research.
- Combining massive MIMO with quantum communication at frequencies above 300 GHz would be an interesting field for research.
- Given that in the existing market, phones that are available do not support the infrastructure of massive MIMO. Creating cheaper phones that can afford this

technology would be a challenge for Smartphone manufacturers. The massive MIMO design framework that can be incorporated with the existing 4G network is an outstanding field of research.

- A fascinating area of research is also the study of future potential technologies enabling 6G networks, for example, holographic radio, visible-light communication, and THz communication.

6 Conclusion

A powerful cellular spectrum capable of supporting the enormous rise in traffic for wireless data is desperately needed. The solution to this worldwide demand is the massive MIMO wireless technology. Massive MIMO technology puts together transmitter antennas and receiver antennas for providing higher energy and spectral efficiency by reasonably easy processing. The technology of massive MIMO has been researched in limited amounts, considering the global need for an effective spectrum. Thus, this new wireless access technology is still facing some open research challenges.

A comprehensive overview of massive MIMO systems is given in this paper. While massive MIMO provides enormous advantages, there are still many implementation problems for 5G and 6G networks, such as precoding, pilot contamination, user scheduling, channel estimation, signal detection, hardware impairments, and energy management, which must be resolved before we can attain the promised benefits. In addition, for massive MIMO system this paper also addresses the usage of machine learning technology and deep learning. We expect that the researchers presently working on 5G and networks beyond will be inspired by this paper to discover new direction researchers and open issues to solve in the upcoming years.

References

1. H. Ngo, A. Ashikhmin, H. Yang, E. Larsson, T. Marzetta, Cell-free massive MIMO versus small cells. IEEE Trans. Wireless Commun. **16**, 1834–1850 (2017)
2. E. Björnson, E. Larsson, T. Marzetta, Massive MIMO: ten myths and one critical question. IEEE Commun. Mag. **54**, 114–123 (2016)
3. T. Van Chien, E. Björnson, Massive MIMO communications
4. E. Larsson, O. Edfors, F. Tufvesson, T. Marzetta, Massive MIMO for next generation wireless systems. IEEE Commun. Mag. **52**, 186–195 (2014)
5. R. Chataut, R. Akl, Massive MIMO systems for 5G and beyond networks—overview, recent trends, challenges, and future research direction. Sensors. **20**, 2753 (2020)
6. L. Lu, G. Li, A. Swindlehurst, A. Ashikhmin, R. Zhang, An overview of massive MIMO: benefits and challenges. IEEE J. Select. Top. Signal Proc. **8**, 742–758 (2014)
7. J. Jose, A. Ashikhmin, T. Marzetta, S. Vishwanath, Pilot contamination and precoding in multi-cell TDD systems. IEEE Trans. Wireless Commun. **10**, 2640–2651 (2011)

8. B. Gopalakrishnan, N. Jindal, An analysis of pilot contamination on multi-user MIMO cellular systems with many antennas. in *2011 IEEE 12th International Workshop on Signal Processing Advances in Wireless Communications.* (2011)
9. N. Krishnan, R. Yates, N. Mandayam, Cellular systems with many antennas: large system analysis under pilot contamination. in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton).* (2012)
10. M. Moqbel, W. Wangdong, A. Ali, MIMO channel estimation using the LS and MMSE algorithm. IOSR J. Electron. Commun. Eng. 12, 13–22 (2017)
11. X. Wu, L. Gu, W. Wang, X. Gao, Pilot design and AMP-based channel estimation for massive MIMO-OFDM uplink transmission. in *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC).* (2016)
12. N. Fatema, G. Hua, Y. Xiang, D. Peng, I. Natgunanathan, Massive MIMO linear precoding: a survey. IEEE Syst. J. **12**, 3920–3931 (2018)
13. C. Studer, M. Wenk, A. Burg, MIMO transmission with residual transmit-RF impairments. in *2010 International ITG Workshop on Smart Antennas (WSA).* (2010)
14. U. Gustavsson, C. Sanchez-Perez, T. Eriksson, F. Athley, G. Durisi, P. Landin, K. Hausmair, C. Fager, L. Svensson, On the impact of hardware impairments on massive MIMO. in *2014 IEEE Globecom Workshops (GC Wkshps).* (2014)
15. F. Fernandes, A. Ashikhmin, T. Marzetta, Inter-cell interference in noncooperative TDD large scale antenna systems. IEEE J. Sel. Areas Commun. **31**, 192–201 (2013)
16. K. Appaiah, A. Ashikhmin, T. Marzetta, Pilot contamination reduction in multi-user TDD systems. in *2010 IEEE International Conference on Communications.* (2010)
17. V. Saxena, G. Fodor, E. Karipidis, Mitigating pilot contamination by pilot reuse and power control schemes for massive MIMO systems. in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring).* (2015)
18. T. Lee, H. Kim, S. Park, S. Bahk, Mitigation of sounding pilot contamination in massive MIMO systems. in 2014 IEEE International Conference on Communications (ICC). (2014)
19. Y. Li, Y-H. Nam, B.L. Ng, J. Zhang, A non-asymptotic throughput for massive MIMO cellular uplink with pilot reuse. in *2012 IEEE Global Communications Conference (GLOBECOM).* (2012)
20. R. Muller, L. Cottatellucci, M. Vehkapera, Blind pilot decontamination. IEEE J. Select. Top. Signal Proc. **8**, 773–786 (2014)
21. L. Cottatellucci, R. Muller, M. Vehkapera, Analysis of pilot decontamination based on power control. in *2013 IEEE 77th Vehicular Technology Conference (VTC Spring).* (2013)
22. F. Banoori, J. Shi, K. Khan, R. Han, M. Irfan, Pilot contamination mitigation under smart pilot allocation strategies within massive MIMO-5G system. Phys. Commun. 47, 101344 (2021)
23. X. Gao, L. Dai, Y. Hu, Z. Wang, Z. Wang, Matrix inversion-less signal detection using SOR method for uplink large-scale MIMO systems. in *2014 IEEE Global Communications Conference.* (2014)
24. H. Ngo, E. Larsson, T. Marzetta, The multicell multiuser MIMO uplink with very large antenna arrays and a finite-dimensional channel. IEEE Trans. Commun. **61**, 2350–2361 (2013)
25. A. Adhikary, A. Ashikhmin, T. Marzetta, Uplink interference reduction in large-scale antenna systems. IEEE Trans. Commun. **65**, 2194–2206 (2017)
26. K. Li, X. Song, M. Ahmad, M. Swamy, An improved multicell MMSE channel estimation in a massive MIMO system. Int. J. Antennas Propag. **2014**, 1–9 (2014)
27. A. Almamori, S. Mohan, Improved MMSE channel estimation in massive MIMO system with a method for the prediction of channel correlation matrix. in *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC).* (2018)
28. X. Ma, Z. Gao, Data-driven deep learning to design pilot and channel estimator for massive MIMO. IEEE Trans. Veh. Technol. **69**, 5677–5682 (2020)
29. J. Jiang, T. Lin, S. Phoong, New subspace-based blind channel estimation for orthogonally coded MIMO-OFDM systems. in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* (2014)

30. C. Shin, R. Heath, E. Powers, Blind channel estimation for MIMO-OFDM systems. IEEE Trans. Veh. Technol. **56**, 670–685 (2007)
31. M. Boloursaz Mashhadi, D. Gündüz, Deep learning for massive MIMO channel state acquisition and feedback. J. Indian Inst. Sci. **100**, 369–382 (2020)
32. D. Verenzuela, E. Bjornson, X. Wang, M. Arnold, S. ten Brink, Massive-MIMO iterative channel estimation and decoding (MICED) in the uplink. IEEE Trans. Commun. **68**, 854–870 (2020)
33. E. Balevi, A. Doshi, J. Andrews, Massive MIMO channel estimation with an untrained deep neural network. IEEE Trans. Wireless Commun. **19**, 2079–2090 (2020)
34. Z. Albataineh, K. Hayajneh, H. Bany Salameh, C. Dang, A. Dagmseh, Robust massive MIMO channel estimation for 5G networks using compressive sensing technique. AEU–Int. J. Electron. Commun. 120, 153197 (2020)
35. M. Ke, Z. Gao, Y. Wu, X. Gao, R. Schober, Compressive sensing-based adaptive active user detection and channel estimation: massive access meets massive MIMO. IEEE Trans. Signal Process. **68**, 764–779 (2020)
36. Y. Jin, J. Zhang, B. Ai, X. Zhang, Channel estimation for mmWave massive MIMO with convolutional blind denoising network. IEEE Commun. Lett. **24**, 95–98 (2020)
37. Y. Chen, L. You, X. Gao, X. Xia, Channel estimation with pilot reuse in IQ imbalanced massive MIMO. IEEE Access **8**, 1542–1555 (2020)
38. Z. Zhang, X. Cai, C. Li, C. Zhong, H. Dai, One-bit quantized massive MIMO detection based on variational approximate message passing. IEEE Trans. Signal Process. **66**, 2358–2373 (2018)
39. W. Ma, C. Qi, Z. Zhang, J. Cheng, Sparse channel estimation and hybrid precoding using deep learning for millimeter wave massive MIMO. IEEE Trans. Commun. **68**, 2838–2849 (2020)
40. D. Neumann, T. Wiese, W. Utschick, Learning the MMSE channel estimator. in *IEEE Transactions on Signal Processing*. (2018), pp 1–1
41. J. Yuan, H. Ngo, M. Matthaiou, Machine learning-based channel estimation in massive MIMO with channel aging. in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. (2019)
42. J. Yuan, H. Ngo, M. Matthaiou, machine learning-based channel prediction in massive MIMO with channel aging. IEEE Trans. Wireless Commun. **19**, 2960–2973 (2020)
43. M. Soltani, V. Pourahmadi, A. Mirzaei, H. Sheikhzadeh, Deep learning-based channel estimation. IEEE Commun. Lett. **23**, 652–655 (2019)
44. Q. Hu, F. Gao, H. Zhang, S. Jin, G. Li, Deep learning for channel estimation: interpretation, performance, and comparison. IEEE Trans. Wireless Commun. **20**, 2398–2412 (2021)
45. M. Costa, Writing on dirty paper (Corresp.). IEEE Trans. Inform. Theory 29, 439–441 (1983)
46. F. Cao, N. Dao, Receive beamformer design for MMSE multiuser tomlinson-harashima precoding. IEEE Commun. Lett. **15**, 293–295 (2011)
47. K. Zu, R. de Lamare, M. Haardt, Multi-branch tomlinson-harashima precoding design for MU-MIMO systems: theory and algorithms. IEEE Trans. Commun. **62**, 939–951 (2014)
48. F. Boccardi, G. Caire, The$p$-sphere encoder: peak-power reduction by lattice precoding for the MIMO Gaussian broadcast channel. IEEE Trans. Commun. **54**, 1703–1703 (2006)
49. S. Shi, M. Schubert, MMSE transmit optimization for multiuser multiantenna systems. in *Proceedings (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing*. (2005)
50. H. Bahrami, T. Le-Ngoc, Maximum ratio combining precoding for multi-antenna relay systems. in *2008 IEEE International Conference on Communications*. (2008)
51. V. Crasmariu, M. Arvinte, A. Enescu, S. Ciochina, Waterfilling power allocation applied in block-diagonalization Multi-User MIMO precoding technique. in *2016 12th IEEE International Symposium on Electronics and Telecommunications (ISETC)*. (2016).
52. N. Sharma, B. Gupta, Low-complexity non-iterative hybrid precoding scheme for millimeter wave massive MIMO systems. Int. J. Electron. 1–17 (2021)
53. M. Mostafa, F. Newagy, I. Hafez, Complex regularized zero forcing precoding for massive MIMO systems. Wireless Pers. Commun. (2021)

54. H. Yang, User scheduling in massive MIMO. in *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. (2018)
55. T. Lyu, Capacity of multi-user MIMO systems with MMSE and ZF precoding. in *2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. (2016)
56. M. Guo, M. Gursoy, Statistical learning based joint antenna selection and user scheduling for single-cell massive MIMO systems. IEEE Trans. Green Commun. Netw. **5**, 471–483 (2021)
57. R. Elliott, S. Sigdel, W. Krzymien, M. Al-Shalash, A. Soong, Genetic and greedy user scheduling for multiuser MIMO systems with successive zero-forcing
58. X. Meng, Y. Chang, Y. Wang, J. Wu, Multi-user grouping based scheduling algorithm in massive MIMO uplink networks. in *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*. (2018)
59. Z. Jiang, S. Chen, S. Zhou, Z. Niu, Joint user scheduling and beam selection optimization for beam-based massive MIMO downlinks. IEEE Trans. Wireless Commun. **17**, 2190–2204 (2018)
60. J. Shi, W. Wang, J. Wang, X. Gao, Machine learning assisted user-scheduling method for massive MIMO system. in *2018 10th International Conference on Wireless Communications and Signal Processing (WCSP)*. (2018)
61. S. Bittner, A. Frotzscher, G. Fettweis, E. Deng, Oscillator phase noise compensation using Kalman tracking. in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. (2009)
62. P.I. Tebe, G. Wen, J. Li, Y. Huang, A.E. Ampoma, K.O. Gyasi, Massive MIMO with transceiver hardware impairments: performance analysis and phase noise error minimization. in *KSII Transactions on Internet and Information Systems*. (2019), p 13
63. Z. Liu, C. Lee, W. Xu, S. Li, Energy-efficient design for massive MIMO with hardware impairments. IEEE Trans. Wireless Commun. **20**, 843–857 (2021)
64. Y. Jeon, N. Lee, S. Hong, R. Heath, One-bit sphere decoding for uplink massive MIMO systems with one-bit ADCs. IEEE Trans. Wireless Commun. **17**, 4509–4521 (2018)
65. A. Elghariani, M. Zoltowski, Successive interference cancellation for large-scale MIMO OFDM. in 2015 IEEE international conference on electro/information technology (EIT). (2015)
66. Y. Liang, G. Pan, Z. Bai, Asymptotic performance of MMSE receivers for large systems using random matrix theory. IEEE Trans. Inf. Theory **53**, 4173–4190 (2007)
67. B. Yin, M. Wu, J. Cavallaro, C. Studer, VLSI design of large-scale soft-output MIMO detection using conjugate gradients. in 2015 IEEE International Symposium on Circuits and Systems (ISCAS). (2015)
68. O. Gustafsson, E. Bertilsson, J. Klasson, C. Ingemarsson, Approximate neumann series or exact matrix inversion for massive MIMO? in *2017 IEEE 24th Symposium on Computer Arithmetic (ARITH)*. (2017)
69. B. Kang, J. Yoon, J. Park, Low-complexity massive MIMO detectors based on richardson method. ETRI J. **39**, 326–335 (2017)
70. J. Zhou, Y. Ye, J. Hu, Biased MMSE soft-output detection based on Jacobi method in massive MIMO. in *2014 IEEE International Conference on Communiction Problem-solving*. (2014)
71. Z. Wu, C. Zhang, Y. Xue, S. Xu, X. You, Efficient architecture for soft-output massive MIMO detection with Gauss-Seidel method. in 2016 IEEE International Symposium on Circuits and Systems (ISCAS). (2016)
72. R. Chataut, R. Akl, U. Dey, Least square regressor selection based detection for uplink 5G massive MIMO systems. in *2019 IEEE 20th Wireless and Microwave Technology Conference (WAMICON)*. (2019)
73. R. Chataut, R. Akl, Huber fitting based ADMM detection for uplink 5G massive MIMO systems. in *2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. (2018b)
74. R. Chataut, R. Akl, Efficient and low complex uplink detection for 5G massive MIMO systems. in *2018 IEEE 19th Wireless and Microwave Technology Conference (WAMICON)*. (2018a)
75. L. Li, H. Hou, W. Meng, Convolutional-neural-network-based detection algorithm for uplink multiuser massive MIMO systems. IEEE Access **8**, 64250–64265 (2020)

76. R. Chataut, R. Akl, Channel gain based user scheduling for 5G massive MIMO Systems. in *2019 IEEE 16th International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT and AI (HONET-ICT)*. (2019)
77. S. Kim, J. Chae, S. Hong, Machine learning detectors for MU-MIMO systems with one-bit ADCs. IEEE Access. **8**, 86608–86616 (2020)
78. S. Kim, S. Hong, A supervised-learning detector for multihop distributed reception systems. IEEE Trans. Veh. Technol. **68**, 1958–1962 (2019)
79. O. Simeone, A very brief introduction to machine learning with applications to communication systems. IEEE Trans. Cogn. Commun. Netw. **4**, 648–664 (2018)
80. C. Jiang, H. Zhang, Y. Ren, Z. Han, K. Chen, L. Hanzo, Machine learning paradigms for next-generation wireless networks. IEEE Wirel. Commun. **24**, 98–105 (2017)
81. A. Zappone, L. Sanguinetti, M. Debbah, User association and load balancing for massive MIMO through deep learning. in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*. (2018)
82. Y. Yang, F. Gao, X. Ma, S. Zhang, Deep learning-based channel estimation for doubly selective fading channels. IEEE Access **7**, 36579–36589 (2019)
83. S. Dorner, S. Cammerer, J. Hoydis, S. Brink, Deep learning based communication over the air. IEEE J. Select. Top. Signal Proc. **12**, 132–143 (2018)
84. T. Lin, B. Horne, P. Tino, C. Giles, Learning long-term dependencies in NARX recurrent neural networks 1996. IEEE Trans. Neural Netw. 7, 1329–1338
85. S. Hochreiter, J. Schmidhuber, Long short-term memory. Neural Comput. **9**, 1735–1780 (1997)
86. M. Alrabeiah, A. Alkhateeb, Deep learning for TDD and FDD massive MIMO: mapping channels in space and frequency
87. A. Waseem, A. Naveed, S. Ali, M. Arshad, H. Anis, I. Qureshi, Compressive sensing based channel estimation for massive MIMO communication systems. Wirel. Commun. Mob. Comput. **2019**, 1–15 (2019)

Millimeter-Wave Dual-Band (32/38 GHz) Microstrip Patch Antenna for 5G Communication

Jyoti Hatte, Shivleela Mudda, K. M. Gayathri, and Rupali B. Patil

Abstract In this paper proposed is a dual-band millimeter-wave antenna for the fifth-generation device applications operating at Ka band is proposed in this paper. The designed antenna uses a substrate layer made up of RT Rogers 5880 material having dimensions of $7.8546 \times 9.0356 \times 0.8\ mm^3$ with dielectric constant of 2.2 and loss tangent 0.0009 on which radiating patch is loaded with surface area of $3.054 \times 4.235\ mm^2$. The antenna resonates at 32 and 38 GHz, respectively. These two frequencies are allotted to 5G communication by ITU. The simulation results of this antenna are in terms of electrical parameters such as Return loss, VSWR, Radiation pattern, and Gain. Performance parameters are further enhanced using defects in ground plane. This antenna can cover both 32 GHz band (31–34 GHz) and 38 GHz band (36.5–40 GHz). The minimum value of return loss, VSWR are −23.6 dB, 1.14 at 32 GHz similarly, −21.5 dB, 1.18 at 38 GHz, respectively. A −10 dB impedance bandwidth obtained at 32 and 38 GHz are 2.8 and 2.85 GHz, respectively. The studied antenna attained the gain of 6.27 and 4.73 dB with almost unidirectional radiation pattern. The propounded antenna is analyzed using HFSS 15.0 software.

Keywords 5G · Millimeter · DGS (defected ground structure) · Dual band · Microstrip antenna · HFSS

J. Hatte (✉) · R. B. Patil
Department of Electronics and Telecommunication Engineering, G H Raisoni College of Engineering and Management, Wagholi, Pune, Maharashtra, India

R. B. Patil
e-mail: rupalib.patil@raisoni.net

S. Mudda · K. M. Gayathri
Department of Electronics and Communication Engineering, Dayananda Sagar College of Engineering and Management, Bangalore, Karnataka, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_17

1 Introduction

In recent years requirement of immense caliber universal wireless communication systems is increased than earlier. This demand is the sweeping force fronting the accelerated research and evolution in the area of intelligence communication technologies all over the world that advances the research area toward the growth of upcoming generation of wireless communication. Fifth-generation (5G) techniques for broadcast systems turn interesting and raise enormous request from the final users. 5G operation facilitated applications will be decided based on the selection of spectrum, propagation properties, antenna technology, transceiver assimilation, and digital signal processing of broadband technology [1, 2]. Abundant research is carried around the world toward the growth and normalization of 5G systems. The expected frequency bands for 5G are sub-6-GHz band and a new millimeter-wave band which promises high data rates with reduced latency. 5G communication claims for larger gain aerial to overcome path loss at mm waves. The ITU issued the powered millimeter-wave band frequency 24–84 GHz in the 2015 conference of a global radio communication, in which main band are 27.5–29.5 GHz, 33.4–36 GHz, 37–40.5 GHz, 42–45 GHz, 47–50.2 GHz, and 59.3–71 GHz [3]. Because at these frequencies the air fading is negligible and hence this leads to maximum data rate with reduced latency [4]. Many scholars are trying to design a compact and broad bandwidth antenna for the mentioned frequencies. The supreme way is to design an antenna resonating at two bands rather than using two separate antennas to enhance bandwidth, which permits dual bands to be occupied with single antenna, and also use of one structure will lessen the place required and price figure of the system. In such a case, microstrip patch antenna (MPA) operating at dual frequencies is widely used for efficient communication, due to its simplicity of design, compactness, and monolithic integration with other circuit components [5, 6]. Microstrip antennas had few drawbacks, like single operating frequency, lower impedance bandwidth, low gain, larger in size, and polarization problems. Many techniques have been specified for improving the performance of traditional microstrip antennas, such as using stacking and using different feeding techniques [7]. Defected Ground Structure (DGS) has achieved popularity among all the methods reported for improving the performance due to its simple framework design. These structures are either periodic slots or aperiodic slots engraved in the ground layer. The current paths are modified by these engraved slots which results in variation of the reactance of feed port [8, 9]. Dual-band MIMO antenna using inverted I-shaped slots in patch with DGS of slot in the partial ground plane is presented which resonates at 28 and 38 GHz [10]. A novel compact dual frequency (38/60 GHz) microstrip patch antenna proposed finds application in 5G mobile handset [11]. Parametric variation of slot length and width gives improvement in resonant frequency and improved impedance match [12].

This paper describes design of defective ground structure based dual-band microstrip patch antenna resonating at 32 and 38 GHz, respectively. We observed good values for performance parameters of return loss¸ VSWR, gain, and radiation pattern that were achieved. These frequencies are under consideration for future fifth-generation (5G) wireless technologies.

2 Antenna Design

Antenna functionality can be affected by height, dielectric constant value, and tangent loss of the substrate. Hence the substrate for design was conscientiously chosen as Rogers RT-duroid −5880 with $\varepsilon_r = 2.2$, $\tan \delta = 0.0009$, and height 0.8 mm. Advantages such as negligible dielectric loss and lower dispersion makes RT-duroid as the most suitable material for millimeter-wave applications [13].

In the first phase of design, copper patch of rectangular shape is placed on top of the substrate layer. Microstrip feeding technique is used as it is simple to design and fabricate.

On the patch one U-shaped slot and another nearly U-shaped slot are etched refer Fig. 1. The dimensions of the substrate are $7.8546 \times 9.0356 \times 0.8$ mm^3, and that of the feed to the patch are of 0.277×3.2 mm. The final dimensions of planned slot are analyzed by optimization and by maintaining the same antenna parameters of the proposed system. The simulation results of the first design are shown in Fig. 3a–e which yielded the reflection loss S_{11} of −14.56 dB, VSWR of 1.46, and gain of 6.38 dB.

In the second phase of the design, dual-band characteristic was obtained by incising three vertical slots in the patch and etching rectangular ring shaped DGS in the ground plane, refer Fig. 2a. The dimensions of the patch were not altered but the horizontal and vertical width of the new three slots were optimized to get the second resonance at 38 GHz. The DGS structure showed the improvement in the S_{11}, VSWR, and impedance characteristics. The antenna system model with dimensions is as shown in Fig. 2b. And the optimized dimensions of the proposed dual-band antenna system are given in Table 1.

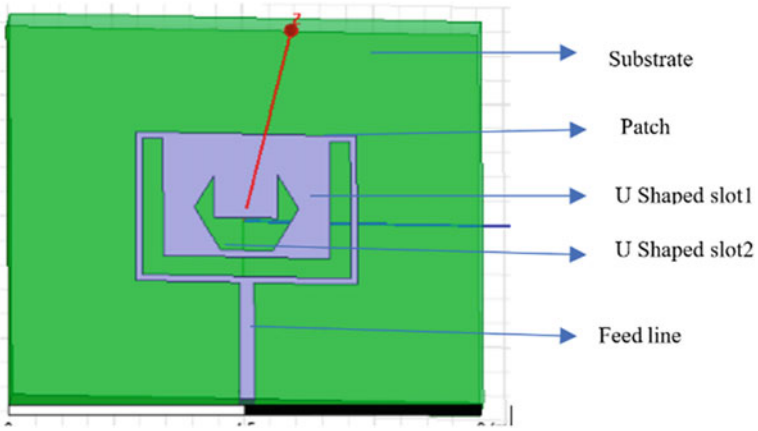
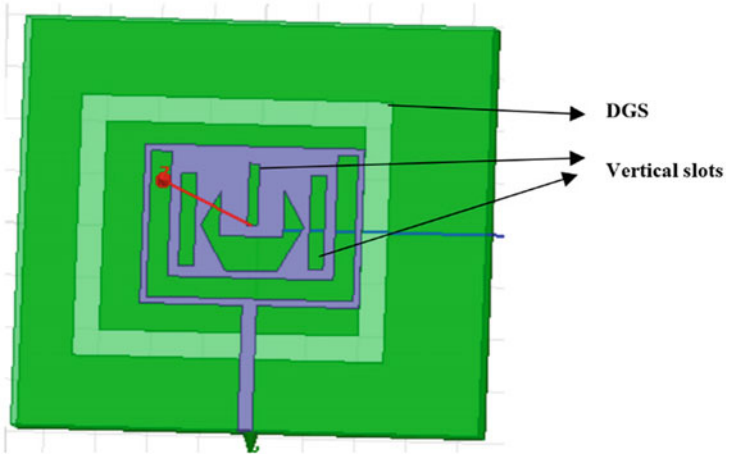
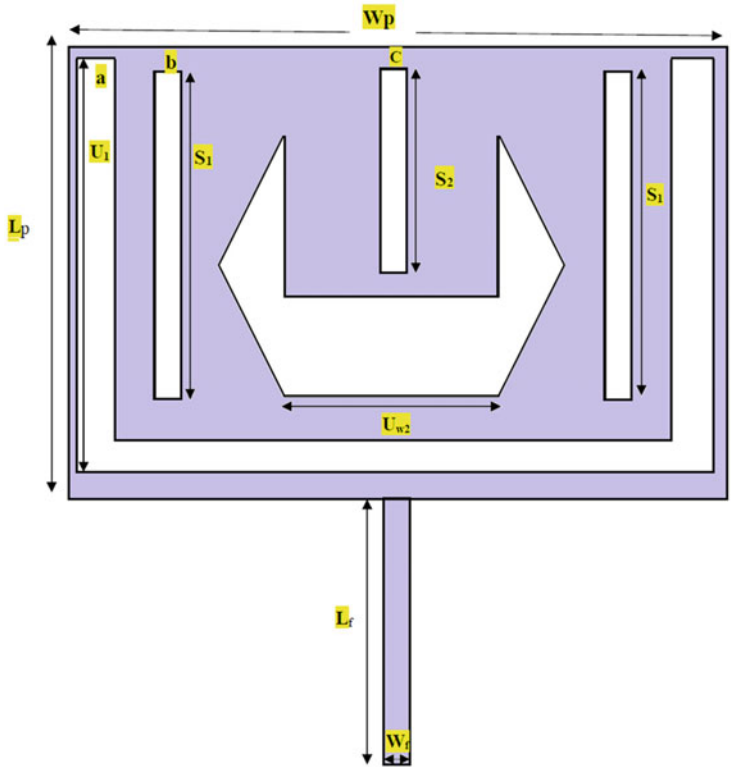


Fig. 1 Proposed single-band antenna design resonating at 32 GHz



(a)

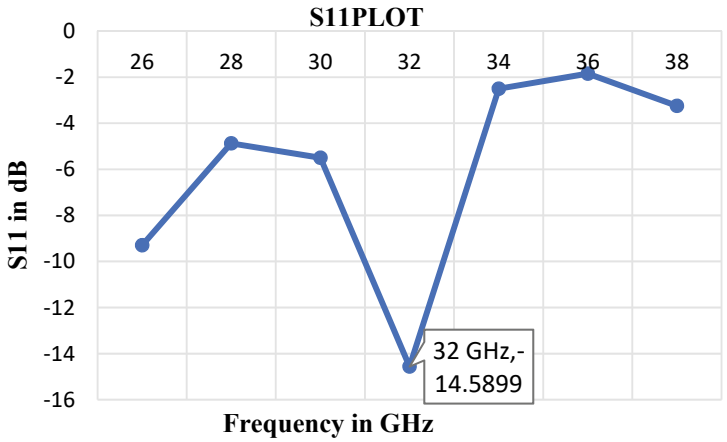


(b)

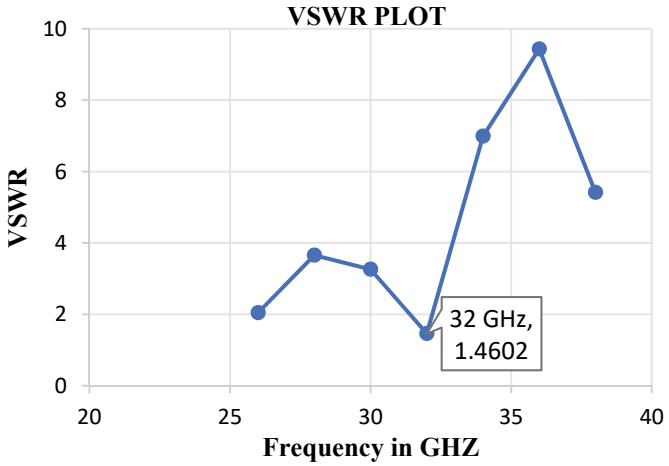
Fig. 2 **a** proposed geometry of 32/38 GHz antenna **b** Proposed dual-band antenna with dimensions

Table 1 Designed values of proposed antenna

Parameter	Dimensions	Parameter	Dimensions	Parameter	Dimensions
W_p	4.235 mm	L_f	3.2 mm	c	0.2 mm
L_p	3.054 mm	W_f	0.277 mm	UW_2	1 mm
U_1	2.8 mm	a	0.4 mm	S_1	1.8 mm
UW_1	3.9 mm	b	0.3 mm	S_2	1.2 mm

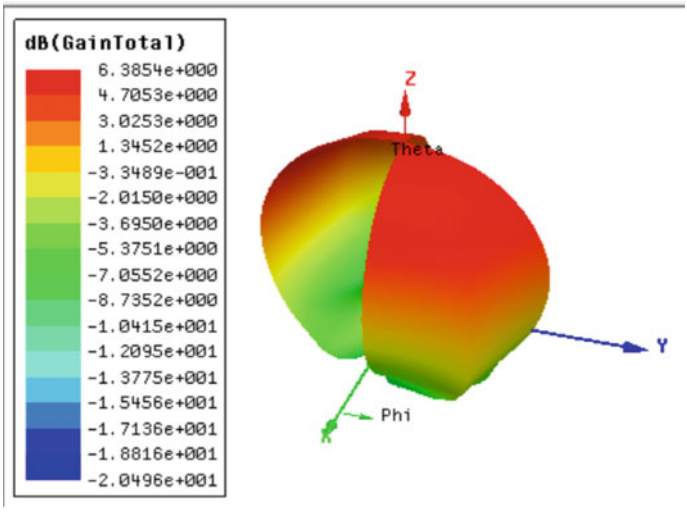


(a)

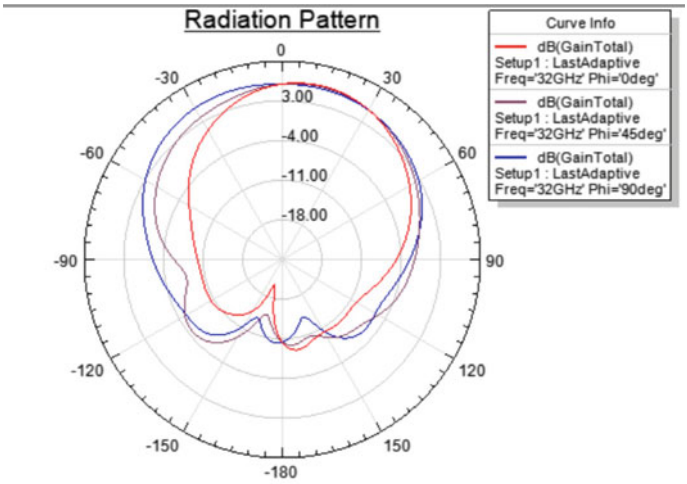


(b)

Fig. 3 **a** S_{11} parameter at 32 GHz. **b** VSWR parameter at 32 GHz. **c** Gain at 32 GHz. **d** Radiation Pattern at 32 GHz. **e** Smith chart at 32 GHz



(c)



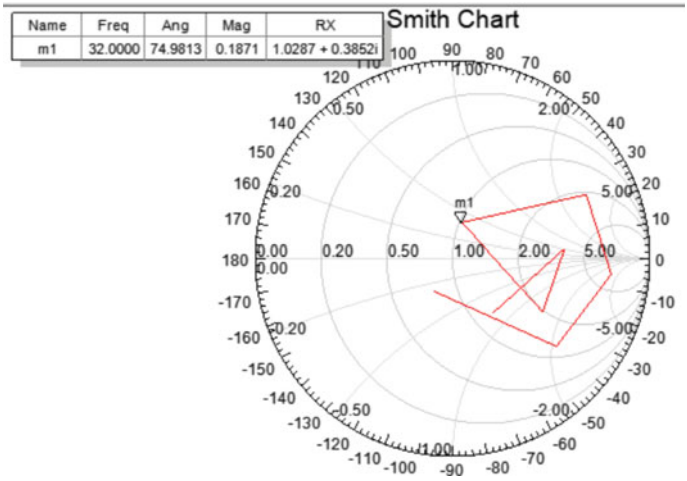
(d)

Fig. 3 (continued)

2.1 Design Method

Microstrip patch antenna can be designed by selecting the required operating frequency and appropriate substrate [14]. The dimensions for the design can be calculated using transmission line model equations.

Rectangular patch width is obtained by using the following equation



(e)

Fig. 3 (continued)

$$W = (C/2f)\sqrt{\left(\frac{2}{(\varepsilon_r + 1)}\right)} \tag{1}$$

where f is the operating frequency

and

$$\varepsilon_{reff} = \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{2}\left[1 + 12\frac{h}{w}\right]^{-\frac{1}{2}}, \frac{w}{h} > 1 \tag{2}$$

Rectangular patch length is calculated as

$$\frac{\Delta L}{h} = 0.412\left(\frac{\varepsilon_{reff} + 1}{\varepsilon_{reff} - 0.258}\right)\left(\frac{\frac{w}{h} + 0.264}{\frac{w}{h} + 0.8}\right) \tag{3}$$

$$L = \frac{c}{2f\sqrt{\varepsilon_{reff}}} - 2\Delta L \tag{4}$$

Length and width of ground plane are calculated as

$$L_g = L + 6h \tag{5}$$

$$W_g = W + 6h \tag{6}$$

3 Simulation Results

In the first step of design antenna resonates at 32 GHz. The simulation result gave $S_{11}$of −14.556, VSWR of 1.46, and gain of 6.38 dB, refer Fig. 3a-c. The plot of radiation pattern and smith chart for impedance measurement are as in Fig. 3d, e, respectively.

Second step of design constitutes dual-band microstrip patch antenna resonating at 32 and 38 GHz. For dual characteristics of antenna, slots are engraved in the patch and optimized. Multiple slots in patch and DGS in ground plane create multiple current paths which make the antenna to resonate at more than one frequency. Simulation result showed the S_{11} parameter of −23.54 and −21.54 dB at 32 and 38 GHz, respectively, as shown in Fig. 4.

For efficient antenna its VSWR should be below 2 leading to S_{11} being less than −10 dB. The VSWR obtained is 1.1426 and 1.1829 at 32 and 38 GHz, respectively, which is shown in Fig. 5. The gain obtained is 6.2789 and 4.7397 dB at 32 and 38 GHz, respectively, refer Fig. 6a,b. Offered gain made the design to fulfil the requirements of 5G.

The radiation pattern of the proposed antenna system for 32 and 38 GHz are as shown in Fig. 7a, b. Radiation patterns are unidirectional for both resonating frequencies. It is one of the prime advantages of the proposed design because stable, unidirectional, or omnidirectional radiation pattern is the key requirement of 5G communication.

The impedance values are 1.0554–0.1253j and 1.1477–0.1031j at 32 and 38 GHz, respectively, as shown in the smith chart of Fig. 8. Impedance values shows good match between patch and feedline.

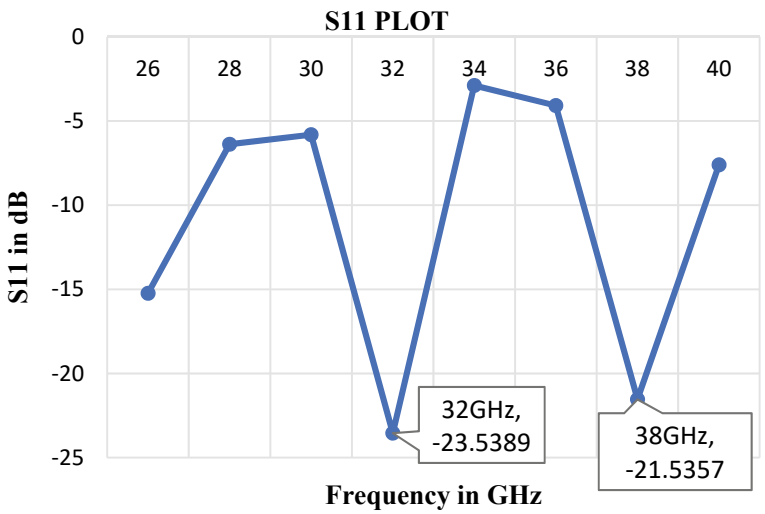


Fig. 4 Reflection coefficient of simulated dual-band antenna

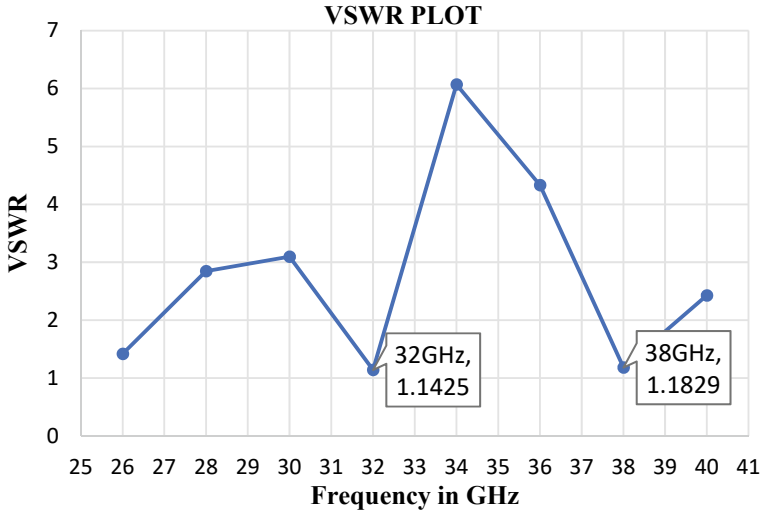


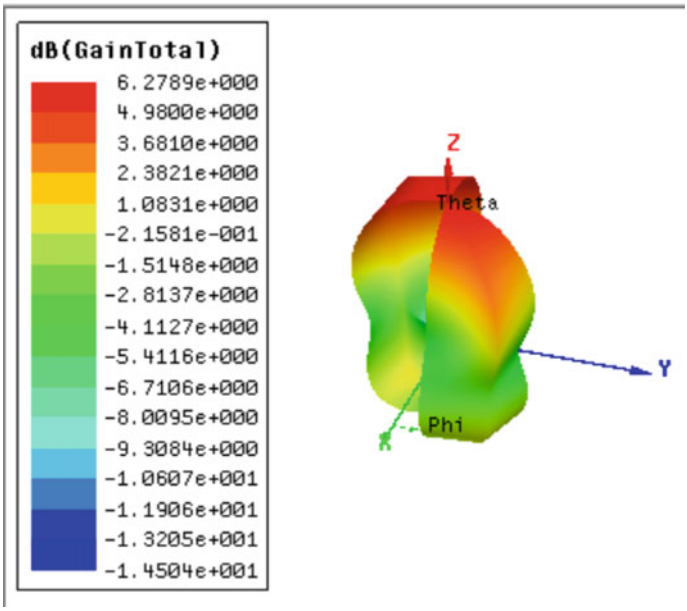
Fig. 5 VSWR of simulated dual-band antenna

4 Conclusion

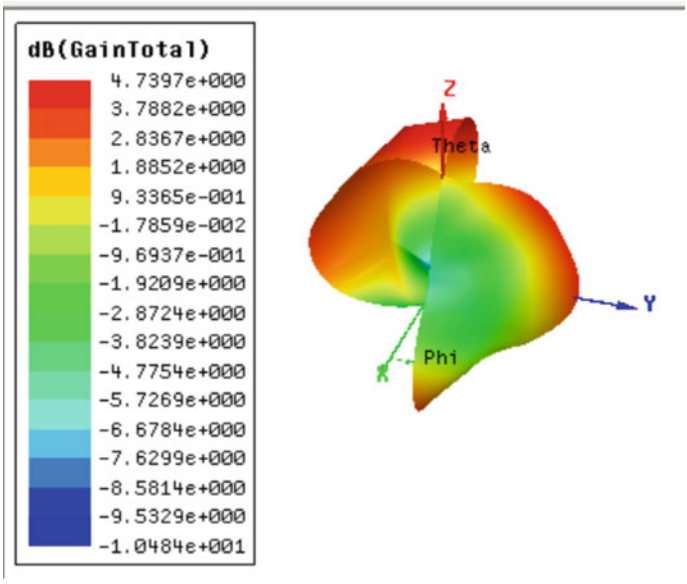
A rectangular microstrip antenna showing resonance operation at dual band has been designed and presented in this paper. The antenna supports a good dual frequency operation in the range from 32 to 38 GHz. The radiation pattern is directional with high gain. The designed simple dual-band antenna is good option for operating in millimeter frequency band. Introduction of DGS has shown the improvement in the return loss to large extent and also there is improvement in VSWR characteristics with slight reduction in the gain. The proposed dual-band antenna element which has the return loss of −23.54 dB at 32 GHz and −21.54 dB at 38 GHz with compact size makes it a developing contender for MIMO applications in mm-wave band of 5G communications.

5 Future Scope

The proposed design can be extended for further improvement in the bandwidth, gain, and radiation parameter, so that it can be implemented for MIMO antenna system in 5G applications. It is also possible to make antenna radiation pattern omnidirectional by introducing partial ground plane technique so that it can be used for portable devices such as mobile handsets, laptops, tabs, and other direction showing hand held devices.

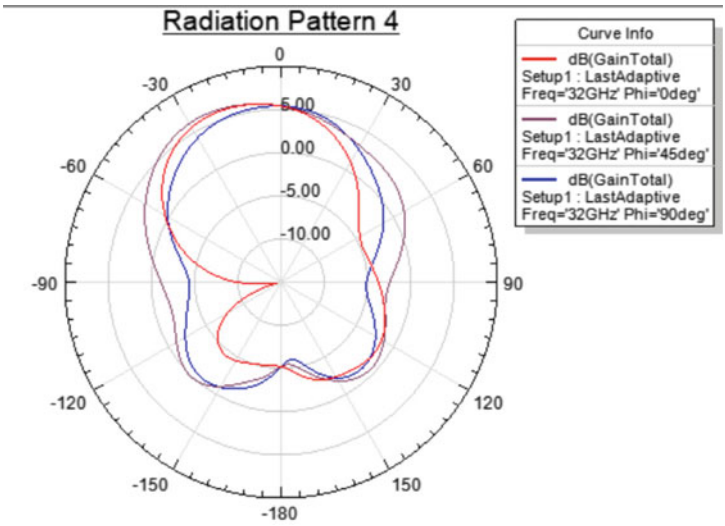


(a)

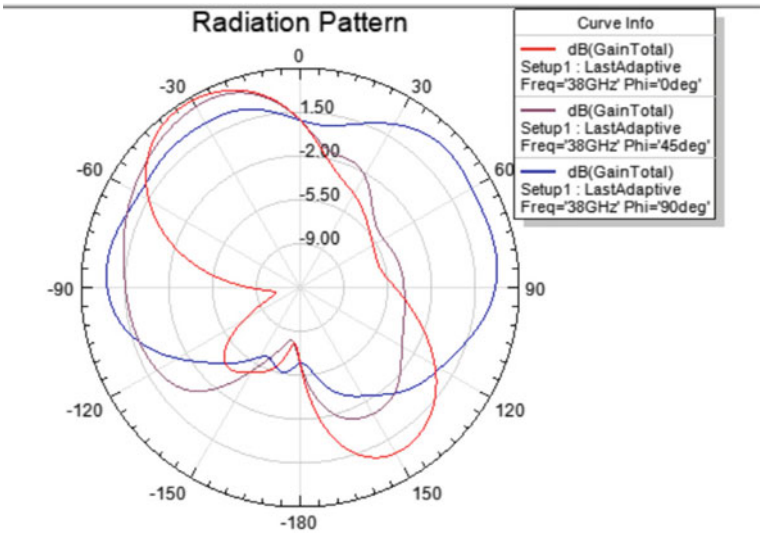


(b)

Fig. 6 **a** 3D Gain of simulated antenna at 32 GHz. **b** 3D Gain of simulated antenna at 38 GHz



(a)



(b)

Fig. 7 **a** Simulated radiation pattern of antenna at 32 GHz. **b** Simulated radiation pattern of antenna at 38 GHz

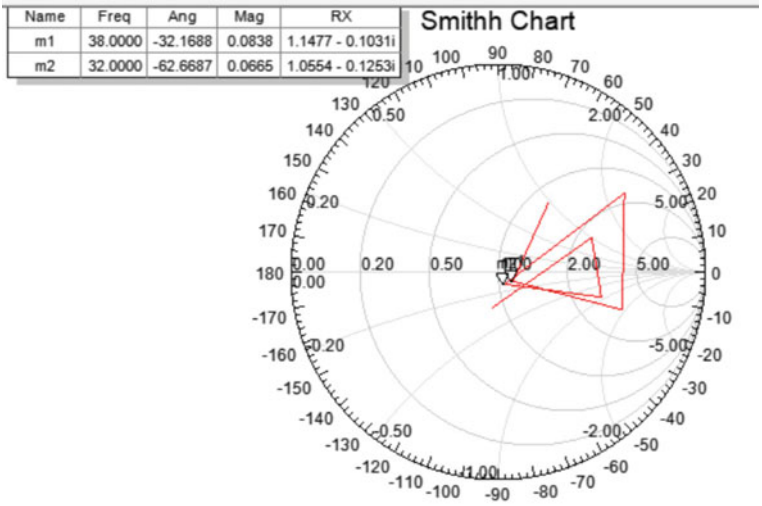


Fig. 8 Smith chart of antenna at 32 GHz and at 38 GHz

References

1. M. Masoudi, Green mobile networks for 5G and beyond. IEEE Access **7**, 107270–107299 (2019)
2. R. Vanitha, S. Talwar, *Towards 5G: applications, requirements and candidate technologies* (Wiley, Hoboken, NJ, USA, 2017)
3. C. Seker, M.T. Güneser, A single band antenna design for future millimetre wave wireless communication at 38 GHz. Eur. J. Eng. Form. Sci. **2**, 35–39 (2018)
4. M.L. Hakim, M.J. Uddin, M.J. Hoque, 28/38 GHz dual-band microstrip patch antenna with DGS and stub-slot configurations and its 2x2 MIMO antenna design for 5G wireless communication. in *2020 IEEE Region 10 Symposium (TENSYMP).* (Dhaka, Bangladesh, 2020)
5. F.A.L.B Konkyana, B.A. Sudhakar, A review on micro strip antennas with defected ground structure techniques for ultra-wideband applications. in *2019 International Conference on Communication and Signal Processing (ICCSP), Chennai, India.* (2019) pp. 0930–0934. https://doi.org/10.1109/ICCSP.2019.8697941
6. B. Biglarbegian, M. Fakharzadeh, D. Busuioc, M.R. Nezhad-Ahmadi, S. Safavi-Naeini, Optimized microstrip antenna arrays for emerging millimeter-wave wireless applications. IEEE Trans. Antennas Propag. **59**, 1742–1747 (2011)
7. F. Zavosh, J. Aberle, Improving the performance of microstrip-patch antennas. IEEE Antennas Propag Mag **38**, 7–12 (1996) https://doi.org/10.1109/74.537361
8. M.K. Khandelwal, B.K. Kanaujia, S. Kumar, Defected ground structure: fundamentals, analysis, and applications in modern wireless trends. Int. J. Antennas Propag. (2017). https://doi.org/10.1155/2017/2018527
9. S. Mudda, K.M. Gayathri, M. Mudda, Compact high gain microstrip patch millimeter wave multi-band antenna for future generation portable devices communication. Int. Conf. Emerg. Smart Comp. Inform. (ESCI) **2021**, 471–476 (2021). https://doi.org/10.1109/ESCI50559.2021.9396776
10. H.M. Marzouk, M.I. Ahmed, A.H.A. Shaalan, Novel dual-band 28/38 GHz MIMO antennas for 5G mobile applications. Prog. Electromag. Res. C **93**, 103–117
11. M.H. Sharaf, A.I. Zaki, R.K. Hamad, M.M.M. Omar, A Novel dual-band (38/60 GHz) patch antenna for 5G mobile handsets. Sensors **20**, 2541 (2020). https://doi.org/10.3390/s20092541

12. J.S. Hatte, R.B. Patil (2020) Design methods of MIMO antenna for 5G new radio applications in mobile terminals-a review. Int. J. Emerg. Technol. **11**(3), 738–745
13. P. Mane, S.A. Patil, P.C. Dhanawade, Comparative study of micro strip antenna for different substrate material at different frequencies. Int. J. Emerg. Eng. Res. Technol. **2**(9), 18–23
14. H. Werfelli, K. Tayari, M. Chaoui, M. Lahiani, H. Ghariani (2016) Design of rectangular microstrip patch antenna. 798–803. https://doi.org/10.1109/ATSIP.2016.7523197

Design and Analysis of Single Band and Wideband Wineglass-Shaped Patch Antenna for WLAN and Satellite Applications

Narbada Prasad Gupta, Parulpreet Singh, Sanjay Kumar Sahu, and Shelej Khera

Abstract The present work focuses around plan and investigation of a wineglass-formed planar patch antenna (PPA) with single band and wideband capacities. The proposed antenna works at two diverse frequency bands, for example, one band is centered at 3.62 GHz and another at 7.45 GHz. It has a single band frequency operation at the mid frequency of 3.62 GHz and wideband operation from 6.4 to 9.3 GHz. Subsequently, it is appropriate for both single band and wideband applications. The antenna proposed, here, has been planned utilizing FR4 promptly accessible substrate with dielectric consistent of 4.5. Height of the substrate is 0.8 mm and size of substrate is 32×26 mm^2. The antenna has been mimicked utilizing CST MWS RF Simulator. The proposed structure is upgraded to have better return loss, VSWR, gain, and radiation characteristics for the planned frequencies. The proposed antenna is discovered to be appropriate for WiMax, Mobile WiMax, RADAR, and Satellite applications.

Keywords Patch antenna · Wineglass-shaped antenna · Single band and wideband antenna

1 Introduction

In modern communication system, it has been realized in fact attracted much attention that there is a need of antennas for wideband or multiband operations. Ultra-Wide Band (UWB) has been flung in the frequency range from 3.1 to 10.6 GHz, which

N. P. Gupta (✉) · P. Singh · S. K. Sahu · S. Khera
School of Electronics and Electrical Engineering, Lovely Professional University, Phagwara, Punjab, India
e-mail: narbada.24806@lpu.co.in

P. Singh
e-mail: parulpreet.23367@lpu.co.in

S. K. Sahu
e-mail: sanjay.23393@lpu.co.in

S. Khera
e-mail: shelej.22390@lpu.co.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_18

has fascinated attention of huge number of researchers. It has got various advantages of being cost effective, resistant to unadorned multipath and jamming, etc. [1–3]. A wide variety of work on design and optimization of UWB antennas have been done by various researchers as complied in [4]. Most of the researches were focused on specific frequencies [5–7].

A UWB antenna may be a good choice to be used for wide band operations [1, 8]. UWB communication suggests fundamentally different approach to wireless communication if it is related with a traditional narrowband system. It is centered on ultra-wide spectrum, short pulse transmission with relatively low energy, tolerable interference with other users, and slight power spectral density [9]. Design of antenna for the transmission of short pulses is a key task. As far as system design is concerned, response of antenna should be such that it covers complete working bandwidth. Also, the antenna should be non-responsive to signals external to the stated band [10]. Performance evaluation characteristics like radiation pattern, return loss, and gain is not sufficient exclusively. But still it provides a significant idea about the design. For short-range, high-speed, and low-power UWB communication, it is also important that waveform distortion during the transient transmission, propagation, and reception must be evaluated. This work is beyond the scope of the work presented here. Variety of antennas have been freshly suggested for UWB applications as given in [9, 10]. Authors in [11] presented a wineglass-shaped patch antenna for wireless communication with woodpile EGB structure. In this paper, EBG structure has been used to enhance the gain characteristics of the structure. Author in [12] presented reconfigurable antenna with selective and wideband capabilities. Various authors have proposed multiband operable structure [14, 15], which further improves the applicability of the antenna for variety of applications. But all of these proposed antennas have their pros and cons like complicated structure, large antenna size, typical substrate material, etc. The purpose of the present research is to propose a design which is novel, miniaturized, and has dual-band capabilities. The dual-band capability has been focused while designing the antenna for single frequency as well as wideband frequency at the same time.

In this paper, a novel but very simple design of single and Wide Band antenna having a shape of wine glass is presented on FR4 substrate which is easily available in the open market. The proposed antenna is appropriate to be used for WiMax and satellite applications. In the next section, design of antenna along with all its simulated results has been presented. Further, design results are compiled and discussed. In the last section, results and related conclusions have been discussed.

2 Design of Wineglass-Shaped Single and Wideband Patch Antenna

The wineglass-shaped antenna is a unique structure proposed in this paper. It has lower width in the bottom and larger width in the upper portion, which depicts a wineglass. The proposed shape of wineglass has smooth curvature along the length; therefore, it becomes like a travelling wave structure when implemented as an antenna. Larger width at the upper portion is a decisive factor for the wideband characteristics. In fact the quality factor of the antenna depends on the antenna dimensions like length and width [11]. The author has proposed a wineglass-shaped microstrip patch antenna structure of patch size 41.56 by 26.53 mm^2 in [11]. In this paper, the size of the patch is quite smaller, i.e., 13.5 by 11.99 mm^2. Also, the proposed structure is fed with line feed, whereas author in [11] used CPW feeding. Although the CPW offers wideband impedance matching, but in the proposed structure, it is of less significance due to single and wideband capability.

Initially, a simple patch antenna with full ground in the back side of the substrate has been designed using standard design equations given in [13]. The simple patch is designed with a substrate of given thickness (h = 0.8 mm). Size of the substrate is 32 × 26 mm.

For a rectangular MSA, its width W and the length L are [13]

$$W = \frac{c}{2f_r}\left(\frac{\varepsilon_r + 1}{2}\right)^{-0.5} \tag{1}$$

$$l = \frac{c}{2f_r\sqrt{\varepsilon_{eff}}} - 2\Delta l \tag{2}$$

$$\varepsilon_r = \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{2}\left(1 + \frac{12h}{W}\right)^{-0.5} \tag{3}$$

$$\Delta l = 0.412h\left(\frac{0.262 + \left(\frac{W}{h}\right)}{0.813 + \left(\frac{W}{h}\right)}\right)\left(\frac{\varepsilon_{eff} + 0.3}{\varepsilon_{eff} - 0.258}\right) \tag{4}$$

where

c = Speed of light = 3 X 10^{11} mm/s.

f_r = The operational frequency in GHz.

ε_r = Dielectric property of the substrate.

ε_{eff} = Effective dielectric constant.

Δl = The line allowance in mm.

h = Depth of the substrate in mm.

3 50 Ω Feed Line

For a proposed characteristic impedance Z_o and dielectric constant ε_r, the ratio of W and d could be found as [4]

$$\frac{W}{d} = \left[\frac{8l^A}{l^{2A} - 2}\right] \tag{5}$$

$$\frac{W}{d} = \left(\frac{2}{\prod}\right)\left[B - 1 - \ln(2B - 1) + \left\{\left(\frac{\varepsilon_r - 1}{2\varepsilon_r}\right)\left(\ln(B - 1) + 0.39 - \left(\frac{0.61}{\varepsilon_r}\right)\right)\right\}\right] \tag{6}$$

where l—line length and

$$A = \left(\frac{Z_0}{60}\right)\left[\left(\frac{\sqrt{\varepsilon_r + 1}}{2}\right) + \left(\frac{\varepsilon_r - 1}{\varepsilon_r + 1}\right)\left\{0.23 + \left(\frac{0.11}{\varepsilon_r}\right)\right\}\right] \tag{7}$$

$$B = \left[\frac{377\prod}{2Z_0\sqrt{\varepsilon_r}}\right] \tag{8}$$

$$\varepsilon_{re} = \left[\frac{\varepsilon_r + 1}{2}\right] + \left[\left(\frac{\varepsilon_r - 1}{2}\right)\left(\frac{1}{\sqrt{1 + \frac{12H}{w}}}\right)\right] \tag{9}$$

$$\lambda_g = \frac{\lambda_0}{\sqrt{\varepsilon_{re}}} \tag{10}$$

$$L = \frac{\lambda_g}{4} \tag{11}$$

After designing the simple patch antenna, its bandwidth has been enhanced using techniques provided in [12]. Some of the techniques which are found to be suitable are

a. Using partial ground
b. Truncating the edges
c. Modifying the substrate height
d. Using parasitic elements.

Out of these techniques, partial ground method has been used to improve the bandwidth of antenna at two different center frequencies.

Figure 1 presents two different views (front and back) of the designed antenna. The antenna has been realized using FR-4 substrate with relative dielectric constant of 4.5 and loss tangent 0.0005.

Figure 2 shows parametric study of ground length. The length of the ground has been made varying between 9.05 and 10.1 mm. It is observed from the results that increment in the ground length leads to omission of wideband characteristics of the proposed antenna. So, to design the prototype the optimized value of ground length

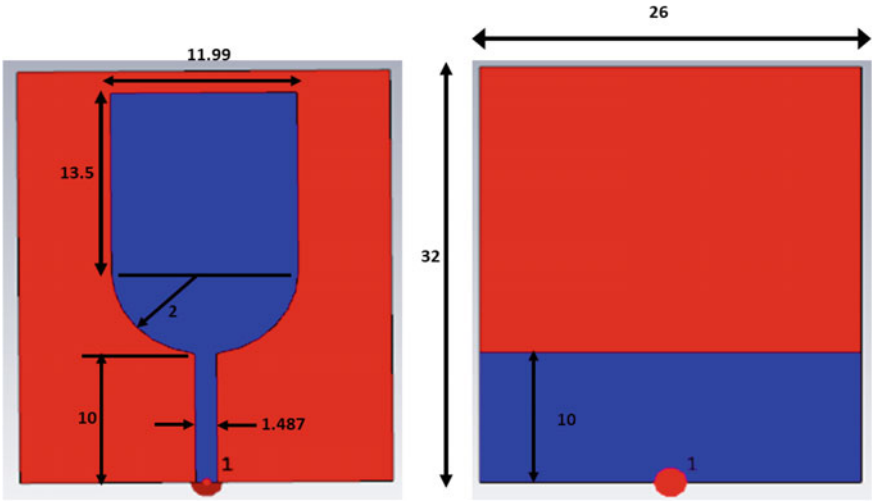


Fig. 1 Geometry of wineglass-shaped patch antenna

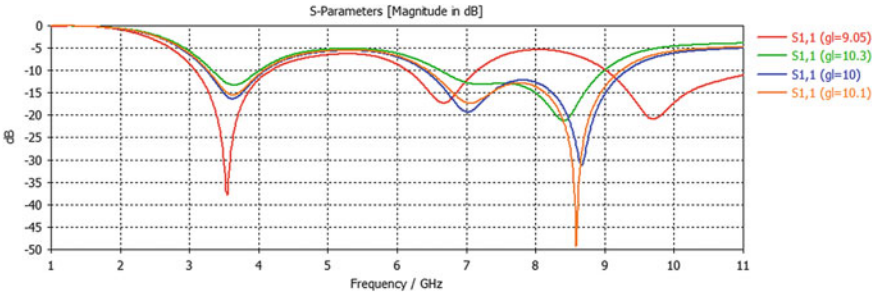


Fig. 2 Parametric study for the proposed antenna

has been taken as 10 mm. It is considered such that both the single band and wideband characteristics meet the required characteristics.

Figure 3 shows the optimized S_{11} parameter of the proposed wineglass-shaped patch antenna. It is shown here that the antenna is resonating at 3.62, 6.99, and 8.65 GHz. The corresponding S_{11} is also shown in the figure. −10 dB bandwidth of the antenna resonating at 3.62 GHz is between 3.24 and 4.09 GHz, which is 850 MHz. Similarly, the antenna shows wideband characteristics between 6.33 and 9.33 GHz. Total bandwidth is 3 GHz.

Figure 4 shows the optimized VSWR for the proposed structure. It has been shown that the VSWR is below 1.4 for the entire band of interest. VSWR is 1.35 at 3.6 GHz, 1.24 at 6.99 GHz, and 1.05 at 8.66 GHz frequencies.

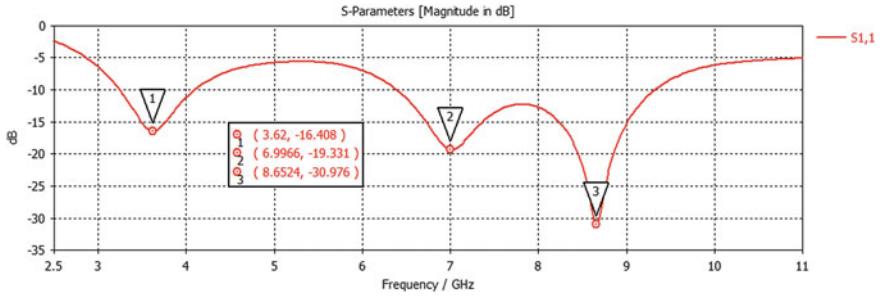


Fig. 3 Optimized S_{11} of wineglass-shaped patch antenna

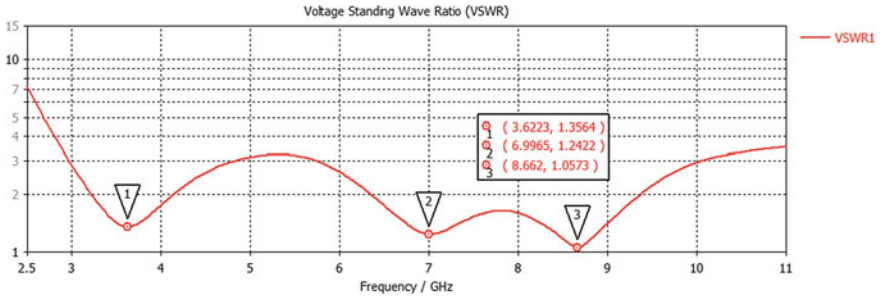


Fig. 4 Optimized VSWR of wineglass-shaped patch antenna

Figure 5 shows radiation efficiency of the proposed wineglass-shaped patch antenna for diverse ground lengths. It has been shown that radiation efficacy is more than 95% for entire frequency band of interest.

Gain is of utmost importance in the design of wideband antenna. Figure 6 represents the gain of the proposed antenna. It has been shown here that gain is ranging between 2 and 5 dB.

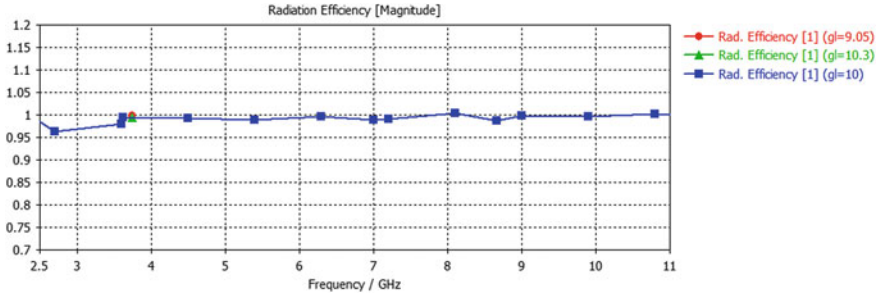


Fig. 5 Radiation efficiency of wineglass-shaped patch antenna

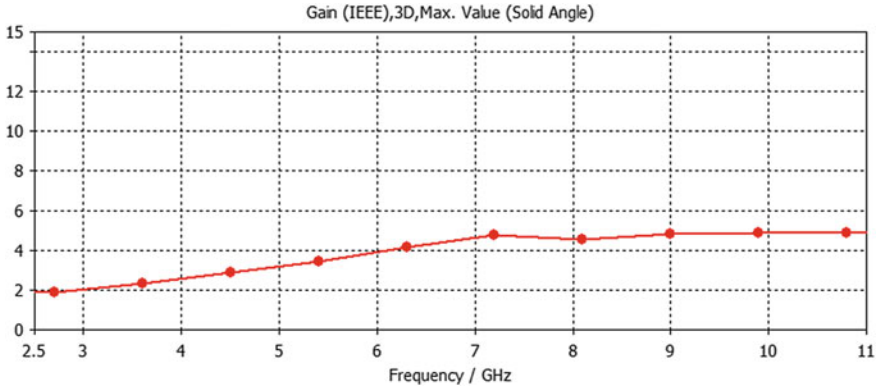


Fig. 6 Gain of wineglass-shaped patch antenna

Figures 7 and 8 show 3-D radiation pattern and E-plane and H-plane polar plot of the proposed wineglass0shaped patch antenna at 3.62 GHz. It has been shown here that the gain of the antenna is 2.35 dB.

Figures 9 and 10 show 3-D radiation pattern and E-plane and H-plane polar plot of proposed wineglass-shaped patch antenna at 7 GHz. It has been shown here that the gain of the antenna at this resonance frequency is 4.69 dB.

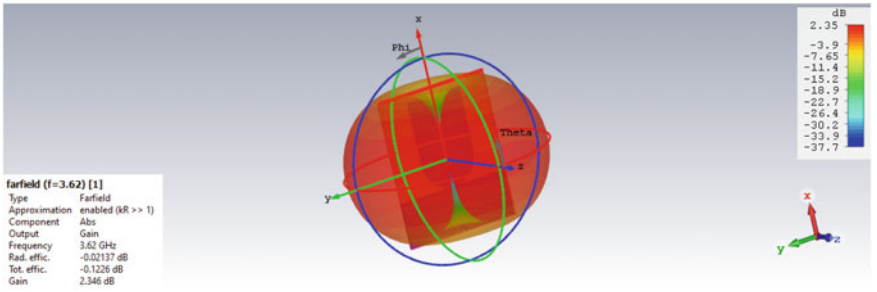


Fig. 7 3-D radiation pattern of wineglass-shaped patch antenna at 3.62 GHz

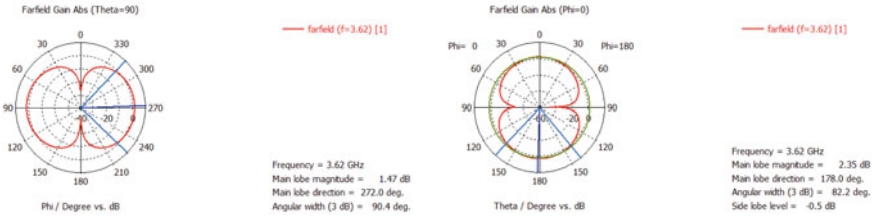


Fig. 8 E-plane and H-plane polar plot of wineglass-shaped patch antenna at 3.62 GHz

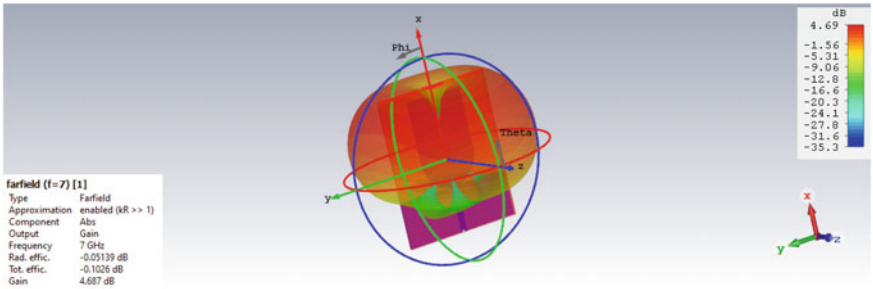


Fig. 9 3-D radiation pattern of wineglass-shaped patch antenna at 7 GHz

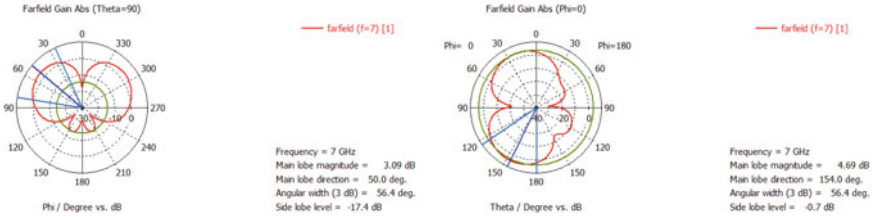


Fig. 10 E-plane and H-plane polar plot of wineglass-shaped patch antenna at 7 GHz

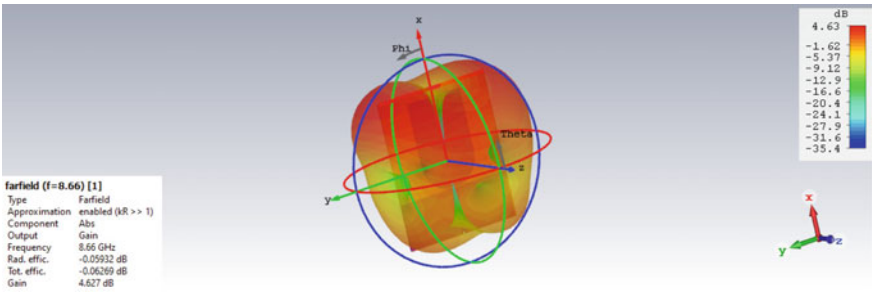


Fig. 11 3-D radiation pattern of wineglass shaped patch antenna at 8.66 GHz

Figures 11 and 12 show 3-D radiation pattern and E-plane and H-plane polar plot of the proposed wineglass-shaped patch antenna at 8.66 GHz. It has been shown here that the gain of the antenna is 4.63 dB at the resonance.

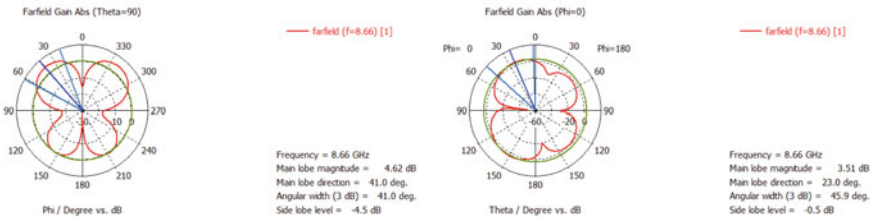


Fig. 12 E-plane and H-plane polar plot of wineglass-shaped patch antenna at 8.66 GHz

4 Conclusion and Discussion

In this paper, a novel miniaturized wineglass-shaped planner patch antenna has been introduced. The proposed antenna has been planned utilizing FR4 substrate having dielectric constant of 4.4 and thickness 0.8 mm, which is effectively accessible and famous among the researchers as well. The size of antenna substrate is $32 \times 26\ \text{mm}^2$. The proposed antenna has been shown to work at two diverse frequency bands, for example, one band is focused on 3.62 GHz and another at 7.45 GHz. It is shown that the antenna is centered at 3.62, 6.99, and 8.65 GHz. The corresponding S_{11}, transmission efficiencies, gain, and radiation pattern have been introduced. The—10 dB bandwidth of the antenna reverberating at 3.62 GHz is from 3.24 to 4.09 GHz, which is 850 MHz. Essentially, the proposed antenna shows wideband attributes from 6.33 to 9.33 GHz. Complete bandwidth is 3 GHz. Gain of the antenna is varying between 2.4 and 5 dB. The transmission effectiveness is more that 95% for the whole band of interest. Subsequently, it is appropriate for both single band and wideband applications. The proposed antenna is discovered to be appropriate for WiMax, Mobile WiMax, RADAR, and Satellite applications.

References

1. First report and order, revision of part 15 of the commission's rules regarding ultra-wideband transmission systems FCC. FCC02-48 (2002)
2. M. Oppermann et al., *UWB Theory and Applications*, vol. 1 (New York, Wiley, 2004), pp. 3–4
3. M. Ghavami, L. Michael, R. Kohno, *Ultra-Wideband Signals and Systems in Communication Engineering* (John Wiley & Sons, 2004)
4. P.S. Hall, Y. Hao, Antennas and Propagation for Body Centric Communications. In *European Conference on Antennas and Propagation (EuCAP)* (2006)
5. A.C. Gareth, G.S. William, Antennas for over-body-surface communication at 2.45 GHz. IEEE Trans. Ant. Prop. **57**(4) (2009)
6. W.S. Yeoh, K.L. Wong, W.S.T. Rowe, Wideband miniaturized half bowtie printed dipole antenna with integrated balun for wireless applications. IEEE Trans. Ant. Prop. **59**(1), 339–342 (2011)
7. N.P. Gupta, R. Maheshwari, M. Kumar, Advancement in ultra-wideband antennas for wearable applications. Int. J. Sci. Eng. Res. **4**, 341–348 (2013)

8. A. Mukherjee, H.M. Kwon, Compact multi-user wideband MIMO system using multiple-mode microstrip antennas. in *Proceedings of Vehicular Technology Conference Spring* (2007), pp. 584–588
9. N.P. Gupta, M. Kumar, Radiation performance improvement in wearable UWB antenna through slot insertion technique. in *IEEE CSNT* (Gwalior, India, 2015), pp. 83–87
10. R. Mallahzadeh, S. Es'haghi, A. Alipour, Design of an E shaped MIMO antenna using IWO algorithm for wireless applications at 5.8GHz. Prog. Electromagn. Res. **90**, 187–203 (2009)
11. R.C. Mahajan, V. Vyas, Wine glass shaped microstrip antenna with woodpile structure for wireless applications. Majlesi J. Electr. Eng. **13**(1), 37–45 (2019)
12. N.P. Gupta, M. Kumar Development of a reconfigurable and miniaturized CPW antenna for selective and wideband communication. Wirel. Pers. Commun. **95**(3), 2599–2608 (2017)
13. C.A. Balanis, *Antenna Theory and Analysis*, 2nd edn. (Wiley, New York, 1997)
14. D. Abdul Rahim, P.K. Malik, V.S. Ponnapalli, Sankar Ponnapalli, Design and analysis of multi band fractal antenna for 5G vehicular communication. Test Eng. Manag. **83**, 26487–26497 (2020). ISSN: 0193-4120
15. P.K. Malik, M. Singh, Multiple bandwidth design of micro strip antenna for future wireless communication. Int. J. Recent Technol. Eng. **8**(2), 5135–5138 (2019). ISSN: 2277-3878. https://doi.org/10.35940/ijrte.B2871.078219

ECICM: An Efficient Clustering and Information Collection Method in Heterogeneous Wireless Sensor Networks

Samayveer Singh, Aruna Malik, and Pradeep Kumar Singh

Abstract The wireless sensor networks (WSNs) are deployed in many applications such as remote monitoring, precision agriculture, vaccine monitoring, etc. due to the advancement of the electronic circuitry of sensor nodes. These devices are low in size, and have low complex and battery power. A battery mechanism is required because sensor nodes have very limited energy. In these networks, longevity is required for gathering the data from the environment for a longer time on the real-time bases. To maintain the longevity in the network, in this paper, an efficient clustering and data collection method for heterogeneous WSNs is discussed. The proposed work considers four clustering parameters such as cluster heads (CH) distance to sink, density of the cluster heads, average distance of the cluster, and network residual energy are considered for electing the cluster heads dynamically. A threshold-based probability formula which correlates all four parameters is proposed for electing the cluster head. The heterogeneous nodes are considered for deploying the sensor nodes. The performance of the proposed method is compared with the existing work by considering alive and dead nodes, throughput, residual energy, and cluster heads per round as metrics. In the simulation results, the proposed method performs better than the existing protocol on all the matrices.

Keywords Clustering · Energy efficiency · Wireless sensor networks · Network lifespan · Node density

1 Introduction

In the current era, various wireless technologies like Internet of Things (IoTs), Cloud Computing (CC), Edge Computing (EC), Internet of Clouds (IoCs), Wireless Sensor Networks (WSNs) etc. are helping to connect the physical word with the cyber world

S. Singh · A. Malik
Department of Computer Science & Engineering, Dr. B R Ambedkar National Institute of Technology Jalandhar, Jalandhar, Punjab, India

P. K. Singh (✉)
Department of Computer Science, KIET Group of Institutions, Delhi-NCR, Ghaziabad, UP, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_19

devices. The devices are deployed in the physical world where data are gathered from the environment and forwarded for the further process using the cyber world. There are many applications in the current scenarios where we can deploy such type of technologies such as real-time monitoring of vaccines, real-time fertilizer monitoring, border surveillances, real-time building structure health monitoring, precision agriculture, crops threshing, underwater monitoring system, etc. These wireless technologies are deployed for solving the above-mentioned problems and provide very efficient and effective solutions [1].

In WSNs, sensor nodes are deployed in the monitoring filed for collecting the data or information according to the defined applications. The sensor nodes gathered the information from the field and forwarded the information to the sink. After collecting the data by the sink, it forwards the data to the server with the help of the Internet connection or the end user server. The deployment of nodes can be dynamic or static in WSNs. In the static deployment of nodes, sensor nodes are deployed by the human being manually. This type of deployment is not useful in the harsh environment where human intervention is not possible, whereas dynamic deployment of nodes can be done by the aircraft as per the requirement of the applications. This type of deployment is also called random deployment of sensor nodes [2–4].

Generally, sensor nodes have various capabilities like energy, link, memory, microprocessor, and other various computing capabilities. If all the nodes have same capabilities and are deployed in the monitoring field then they are called homogeneous nodes. If the nodes have different capabilities such networks are called heterogeneous networks. As discussed in many research papers [5, 6] heterogeneous networks are more capable than that of the homogeneous networks. The cost of the heterogeneous networks is higher than the homogeneous networks because these heterogeneous networks consist of higher networks capabilities as compared to the homogeneous networks. The proportion of the cost of heterogeneous networks is not much higher than the performance proportion of the network. These networks perform better than the other networks but on addition of the cost in term of capabilities of the networks [5, 6].

There are many issues in the WSNs including routing, deployment, energy efficiency, fault tolerance, calibration, stability, flexibility, etc. These problems can be easily solved by incorporating the clustering process with the nodes in the monitoring area. In this process, the total number of deployed sensor node is divided in groups and for each group a master node is selected which collects the data from the other nodes. Initially some of the nodes are chosen as the master node or cluster head (CH) [7]. These CHs cover most of the sensor by initializing the hello message and make a communication with them. If some of the nodes does not have any CH, then some random nodes are chosen as CH. After that each node in the field has CH with the condition only one sensor node can communicate with only one CH. After selecting the CH for the nodes, data collection will take place and after certain amount of time, the rotation on the CH will be performed, which helps in making the networks more efficient and balances the load among the networks efficiently [8].

In this paper, an energy efficient clustering and data collection method is discussed for prolonging the lifetime of the WSNs. The proposed work considers four clustering

parameters such as cluster heads (CHs) distance to sink, density of the cluster heads, average distance of the cluster, and network residual energy for electing the cluster heads dynamically. A threshold-based probability formula which correlates all four parameters is proposed for electing the cluster head. The heterogeneous nodes are considered for deploying the sensor nodes. The performance of the proposed method is compared with the existing work by considering alive and dead nodes, throughput, residual energy, and cluster heads per round as metrics.

The organization of the paper is given as follows: Sect. 2 discusses the literature review of the existing techniques. The systems model discusses in the Sect. 3 of the paper. Section 4 discusses the proposed method and Sect. 5 discusses the simulation results of the paper. Finally, paper is concluded in Sect. 6 with its future scope.

2 Literature Review

In this section, a comprehensive review of the existing clustering and data collection algorithms are discussed. One of the very first protocols which are discussed for WSNs is called low energy adaptive clustering hierarchy (LEACH) protocol [1]. In the LEACH, clusters are selected based on the probability formulation. If the probability value of the node is higher than the defined value of threshold, the node is considered the cluster heads (CH), otherwise node may be CH in the other node. Lindsey et al. discuss a protocol which collects data in the chaining form [2]. This approach starts to collect data from the farthest nodes and forwards that data collection to the nearer nodes, and same process follows till the data reached to the sink. However, this data collection process is not suitable for large networks. In paper [3], Chand et al. discuss a heterogeneous HEED protocol for WSNs. This approach considers three types of nodes for collecting data and fuzzy-based clustering method. However, this approach is suitable for large networks. In paper [4], Singh et al. discuss a multilevel heterogeneous network model for WSNs. This method considers an approach which consists of multiple type of heterogeneity of nodes. It can define a general model for the heterogeneity. In paper [5], Singh et al. discuss a performance investigation of energy-efficient HetSEP for prolonging lifetime in WSNs. This method shows the comparative study of the stable election protocol with three level of heterogeneity. In paper [6], Singh et al. discuss an energy aware data gathering and clustering technique based on nature-inspired optimization in WSNs. This paper discusses a clustering method by considering four different parameters such as residual energy, density, average energy, and distance. In paper [7], Ramteke et al. discusses a particle swarm optimization (PSO) and genetic mutation (GM)-based routing technique for IoT-based homogeneous software-defined WSNs. This approach discusses PSO and GM for cluster head election. However, this method does not consider the heterogeneity in the networks. In paper [8], Singh et al. discusses a method called OSEP, an optimized stable election protocol in heterogeneous WSNs. This method is the extension of [5] by considering the three levels of heterogeneity.

In paper [9], Malik et al. discuss a method called DACHE, a data aggregation-based effective and optimized cluster head election routing protocol for HWSNs. This method uses the data aggregation with the cluster process with heterogeneity. However, this method performs very well as compared to the existing methods.

In paper [10], Singh et al. discuss a clustering-based optimized stable election protocol in WSNs. This clustering process considers the residual energy, distance, density parameters for cluster head election. However, it does not consider the average energy of the nodes for cluster head election. In paper [11], Singh et al. discuss an effective analysis and performance investigation of energy heterogeneity in WSNs. This paper uses a comparative study of the stable election protocol, deterministic energy-efficient clustering, and hybrid energy efficient distributed. In paper [12], Singh et al. discuss a method called optimized cluster head election protocol for heterogeneous WSNs (OCHEP). This method uses the optimized cluster method which is based on the probability and its threshold value. In paper [13], Singh et al. discuss energy-efficient clustering protocol using fuzzy logic for heterogeneous WSNs. This method uses a fuzzy logic system by considering residual energy, distance, and density parameters for cluster head election. However, it is not defined as the general model for heterogeneity. In paper [14, 15], a stable period enhancement for zonal (SPEZ)-based clustering in heterogeneous WSN is discussed. This paper uses the fuzzy-based clustering for electing the cluster heads efficiently. However, this paper does not consider the average energy of the nodes as clustering parameters. In paper [16], the authors discuss an energy-efficient cross layer based adaptive threshold routing protocol for WSN. This paper discusses an adaptive method for clustering and data collection. However, the results are not satisfactory in this paper and lifetime may be improved by adding more concepts. In paper [17], a method is discussed for selective α-coverage based heuristic in wireless sensor networks. This paper discusses a work for target coverage problem. It does not consider the data collection process among the nodes and sinks.

In this paper, an energy-efficient clustering and data collection method are discussed for prolonging the lifetime of the WSNs. The proposed work considers four clustering parameters such as cluster heads (CH) distance to sink, density of the cluster heads, average distance of the cluster, and network residual energy for electing the cluster heads dynamically. A threshold-based probability formula which correlates all four parameters is proposed for electing the cluster head. The heterogeneous nodes are considered for deploying the sensor nodes.

3 Assumptions of the Network Energy Model and Radio Dissipation Model

Following are the assumptions of the network energy model and radio dissipation model.

- Nodes are not movable and an ID is fixed for node identification.

- Initial energy of nodes is defined.
- Location of the sink is fixed and symmetric connections are used between the sensor nodes and sink for communication.
- Heterogeneous nodes are deployed in the network.

The proposed network consists of three types of nodes called nrm, int, and adv nodes. The energies and number of the nrm, int, and adv nodes are denoted as E_1, E_2, and E_3 and n_{nrm}, n_{int}, n_{adv}, with the condition $n_{nrm} > n_{int} > n_{adv}$, respectively. The total energy of the network is calculated as follows:

$$E_{Tot} = \chi \times n_n \times E_1 + \chi^2 \times n_n \times E_2 + \left(1 - \chi - \chi^2\right) \times n_n \times E_3 \tag{1}$$

where χ is the model parameter.

The number of nrm, int, and adv nodes are $\chi \times n_n$, $\chi^2 \times n_n$, and $(n_n - (\chi * n_n + \chi^2 * n_n))$ with the E_1, E_2, and E_3 energy, respectively. When we are putting $\chi = 0$, in (1) it gives one type of nodes with the following energy of the networks, i.e., $E_{Tot} = n_n * E_3$ in 1-level heterogeneity. This is called the adv nodes instead of nrm nodes where we impose a condition for changing the adv node energy into nrm nodes energy as

$$\chi = \frac{E_3 - E_1}{\beta * f(E_2,\ E_3)} \tag{2}$$

When we are putting $1 - \chi - \chi^2 = 0$ in Eq. (1), it gives two types of sensors, namely, nrm and int nodes. This relation $1 - \chi - \chi^2 = 0$ has two solutions, i.e., $\left(\left(\sqrt{5}\right) - 1\right)/2$ and $\left(\left(\sqrt{5}\right) + 1\right)/2$. The value $\left(\left(\sqrt{5}\right) - 1\right)/2$ lies in the range between 0 and 1. Thus, it may be considered as the true value for solving the expressions in second level of heterogeneity.

In third level of heterogeneity, upper bond is $\left(\left(\sqrt{5}\right) - 1\right)/2$ denoted by χ_{ub} and let the lower bound be χ_{lb}. The range of χ is $\chi_{lb} < \chi < \chi_{ub}$ and function value f is considered as $(E_3 - E_2)$ from (2). Thus, χ_{lb} as $\chi_{lb} < \chi < \chi_{ub}$. By using the values χ and χ_{ub}, calculate χ_{lb} as below:

$$\chi_L < \frac{E_3 - E_1}{\beta * (E_3 - E_2)} < \left(\left(\sqrt{5}\right) - 1\right)/2 \tag{3}$$

Let $E_2 = \alpha_1 + E_1$ and $E_3 = \alpha_2 + E_2$ and we have the following relation $\chi_{lb} < \frac{\alpha_2 + \alpha_1}{\beta * \alpha_2}$

$$\frac{\alpha_2}{\alpha_1} < \frac{1}{\beta * \chi_{lb} - 1} = -\frac{\alpha_2}{\alpha_1} \geq \frac{1}{1 - \beta * \chi_{lb}} \tag{4}$$

By putting $-\frac{\alpha_2}{\alpha_1} = 0$ in (4) because LHS is negative and get the following relation:

$$1 - \beta * \chi_{lb} < 0 = \frac{1}{\beta} < \chi_{lb} \tag{5}$$

Putting the value in (3) as follows:

$$(E_3 - E_1) \leq \frac{\beta * \left(\left(\sqrt{5}\right) - 1\right)}{2} * (E_3 - E_2) \tag{6}$$

$$\beta * \left(\left(\sqrt{5}\right) - 1\right) * E_2 - 2 * E_1 \leq \left(\beta * \left(\left(\sqrt{5}\right) - 1\right) - 2\right) * E_3 \tag{7}$$

The nrm, int, and adv nodes energy are given as e_1, $e_2 = e_1 * (1 + \omega)$, and $e_3 = e_1 * (1 + \eta)$, respectively. And the value of coefficients ω and η are 0.06 and 0.11, respectively.

The energy consumption in transmitting, receiving, and sensing L-bit data between the sensor nodes and BS are given as follows [1]:

$$E_TXS = L * \in_{elec} + L * \in_{(fs)} * d^2 \quad \text{if } d \leq d_0 \tag{8}$$

$$E_{TXL} = L * \in_{elec} + L * \in_{mp} * d^4 \text{ if } d > d_0 \tag{9}$$

$$E_{Rx} = L * \in_{elec} \tag{10}$$

$$E_{Sx} = L * \in_{elec} \tag{11}$$

where $\in_{elec}$, $\in_{fs}$, and $\in_{mp}$ are the energy consumption in different spaces and d_0 threshold distance is given below:

$$d_0 = \sqrt{\frac{\in_{fs}}{\in_{mp}}} \tag{12}$$

4 ECICM: An Efficient Clustering and Information Collection Method in HWSNs

This work discusses an efficient clustering and data collection method for heterogeneous WSNs. In this work, four clustering parameters such as cluster heads (CH) distance to sink, density of the cluster heads, average distance of the cluster and network residual energy are considered for electing the cluster heads dynamically. This CH election method consumes very less energy as compared to the latest papers

in data collection. The process of the efficient clustering with data collection is divided into rounds. Initially, a predefined percentage of nodes are used as cluster heads (e.g., 5% nodes) by considering a constant as two CHs cannot be equipped in the range of each. The selection of CH depends on the threshold value between 0 and 1 initially but in the latter rounds it is calculated as follows:

$$T(n_i) = \frac{p_i}{1 - p_i\left[rmod\left(\frac{1}{p_i}\right)\right]} \times \frac{\mathrm{Dist_{avg}}}{\mathrm{Dist_{BS}}} \times \left[\frac{\mathrm{E_{resi}}}{\mathrm{E_{init}}} + \left(rdiv\left(\frac{1}{p_i}\right)\right)\left(1 - \frac{\mathrm{E_{resi}}}{\mathrm{E_{init}}}\right)\right] \tag{13}$$

where $T(n_i)$ indicates the threshold value for a particular type of nodes. p_i and r indicate the probability of the particular node and the round. $\mathrm{Dist_{avg}}$, $\mathrm{Dist_{BS}}$, $\mathrm{E_{resi}}$, and $\mathrm{E_{init}}$ indicate the average distance from the CH, distance between the CH and nodes, residual energy of the network, and initial energy of the CH in the networks, respectively.

The total energy of the networks (E_{Tol}) is given as follows:

$$\begin{aligned} E_{Tol} &= \chi \times n_n \times E_1 + \chi^2 \times n_n \times E_2 + \left(1 - \chi - \chi^2\right) \times n_n \times E_3 \\ &= n_n \times \left(\chi \times E_1 + \chi^2 \times E_2 + \left(1 - \chi - \chi^2\right) \times E_3\right) \end{aligned} \tag{14}$$

$$E_{Tol} = E_1 \times n_n \times \left(\chi + \chi^2 \times E_2/E_1 + \left(1 - \chi - \chi^2\right) \times E_3/E_1\right) \tag{15}$$

E_1 is defined as the initial energy of a node and $E_1 \times n_n$ is the total network energy. $(\chi + \chi^2 \times E_2/E_1 + (1 - \chi - \chi^2) \times E_3/E_1)$ is indicating the incremental factor due to their energy heterogeneity which means heterogeneous network is having $(\chi + \chi^2 \times E_2/E_1 + (1 - \chi - \chi^2) \times E_3/E_1)$ times more energy than that of the homogeneous nodes. The nodes become CH after $1/\mathrm{p}_i \times (\chi + \chi^2 \times E_2/E_1 + (1 - \chi - \chi^2) \times E_3/E_1)$ rounds and the average CH in the heterogeneous networks are as $n_n \times 1/\mathrm{p}_i \times (\chi + \chi^2 \times E_2/E_1 + (1 - \chi - \chi^2) \times E_3/E_1)$.

The weighted probabilities of the proposed protocol ECICM nrm, int, and adv nodes are p_{nrm}, $p_{int,}$ and p_{adv} given as follows:

$$\begin{aligned} p_{nrm} &= \frac{\mathrm{p_i} \times \mathrm{E_1}}{\left(\chi \times E_1 + \chi^2 \times E_2 + \left(1 - \chi - \chi^2\right) \times E_3\right)} \\ &= \frac{\mathrm{p_i}}{\left(\chi + \chi^2 \times E_2/E_1 + \left(1 - \chi - \chi^2\right) \times E_3/E_1\right)} \end{aligned} \tag{16}$$

$$\begin{aligned} p_{int} &= \frac{\mathrm{p_i} \times \mathrm{E_1}}{\left(\chi \times E_1 + \chi^2 \times E_2 + \left(1 - \chi - \chi^2\right) \times E_3\right)} \\ &= \frac{\mathrm{p_i}}{\left(\chi + \chi^2 \times E_2/E_1 + \left(1 - \chi - \chi^2\right) \times E_3/E_1\right)} \end{aligned} \tag{17}$$

$$p_{adv} = \frac{\mathrm{p_i} \times \mathrm{E_l}}{(\chi \times E_1 + \chi^2 \times E_2 + (1 - \chi - \chi^2) \times E_3)} = \frac{\mathrm{p_i}}{(\chi + \chi^2 \times E_2/E_1 + (1 - \chi - \chi^2) \times E_3/E_1)} \tag{18}$$

The threshold value of the nrm, int, and adv nodes are given as follows:

$$T(nrm_i) = \frac{p_{nrm_i}}{1 - p_{nrm_i}\left[rmod\left(\frac{1}{p_{nrm_i}}\right)\right]} \times \frac{\mathrm{Dist_{avg}}}{\mathrm{Dist_{BS}}} \times \left[\frac{\mathrm{E_{resi}}}{\mathrm{E_{init}}} + \left(rdiv\left(\frac{1}{p_{nrm_i}}\right)\right)\left(1 - \frac{\mathrm{E_{resi}}}{\mathrm{E_{init}}}\right)\right] \tag{19}$$

$$T(int_i) = \frac{p_{int_i}}{1 - p_{int_i}\left[rmod\left(\frac{1}{p_{int_i}}\right)\right]} \times \frac{\mathrm{Dist_{avg}}}{\mathrm{Dist_{BS}}} \times \left[\frac{\mathrm{E_{resi}}}{\mathrm{E_{init}}} + \left(rdiv\left(\frac{1}{p_{int_i}}\right)\right)\left(1 - \frac{\mathrm{E_{resi}}}{\mathrm{E_{init}}}\right)\right] \tag{20}$$

$$T(adv_i) = \frac{p_{adv_i}}{1 - p_{adv_i}\left[rmod\left(\frac{1}{p_{adv_i}}\right)\right]} \times \frac{\mathrm{Dist_{avg}}}{\mathrm{Dist_{BS}}} \times \left[\frac{\mathrm{E_{resi}}}{\mathrm{E_{init}}} + \left(rdiv\left(\frac{1}{p_{adv_i}}\right)\right)\left(1 - \frac{\mathrm{E_{resi}}}{\mathrm{E_{init}}}\right)\right] \tag{21}$$

The $T(n_{nrm})$, $T(n_{int})$, and $T(n_{adv})$ indicate the threshold functional to nrm, int, and adv nodes, respectively. By using this procedure, a cluster efficient cluster nodes are selected for collecting the data.

5 Experimental Results and Discussion

In this paper, the ATEER [16] protocol and the proposed method are simulated by considering the different metrics such as number of alive nodes, dead nodes, packets sent to BS, and number of CH. The number of alive nodes refers to the time whenever a single node is alive in the network. The number of dead nodes defines as the time when all the sensor nodes are dead. The packet sent to BS refers to the number of packets sent to the base station per unit time. It is also called the threshold of the networks. The number of CH refers to the cluster per round in the networks. This work discusses an efficient clustering and data collection method for heterogeneous WSNs. In this work, four clustering parameters such as cluster heads (CH) distance to sink, density of the cluster heads, average distance of the cluster, and network residuals energy are considered for electing the cluster heads dynamically. The ATEER [16] protocol and the proposed method are implemented in MATLAB with the 11th Gen

core Ii7 operating system and 16 GB RAM machine. In this network, 100 number of sensor nodes are deployed in the area of 100m × 100m. Three types of nodes are considered as nrm, int, and adv with 50, 30, and 20 in numbers and with their respective energy 0.10 J, 0.16 J, and 0.21 J, respectively. The message length is considered as 4000 bits, threshold distance value is considered as 75 m, and 25 m is considered as cluster range. The depletion of the energy is 50 nJ/bits in management of the circuit, 50nJ/bit, $10pJ/bit/m^2$and$0.0013pJ/bit/m^4$ to transmit the signal in the shorter and long distance, respectively.

Figure 1 shows a simulation instance for the ATEER [16] protocol and the proposed method which indicates the deployed sensor nodes and sink. This figure shows the cluster heads and their respective sensor nodes and cluster heads connection to the BS while transferring the data.

Figure 2 shows the simulation results in terms of number of rounds with respect to the number of alive nodes for the ATEER [16] protocol and the proposed method. The proposed method have higher alive nodes in number of rounds as compare to the ATEER [16] protocol. Also the stability period of the proposed work is much higher than the existing protocol ATEER [16] because of the first node death after a longer time in the network. The first node depicted its energy in case of the proposed method and the ATEER [16] is 1202 and 3012 rounds, respectively. It is evident from the result that the increment in the stability period is increased by 150% as compared to the ATEER [16]. The reason behind this gigantic improvement is not only the distance and residual energy parameters but also the average energy and node density which helps in selecting the effective cluster heads. This effective clustering

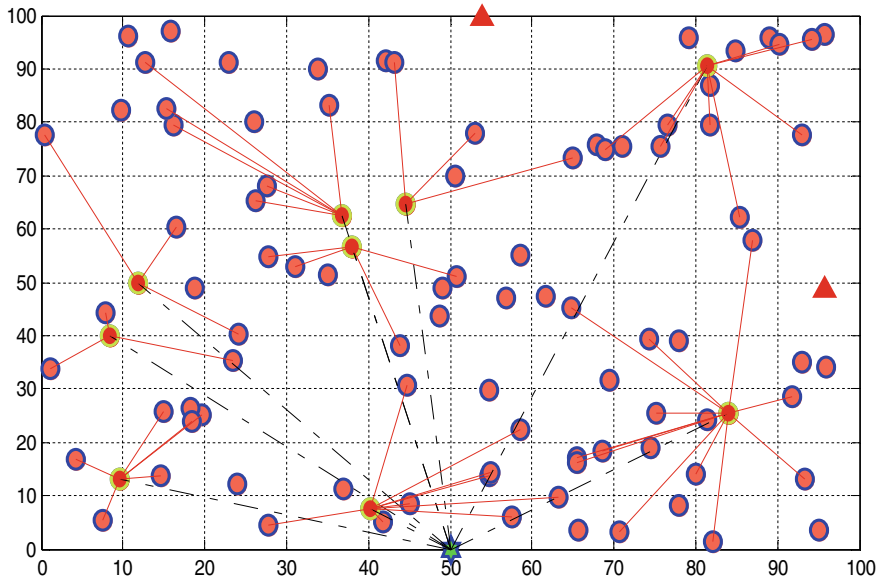


Fig. 1 Number of deployed sensor nodes and respective sink

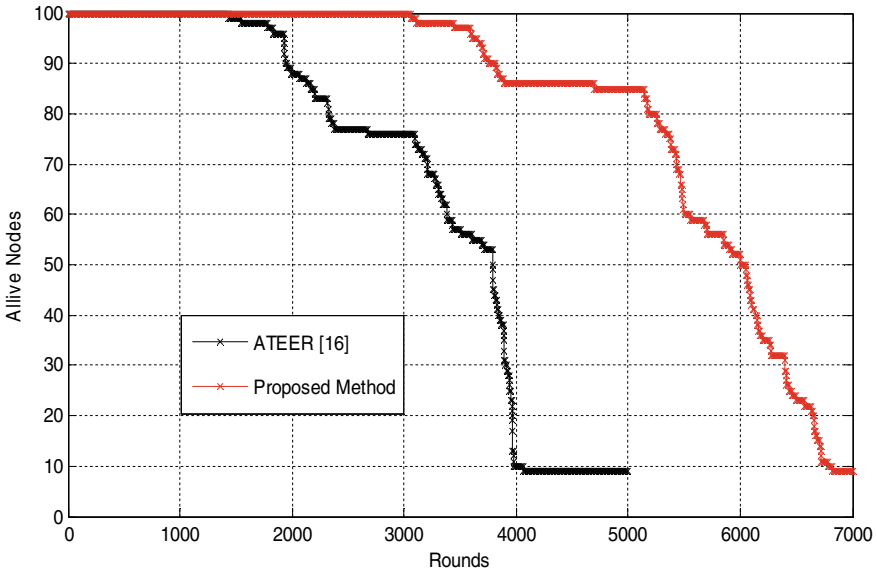


Fig. 2 Number of rounds versus alive nodes

reduces the distance among the cluster heads and the sink for collecting the data in inter-and intra-cluster communication.

Figure 3 shows the simulation results in terms of number of rounds with respect to the number of dead nodes for the ATEER [16] protocol and the proposed method. The first node is depicted dead after 1202 and 3012 rounds, half nodes are dead after 3810 and 6000 rounds, and 4983 and 6932 rounds for both the proposed and ATEER [16] method. This improvement in the results is because of the energy-efficient CH selection which reduces energy expenditure in the network. In addition to efficient CH election, the proposed method also considered the average distance among the nodes and sink for transmitting the data. The average distance reduces the intra-cluster distance and inter-cluster distance which further reduces the communication cost of nodes. This improves the overall network lifetime due to its energy preservation.

Figure 4 shows the simulation results in terms of the number of rounds with respect to the number of packets sent to BS for the ATEER [16] protocol and the proposed method. The proposed method is transmitting more number of packets to the base station as compared to the existing method. The improvement is because of the reduction in the intra-cluster distance and inter-cluster distance which further decreases the communication cost of nodes. Because of the network longevity, more number of packets is transmitted to the BS. Single hop communication has also reduced the energy consumption and increases the network lifetime.

Figure 5 shows the simulation results in terms of number of rounds with respect to the number of cluster heads for the ATEER [16] protocol and the proposed method. The cluster range for the ATEER [16] protocol and the proposed method is varying

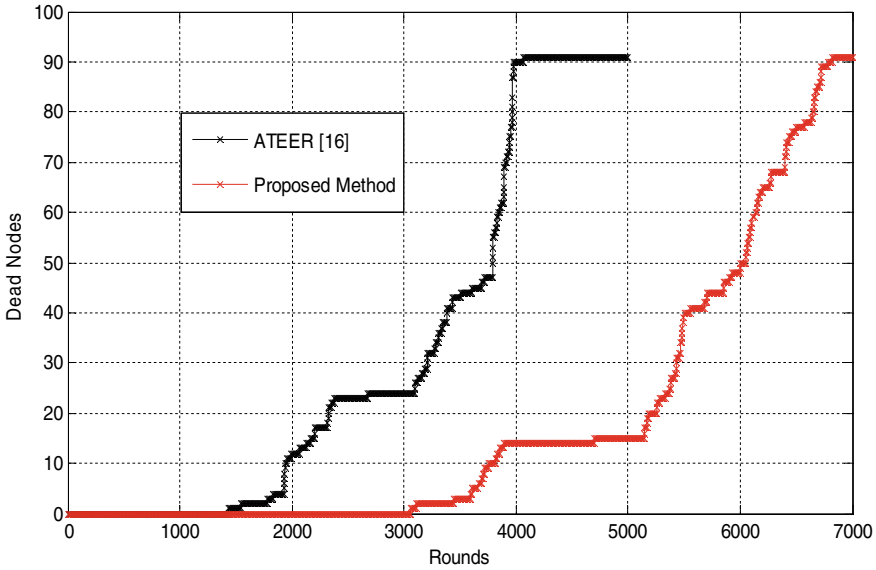


Fig. 3 Number of rounds versus dead nodes

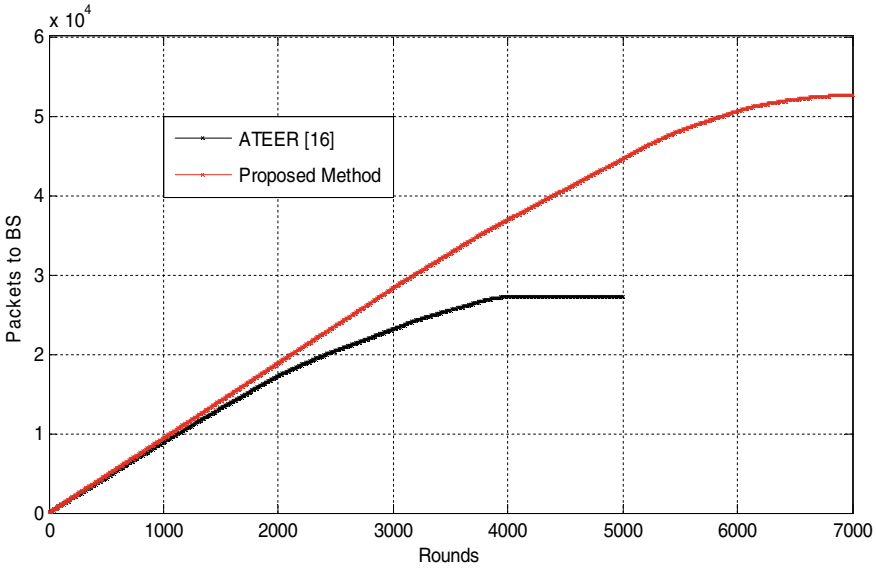


Fig. 4 Number of rounds versus packets sent to BS

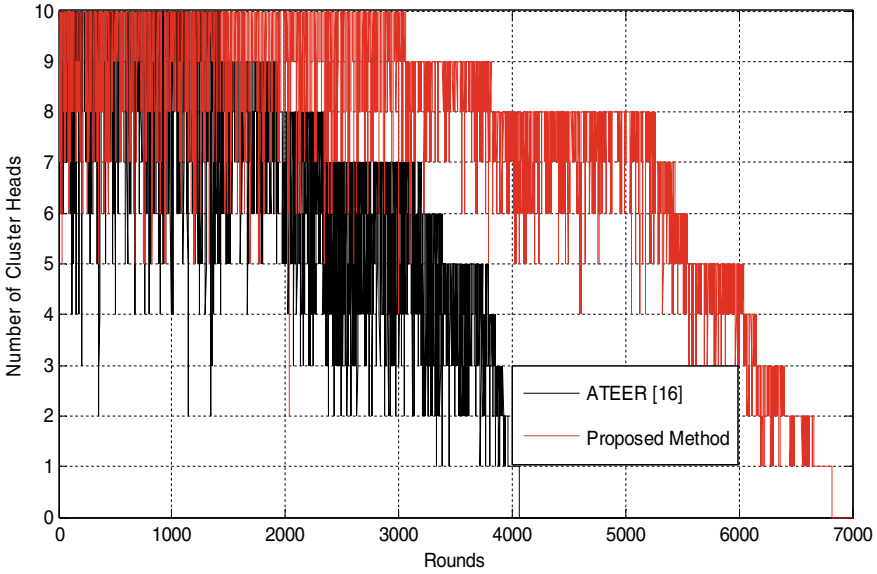


Fig. 5 Number of rounds versus cluster heads

between the 0 and 10. The average number of cluster heads in the ATEER [16] protocol is 8 or 9, whereas the average number of cluster heads in the proposed protocol is 5 or 6. In most of the rounds the number of cluster heads is varying by 1. The proposed method works for long when compared to the ATEER [16] protocol. Thus, approximately 10% nodes of the total number of nodes are more than sufficient for the cluster heads. If we have very large number of cluster heads, then the lifetime of the networks will not be well as discussed in the other word. Therefore, the cluster heads should be defined in the well-structured range.

6 Conclusion

In this paper, an efficient clustering and data collection method for heterogeneous WSNs is discussed. The proposed work considers four clustering parameters such as cluster heads (CHs) distance to sink, density of the cluster heads, average distance of the cluster, and network residual energy for electing the cluster heads dynamically. A threshold-based probability formula which correlates all four parameters is proposed for electing the cluster head. The heterogeneous nodes are considered for deploying the sensor nodes. The first node depicted its energy in case of the proposed method and the ATEER [16] is 1202 and 3012 rounds, respectively. It is evident from the result that the increment in the stability period is increased by 150% as compared to the ATEER [16]. The reason behind this gigantic improvement is not only the distance

and residual energy parameters but also the average energy and node density which helps in selecting the effective cluster heads. This effective clustering reduces the distance among the cluster heads and the sink for collecting the data in inter-and intra-cluster communication.

References

1. W.B. Heinzelman, A.P. Chandrakasan, H. Balakrishnan, An application-specific protocol architecture for wireless microsensor networks. IEEE Trans. Wirel. Commun. **1**(4), 660–670 (2002)
2. S. Lindsey, C.S. Raghavendra, K.M. Sivalingam, Data gathering algorithms in sensor networks using energy metrics. IEEE Trans. Parallel Distrib. Syst. **13**(9), 924–935 (2002)
3. S. Chand, S. Singh, B. Kumar, Heterogeneous HEED Protocol for Wireless Sensor Networks. Wirel. Pers. Commun. **77**, 2117–2139 (2014)
4. S. Singh, S. Chand, B. Kumar, Multilevel heterogeneous network model for wireless sensor networks. Telecommun. Syst. **64**, 259–277 (2017)
5. S. Singh, P.K. Singh, Performance investigation of energy efficient HetSEP for prolonging lifetime in WSNs. in *Futuristic Trends in Network and Communication Technologies. FTNCT 2018*, ed. by P. Singh, M. Paprzycki, B. Bhargava, J. Chhabra, N. Kaushal, Y. Kumar. Communications in Computer and Information Science, vol. 958 (Springer, Singapore, 2019)
6. S. Singh, An energy aware clustering and data gathering technique based on nature inspired optimization in WSNs. Peer-to-Peer Netw. Appl. **13**, 1357–1374 (2020)
7. R. Ramteke, S. Singh, Particle swarm optimization and genetic mutation based routing technique for IoT-based homogeneous software-defined WSNs. in *Recent Innovations in Computing. ICRIC 2020*, ed. by P.K. Singh, Y. Singh, M.H. Kolekar, A.K. Kar, J.K. Chhabra, A. Sen. Lecture Notes in Electrical Engineering, vol. 701 (Springer, Singapore, 2021)
8. S. Singh, P.K. Singh, A. Malik, OSEP: an optimized stable election protocol in heterogeneous wireless sensor networks. in *Recent Innovations in Computing. ICRIC 2020*, ed. by P.K. Singh, Y. Singh, M.H. Kolekar, A.K. Kar, J.K. Chhabra, A. Sen. Lecture Notes in Electrical Engineering, vol. 701 (Springer, Singapore, 2021)
9. Malik A., Singh S., Singh P.K. (2021). DACHE: A Data Aggregation-Based Effective and Optimized Cluster Head Election Routing Protocol for HWSNs. In: Singh P.K., Singh Y., Kolekar M.H., Kar A.K., Chhabra J.K., Sen A. (eds) Recent Innovations in Computing. ICRIC 2020. Lecture Notes in Electrical Engineering, vol 701. Springer, Singapore.
10. S. Singh A clustering-based optimized stable election protocol in wireless sensor networks. in *Applications in Ubiquitous Computing*, ed. by R. Kumar, S. Paiva. EAI/Springer Innovations in Communication and Computing, (Springer, Cham, 2021)
11. S. Singh, R. Kumar, P.K. Singh, An effective analysis and performance investigation of energy heterogeneity in wireless sensor networks. in *Handbook of Wireless Sensor Networks: Issues and Challenges in Current Scenario's*, ed. by P. Singh, B. Bhargava, M. Paprzycki, N. Kaushal, W.C. Hong. Advances in Intelligent Systems and Computing, vol. 1132 (Springer, Cham, 2020)
12. S. Singh, P. Yadav, A. Malik, R. Agrawal, OCHEP: an optimized cluster head election protocol for heterogeneous WSNs. in *Advances in Smart Communication and Imaging Systems*, ed. by R. Agrawal, C. Kishore Singh, A. Goyal. Lecture Notes in Electrical Engineering, vol. 721 (Springer, Singapore, 2021)
13. S. Singh, S. Chand, B. Kumar, Energy efficient clustering protocol using fuzzy logic for heterogeneous WSNs. Wirel. Pers. Commun. **86**, 451–475 (2016)
14. P.S. Mehra, M.N. Doja, B. Alam, Stable period enhancement for zonal (SPEZ)-based clustering in heterogeneous WSN. Smart Innov. Syst. Technol. **79**, 887–896 (2018)
15. P.S. Mehra, M.N. Doja, B. Alam, Stable period extension for heterogeneous model in wireless sensor network. Adv. Intell. Syst. Comput. **638**, 479–487 (2018)

16. R. Singh, A.K. Verma, Energy efficient cross layer based adaptive threshold routing protocol for WSN. AEU Int. J. Electron. Commun. **72**, 166–173 (2017). https://doi.org/10.1016/J.AEUE.2016.12.001
17. M.S. Chand, B. Kumar, *Selective α-Coverage Based Heuristic in Wireless Sensor Networks, Wireless Personal Communication*, vol. 97 (Springer, 2017), pp. 1623–1636

Exploring Trust in SDN Along with Network Monitoring

Gaurav Sharma and Sushopti Gawade

Abstract Software-defined network is a new paradigm that brings programmability to networking. This is achieved by separating the Control Plane and the Data Plane, which also provides more flexibility in managing the networks. However, this also gives rise to certain security issues that need to be addressed. To address the security issues in SDN, many researchers have proposed novel solutions. One direction of introducing security mechanisms to SDN is the trust-based approach. In this paper, we discuss the challenges in SDN security and present a detailed survey of the existing trust-based methods proposed to secure software-defined networks. These methods are based on different techniques and focus only on only a part of the SDN architecture. The paper, then, explores the use of network monitoring at the different planes to develop a trust-based mechanism that can be used at all the planes of the SDN architecture and provide a comprehensive solution.

Keywords SDN · Trust · Security · Control Plane · Data Plane · Network Monitoring

1 Introduction

Software-defined network (SDN) has emerged as a new paradigm in networking that has introduced programmability in networking. In this approach, the Control Plane and the Data Plane are separated, and this gives more flexibility in managing the networks [1]. This separation also allows one controller in the Control Plane to control multiple forwarding devices, i.e., switches. SDN Controllers at the Control Plane can be programmed as they are essentially servers that give APIs. This facility allows any application to be integrated with the controller. This is possible as all the

G. Sharma (✉)
Pillai HOC College of Engineering and Technology, Navi Mumbai, India

S. Gawade
Pillai College of Engineering, Navi Mumbai, India
e-mail: sgawade@mes.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_20

available controllers provide RESTful APIs [2] along with other types of APIs like JAVA, Python, etc.

The role of the Control Plane is to provide different network-related functions. In SDN, the controller is placed at the Control Plane, which communicates with the applications and the switches using the North-Bound and South-Bound APIs, respectively. The controller provides the forwarding rules (flows) to the Data Plane devices and provides data, statistics, and interfaces to manage the network to the applications.

The role of the Data Plane is to provide the forwarding function in the network. It stores the flow rules received by the controller using the South-Bound API, such as OpenFlow protocol. If there are no rules for a particular flow, it asks the controller for the flow rule for the same.

The application plane is where the applications are integrated with the network by using the controller's North-Bound APIs. Restful, Java, Python, and other APIs can be used to enable communication between the application and the controller, as shown in Fig. 1.

With all the advantages SDN brings in, it also brings a new set of challenges related to security. The separation of network functions in Control Plane and Data Plane also creates new vulnerabilities in the SDN architecture.

SDN is still evolving and due to the lack of inherent SDN security mechanisms to fix the vulnerabilities in the SDN architecture, there is a clear need for an external security mechanism that can address those vulnerabilities.

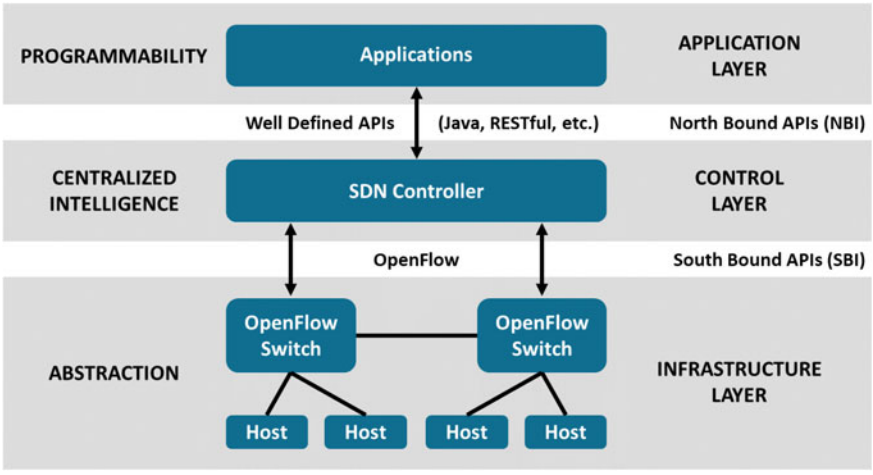


Fig. 1 SDN architecture [1]

2 Challenges in SDN Security

With programmability in networks, SDN brings in a lot of benefits such as traffic programmability, agility, and the ability to create policy-driven network supervision and implement network automation. But with these advantages, SDN also brings a new set of challenges related to security. The separation of network functions in Control Plane and Data Plane also creates new vulnerabilities in the SDN architecture.

The SDN controller can become a single point of failure and a prime target to bring down the network with attacks such as DoS/DDoS. SDN controller holds the topology and traffic-related data about the entire network and may act as a source of information to a hacker if compromised.

A malicious or vulnerable SDN application deployed also poses a great threat to the controller, and the data it holds as the controllers have no mechanisms to assess the eligibility, legality, or trustworthiness of the applications on their own. The absence of authentication and authorization of the applications, conflicts between the applications regarding the flow rules, and opportunities for malicious applications to insert fraud traffic are some of the serious problems with the SDN controllers [3].

Switches in SDN at the Data Plane can also become the target of DoS/DDoS attacks bringing down network segments or can be targeted to modify the flow table stored in the OpenFlow switches through the hosts.

Kreutz et al. [4] have identified the threat vectors in SDN security—(a) forged or fake traffic flows, (b) attacks on vulnerabilities in switches, (c) attacks on Control Plane communication, (d) attacks on and vulnerabilities in controllers, (e) lack of mechanisms to ensure trust between the controller and management applications, (f) attacks on and vulnerabilities in administrative stations, (g) lack of trusted resources for forensics and remediation.

Yoon et al. [5] have presented a comprehensive analysis and categorization of the vectors for potential abuse or attack in SDN. The authors have identified 22 such vectors, which include the attacks on Control Plane, Control Channel, and Data Plane.

Scott-Hayward et al. [6] also present a detailed comparison of the research work at different SDN layers/interfaces. It categorizes the potential attacks to which the SDN architecture is vulnerable.

Akhunzada et al. [7] have come up with a contemporary layered or interface taxonomy of the reported security vulnerabilities, attacks, and challenges of the SDN to illustrate the main categories of security implications of each SDN layer/interface. They also compare and analyze the prominent security solutions presented by various researchers.

Li et al. [8] have focused on the OpenFlow-based SDN with the security challenges and counter-measures. It considers the attacks on confidentiality, integrity, and availability in the SDN and categorizes the security challenges as—switch-related, controller-related, and channel-related security challenges.

Shaghaghi et al. [9] have looked at the most recent improvements in SDNs, identified the most common vulnerabilities, and proposed a new attack taxonomy for

SDNs. They have also presented a thorough analysis of the challenges in securing the SDN Data Plane and Control Plane, as well as recommended solutions.

Recently, Yurekten et al. [10] have presented the threat categorization for SDN based on threat intelligence. It categorizes the possible attacks such as—scanning attacks, spoofing attacks, DoS attacks, sniffing attacks, malware and social engineering attacks, web application attacks, and attacks on the hardware, OS, and processes.

Thus, it can be summarized that the SDN faces security challenges similar to the traditional networks, and more. Also, there is a need for a solution, which can be deployed in the current setting, with Hybrid SDN. A hybrid network also uses traditional, or legacy, network devices. To propose a multi-plane trust-based security architecture, it is important to understand how Hybrid-SDN works and is deployed [11, 12].

3 Literature Review—Trust-Based Security in SDN

Yao et al. [3] have proposed a Trust Management Framework (TMF) for SDN applications in which the trust value of an application is evaluated and is also dynamically updated based on its performance, which is continuously monitored by the Network Performance Monitor (NPM) module with the help of Trusted Platform Module (TPM). The proposed TMF can also detect and solve the flow rule conflicts between applications. It can also detect malicious applications.

Duy et al. [13] have proposed "Trust Trident", a trust-based authentication framework, which manages secure communication between the applications and the controller. The proposed framework work like a plug-in that intercepts the requests from applications opens them for analysis and evaluation of their privileges and if found correct, forwards them to the controller. It authenticates the application by the authentication module and the authorization module checks for the permissions of the application to consume network resources.

Chowdhary et al. [14] have proposed "TRUFL", a distributed trust management framework. The authors have identified threat vectors—Rogue Insider Attack and Compromised Switch Attack. The framework uses Public Key Infrastructure (PKI) to establish trust between only legitimate devices to communicate between different planes. It works over OpenStack and is developed to be deployed in a cloud environment.

Pisharody et al. [15] have proposed a security policy analysis framework to check the vulnerability of flow rules in SDN-based cloud environments. The authors have introduced a framework that monitors and maintains a conflict-free environment, including strategies to automatically resolve conflicts. They have classified and described the possible conflicts among flow rules in SDN's forwarding table that can cause information leakage. The framework detects flow rule conflicts in multiple SDN controllers and works to avoid them.

Aliyu et al. [16] have proposed a trust management mechanism for a secured interaction between the SDN Controller and the network applications. The applications are authenticated and authorization rules, or privileges, that determine network resources that the application can use, are set. Extending their work in Ref. [17], the authors have introduced and implemented a token-based authentication method for authenticating the network applications and assigning them their privilege permissions.

Betgé-Brezetz et al. [18] have proposed an intermediary layer between the controller and the switch called Trusted Oriented Controller Proxy (ToCP). It compares the flow rules from different redundant controllers and installs the most trusted one on the switch. Every controller participates in configuring the paths and has a confidence level. The path is configured only if the cumulative trust level is more than the threshold.

Banse et al. [19], have presented a web-based, secure, independent northbound interface to the controller, which supports the deployment of external applications. The SDN applications and the controller communicate with each other through an encrypted channel. The framework contains an integrated trust manager that enables the restriction of the interface to only authenticated and trusted applications. A configurable permission system allows the enforcement of authorization or permissions to the applications.

Isong et al. [20] have proposed a trust model to protect the controller by allowing only the applications that meet a minimum trust level to communicate with the controller and consume network resources. The researchers have proposed a trust establishment model between the application and the controller and use a trust matrix of applications corresponding to their identity used for effective resources management in SDN. It uses direct trust, which is based on the observation of activities performed by SDN applications when requesting network resources and also on previous interaction between the controller and applications, which was either successful (s) or a failure (f).

Burikova et al. [21] have proposed a trust management framework that establishes and manages the trust between the application and the controller for the SDN-based IoT platform. It includes authentication of the applications by the controller and their authorization (verifying permissions). Applications are also categorized as Security Applications (highest priority) and Non-Security Applications (lowest priority) and get permissions accordingly. Trust is calculated based on factors—reputation, operational risk, information risk, and privacy level.

Cui et al. [22] have proposed an approach to protect the northbound interface by introducing the application authentication system, which addresses the key challenges of resolving conflicts between untrusted applications and requests. The authentication process includes permission checks to verify the access permission of the application.

Scott-Hayward et al. [23] have presented an approach to secure the northbound interface by introducing a permission-based trust management mechanism for SDN—OperationCheckpoint. It ensures that controller operations are available to trusted applications only. However, it has limitations that cannot change the rights of the application dynamically to prevent their malicious behaviors.

Shin et al. [24] have proposed FRESCO, which is an OpenFlow security application development framework. It facilitates rapid design and modular composition of OF-enabled detection and mitigation modules. To enforce flow constraints to defend the network against threats, the authors have introduced the security enforcement kernel [24, 25].

Lamb et al. [26] outline the constraints of trust within SDN. It discusses the different requirements for communication between different network entities. Sood et al. [27] have done a theoretical analysis of policy-based security management or architecture (PbSA), which dynamically controls the network.

Trust-based approaches discussed above focus only on a single plane. Mostly, the focus is on introducing trust between the controller and the applications, in the North-Bound Interface (NBI).

A trust evaluation method for the Data Plane nodes based on direct, indirect, and historical trust values has been presented by Zhao et al. [28]. The method stores the trust scores using Blockchain, hence, they have named it TrustBlock. It considers the security (identity authentication) and reliability (forwarding status) for trust calculation.

Table 1 provides a summary of some prominent trust-based security methods, which were proposed for securing the SDN. It also gives information about the layer

Table 1 Prominent trust-based security methods proposed for SDN

Sr.	Research work	SDN layer/interface	Comment
1	Betgé-Brezetz et al. [18], (2015)	SBI, Data Plane	Uses polling method to get a cumulative trust score of the network Redundant controllers will add to cost and complexity
2	Yao et al. [3], (2018)	Application Plane, NBI	Requires a hardware chip—Trusted Platform Module (TPM)
3	Duy et al. [13], (2019)	Application Plane, NBI	Uses authentication and authorization of applications via a separate Trust Trident REST API
4	Chowdhary et al. [14], (2019)	Data Plane, Control Plane, SBI	Uses PKI-based mechanism to establish trust between the controller and the switches
5	Burikova et al. [21], (2019)	Application Plane, NBI	Uses the factors—reputation and risk for trust calculation of applications by monitoring their behavior
6	Aliyu et al. [17], (2020)	Application Plane, NBI	Uses authentication and authorization to evaluate the trust of network applications

of SDN architecture, and the interface, at which they work. This information is based on the way these proposals work.

4 Monitoring of SDN

One approach toward the security of networks is to monitor network behavior. In traditional networks, this has been used to detect and then mitigate security incidents. Many researchers have used this approach to enhance the security of SDN as well. This approach is well suited for monitoring the behavior of Data Plane devices, i.e., the Switches, and their communication with the Controller.

Giotis et al. [29] and Zaalouk et al. [30] have used and explored traditional monitoring protocols such as NetFlow/IPFIX and sFlow with the SDN protocol, OpenFlow, to monitor the SDN network.

Yu et al. [31] have proposed FlowSense, which uses a push-based approach to receive flow statistics from the switches. In this approach, instead of monitoring the switches by polling actively, their performance is measured by passively capturing and analyzing control messages between the switches and the centralized controller.

Phan et al. [32] have proposed an SDN monitoring framework that separates the monitoring logic from the forwarding logic called SDN-Mon. This approach provides a very fine-grained and efficient way of monitoring, which is independent of the forwarding function, making it more flexible for various management applications.

Fawcett et al. [33] have introduced Tennison, a multi-level distributed monitoring and remediation framework for SDN. It brings in the concept of a separate Tennison Coordinator, which works along with the SDN Controller. It can be used to detect DoS/DDoS attacks, scanning attacks, intrusions and can protect against them. The Tennison Coordinator also uses traditional network monitoring techniques like IPFIX and sFlow for monitoring.

Tsai et al. [34] present an overview of SDN monitoring solutions identifying the challenges and open issues. It provides a comprehensive review of monitoring SDN. It also compares traditional network monitoring with SDN monitoring in terms of Collection, Pre-processing, Transmission, Analysis, Presentation of traffic over the networks.

4.1 Trust Based on SDN Monitoring

Over the years, a lot of work has been done to monitor the performance of the SDN. Traditional network monitoring mechanisms have also been used for SDN monitoring, as discussed in the work mentioned above. However, real-time monitoring of SDN to generate Trust Values for the network entities such as the SDN controller, switches, and applications along with the traditional networking devices, has not been explored. An application based on such a system can be incorporated into existing

tools, like SIEM, and can provide a real-time trust score for all the network entities and provide the security monitoring team with a quick view of the health of the network devices, applications, and segment.

SDN monitoring tasks may include monitoring authentication and authorization processes of the network entities along with monitoring of their behavior and security incidents.

A different set of parameters, mechanisms, and protocols may be used at different planes to calculate a trust score. The idea behind using a trust score is to come up with a single score that can help the security team quickly identify security issues, which may be investigated further. The decisions for network access to certain applications, selecting the best route for forwarding, and the preference of certain devices over the others may be based on the trust scores.

SDN monitoring-based approach may be deployed with less overhead in terms of new technologies and skills as the existing network monitoring tools can be used such as NetFlow/IPFIX and sFlow. Security data, such as security incidents and anomalous behavior, from already deployed tools and techniques like Intrusion Detection Systems, firewalls, and other security mechanisms can also be used for assessing the trust score for the network entities.

For authentication and authorization of network entities, existing Authentication, Authorization, and Accounting (AAA) mechanisms can be used. This will reduce the time, complexity, and need for additional skills required to set up a security solution based on trust at multiple planes of the SDN architecture.

The tasks that can be performed at the different planes, i.e., the Application, Control, and Data Planes, and at the communication interface, i.e., North-Bound and South-Bound Interfaces are described in Fig. 2.

4.2 Test-Bed for SDN with GNS3

For implementing and testing their solutions, researchers deploy an SDN test-bed with multiple nodes. Alternatively, a network simulator or an emulator, like Mininet, is also often used to have a test-bed with a greater number of devices. But such an environment presents very limited functionalities.

To have an SDN environment that can simulate a large-sized network, a basic test-bed with Virtual Machines and Docker Containers in GNS3 was created to understand the configuration and working of SDN.

GNS3 supports Virtual Machines, (VMWare and VirtualBox) and Docker Images. It also supports Cisco traditional network devices and has a ready OpenvSwitch appliance available on its marketplace for free.

The setup of an SDN network using GNS3, as shown in Fig. 3, is more practical for working as Virtual Machines and Docker Containers, which run the real operating systems, are used. This gives us the flexibility to run any tool or protocol, just like on a physical network with real devices, in a controlled environment.

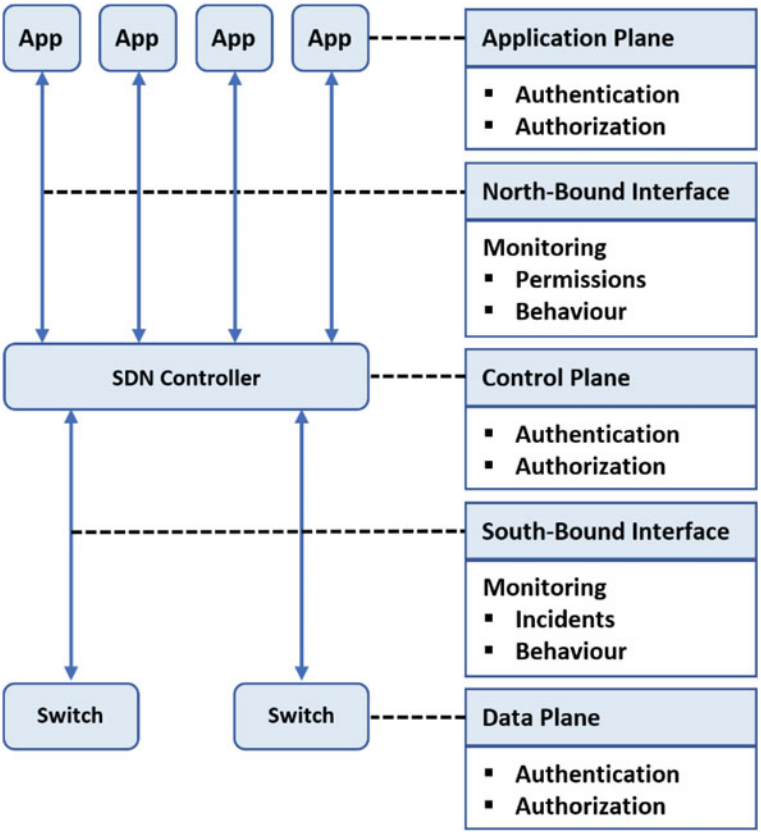


Fig. 2 SDN monitoring tasks

As an example, in Fig. 4, an SDN test-bed having four nodes, which are Docker Containers running Ubuntu Linux OS, are connected to an OpenvSwitch appliance with 16 NICs. The network is connected to the SDN Controller, which can control the entire network. Here, the SDN Controller is the OpenDayLight Controller installed on an Ubuntu Virtual Machine. The VM running the SDN Controller can also connect to other networks and the Internet as well.

5 Conclusion

Many solutions have been proposed for securing the software-defined networks. These solutions follow different approaches like machine learning-based, based on network monitoring, and trust-based, and employ a variety of techniques. A lot of

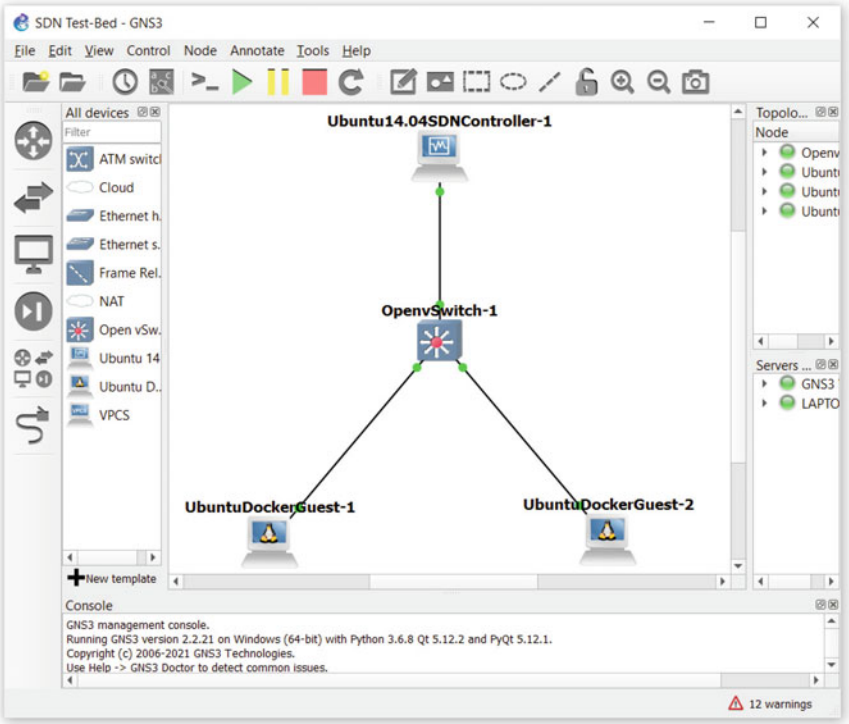


Fig. 3 SDN implementation in GNS3

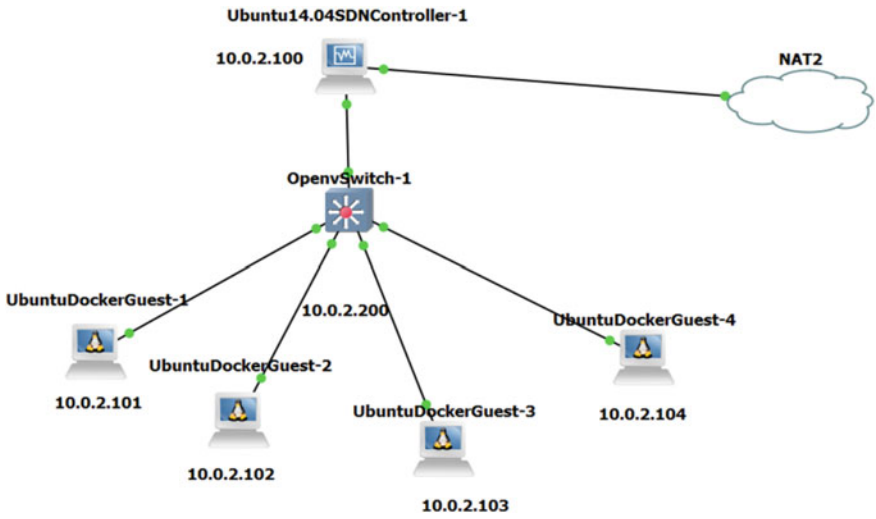


Fig. 4 SDN test-bed in GNS3

work has been done in implementing security in SDN based on trust. But, almost all of them have focused on only a single plane.

Also, the solution based on trust must consider the hybrid deployment of SDN, i.e., using SDN along with the existing network consisting of traditional, or legacy, network devices. A trust-based solution, where the trust values for the different entities, using different mechanisms, can be calculated, may present a novel and practical solution to identify security issues quickly.

The proposed trust-based approach using SDN monitoring can provide a simple way to identify the trustworthiness of any network entity in the SDN. Existing network monitoring tools and techniques can be used to monitor the behavior of different entities to generate a trust score for them which can be used by the security team. It can also be used with hybrid SDN deployments. Such a solution based on trust scores can also be used for designing policies for securing and optimizing the network behavior and also to identify security incidents quickly, which can then be investigated further.

The future work will look into implementing the proposed mechanism in the form of a framework in a multi-controller SDN network. Different network monitoring techniques will be used to identify the best possible set of parameters that can be used for evaluating the trust for network entities at each plane.

References

1. Software-defined networking: the new norm for networks. https://opennetworking.org/sdn-resources/whitepapers/software-defined-networking-the-new-norm-for-networks/. Last accessed 10 Mar 2021
2. Using HTTP methods for RESTful services. https://www.restapitutorial.com/lessons/methods.html. Last accessed 10 Mar 2021
3. Z. Yao, Z. Yan, A trust management framework for software-defined network applications. Concurr. Comput. Pract. Exp. **32**(16), e4518 (2018)
4. D. Kreutz, F.M.V. Ramos, P. Verissimo, Towards secure and dependable software-defined networks. in *Proceedings of the Second ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking (HotSDN 2013)*. IEEE, (ACM, New York, USA, 2013), pp. 55–60
5. C. Yoon, S. Lee, H. Kang, T. Park, S. Shin, V. Yegneswaran, P. Porras, G. Gu, Flow wars: systemizing the attack surface and defenses in software-defined networks. IEEE/ACM Trans. Netw. **25**(6), 3514–3530 (2017)
6. S. Scott-Hayward, S. Natarajan, S. Sezer, A survey of security in software defined networks. IEEE Commun. Surv. Tutor. **18**(1), 623–654 (2016)
7. A. Akhunzada, A. Gani, N.B. Anuar, A. Abdelaziz, M.K. Khan, A. Hayat, S.U. Khan, Secure and dependable software defined networks. J. Netw. Compu. Appl. Elsevier **61**, 199–221 (2016)
8. W. Li, W. Meng, L.F. Kwok, A survey on OpenFlow-based software defined networks: security challenges and countermeasures. J. Netw. Comput. Appl. Elsevier **68**, 126–139 (2016)
9. A. Shaghaghi, M.A. Kaafar, R. Buyya, S. Jha, Software-defined network (SDN) data plane security: issues, solutions, and future directions, in *Handbook of Computer Networks and Cyber Security*. ed. by B. Gupta, G. Perez, D. Agrawal, D. Gupta (Springer, Cham, 2020), pp. 341–387
10. O. Yurekten, M. Demirci, SDN-based cyber defense: a survey. Future Gener. Comput. Syst. Elsevier **115**, 126–149 (2021)

11. S. Vissicchio, L. Vanbever, O. Bonaventure, Opportunities and research challenges of hybrid software defined networks. ACM SIGCOMM Comput. Commun. Rev. **44**(2), 70–75 (2014)
12. S.Y. Sinha, K. Haribabu, A survey: hybrid SDN. J. Netw. Comput. Appl. Elsevier **100**, 35–55 (2017)
13. P.T. Duy, D.T.T. Hien, N. Van Vuong, N.N.H. Au, V.H. Pham, Toward a trust-based authentication framework of northbound interface in software defined networking. in *Industrial Networks and Intelligent Systems. INISCOM 2019*, ed. by T. Duong, N.S. Vo, L. Nguyen, Q.T. Vien, V.D. Nguyen. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 293 (Springer, Cham, 2019), pp. 269–282
14. A. Chowdhary, D. Huang, A. Alshamrani, M. Kang, A. Kim, A. Velazquez, TRUFL: distributed trust management framework in SDN. in *Proceedings of International Conference on Communications (ICC)*. IEEE, (Shanghai, China, 2019).
15. S. Pisharody, J. Natarajan, A. Chowdhary, A. Alshalan, D. Huang, Brew: a security policy analysis framework for distributed SDN-based cloud environments. IEEE Trans. Dependable Secure Comput. **16**(6), 1011–1025 (2019)
16. A.L. Aliyu, P. Bull, A. Abdallah, A trust management framework for network applications within an SDN environment. in *Proceedings of 2017 31st International Conference on Advanced Information Networking and Applications Workshops (WAINA)*. IEEE, (Taipei, Taiwan, 2017), pp. 93–98
17. A.L. Aliyu, A. Aneiba, M. Patwary, P. Bull, A trust management framework for software defined network (SDN) controller and network applications. Comput. Netw. Elsevier **181**, 107421 (2020)
18. S. Betgé-Brezetz, G. Kamga, M. Tazi, Trust support for SDN controllers and virtualized network applications. in *Proceedings of 2015 1st IEEE Conference on Network Softwarization (NetSoft)*. IEEE, (London, UK, 2015), pp. 1–5
19. C. Banse, S. Rangarajan, A secure northbound interface for SDN applications. in *Proceedings of IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE, (Helsinki, Finland, 2015), pp. 834–839
20. B. Isong, T. Kgogo, F. Lugayizi, B. Kankuzi, Trust establishment framework between SDN controller and applications. in *Proceedings of 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*. IEEE, (Kanazawa, Japan, 2017), pp. 101–107
21. S. Burikova, J. Lee, R. Hussain, I. Sharafitdinova, R. Dzheriev, F. Hussain, S. Sharieh, A. Ferworn, A trust management framework for software defined networks-based Internet of Things. in *Proceedings of IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. IEEE, (Vancouver, Canada, 2019), pp. 0325–0331
22. H. Cui, Z. Chen, L. Yu, K. Xie, Z. Xia, Authentication mechanism for network applications in SDN environments. in *Proceedings of the 20th International Symposium on Wireless Personal Multimedia Communications (WPMC2017)*. IEEE, (Bali, Indonesia, 2017), pp. 1–5
23. S. Scott-Hayward, C. Kane, S. Sezer, OperationCheckpoint: SDN application control. in *Proceedings of IEEE 22nd International Conference on Network Protocols*. IEEE, (Raleigh, USA, 2014), pp. 618–623
24. S. Shin, P. Porras, V. Yegneswaran, M. Fong, G. Gu, M. Tyson, FRESCO: modular composable security services for software-defined networks. in *Proceedings of ISOC 20th Annual Network and Distributed System Security Symposium (NDSS)*. IEEE, (San Diego, USA, 2013), pp. 1–16
25. P. Porras, S. Shin, V. Yegneswaran, M. Fong, M. Tyson, G. Gu, A security enforcement kernel for OpenFlow networks. in *Proceedings of the First Workshop on Hot Topics in Software Defined Networks (HotSDN'12)*. IEEE, (ACM, Helsinki, 2012), pp. 121–126
26. C.C. Lamb, G.L. Heileman, Towards robust trust in software defined networks. in *Proceedings of Global Communications Conference (GLOBECOM) Workshops*. IEEE, (Austin, USA, 2014), pp. 166–171
27. K. Sood, K.K. Karmakar, V. Varadharajan, U. Tupakula, S. Yu, Analysis of policy-based security management system in software-defined networks. IEEE Commun. Lett. **23**(4), 612–615 (2019)

28. B. Zhao, Y. Liu, X. Li, J. Li, J. Zou, TrustBlock: an adaptive trust evaluation of SDN network nodes based on double-layer blockchain. PLoS One **15**(3), e0228844 (2020)
29. K. Giotis, C. Argyropoulos, G. Androulidakis, D. Kalogeras, V. Maglaris, Combining OpenFlow and sFlow for an effective and scalable anomaly detection and mitigation mechanism on SDN environments. Comput. Netw. **62**, 122–136 (2014)
30. A. Zaalouk, R. Khondoker, R. Marx, K. Bayarou, Orchsec: an orchestrator-based architecture for enhancing network-security using network monitoring and SDN control functions. in *Proceedings of Network Operations and Management Symposium (NOMS)*. IEEE, (Krakow, Poland, 2014), pp. 1–9
31. C. Yu, C. Lumezanu, Y. Zhang, V. Singh, G. Jiang, H.V. Madhyastha, FlowSense: monitoring network utilization with zero measurement cost. in *Passive and Active Measurement. PAM 2013*, ed. by M. Roughan, R. Chang. Lecture Notes in Computer Science, vol. 7799 (Springer, Berlin, 2013), pp. 31–41
32. X.T. Phan, K. Fukuda, SDN-Mon: fine-grained traffic monitoring framework in software-defined networks. Journal of Information Processing **25**, 182–190 (2017)
33. L. Fawcett, S. Scott-Hayward, M. Broadbent, A. Wright, N. Race, Tennison: a distributed SDN framework for scalable network security. IEEE J. Sel. Areas Commun. **36**(12), 2805–2818 (2018)
34. P. Tsai, C. Tsai, C. Hsu, C. Yang, Network monitoring in software-defined networking: a review. IEEE Syst. J. **12**(4), 3958–3969 (2018)

Improving LoRaWAN Networks Performance Through Optimized Radio Resource Management

Husam Rajab, Xi Tiansheng, and Tibor Cinkler

Abstract The massive number of Internet of Things (IoT) applications is allowed by Low-Power Wide-Area Networks (LPWAN) with an extensive geographical area, low bit rate, and longevity specifications. The LoRa technology is a notable LPWAN that utilizes a proprietary Chirp Spread Spectrum (CSS) physical layer. Simultaneously, the LoRaWAN is defined as the media access control (MAC) protocol for large-scale deployment. It is intended to let low-powered applications to interact with massive requests that are connected through extensive wireless links. The aimed dense network distribution will inescapably lead to a scarcity of radio resources and have an issue with scalability when massive devices are connected to the network. In this paper, we propose a mathematical model to improve the link budget and scalability of LoRaWAN networks. Simulations using Matlab for different network scenarios illustrate the effectiveness of the optimization procedure.

Keywords IoT · LPWAN · Radio resource · LoRaWAN · MAC

1 Introduction

Recently, the Internet of Things (IoT) deployment is expected to compose a massive amount of data that will be used to optimize the consumption of each variety of resources and to enhance the performance of increasingly inter-connected systems. The distribution of the IoT devices pledges to contribute a large-scale, inexpensive, and long-lasting connectivity for each thing that will be helpful and serve for being connected. Today, the IoT is already integrated and leads the business models of various industries and services such as e-health, consumer electronics, automotive,

H. Rajab (✉) · T. Cinkler
Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary
e-mail: husamrajab@tmit.bme.hu

X. Tiansheng
Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, Budapest, Hungary

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_21

smart buildings, supply chains, facility management, connected cities, or manufacturing applications [1, 2]. Many communication machineries can be deployed to connect with IoT devices, e.g., limited-range communication technologies like NFC, ZigBee, Bluetooth, and Z-Wave. This communication diversity provides both decent and adequate communication range and high throughput to deal with the most applications connected to the network. Recently, for outdoor measurements, there was just one possible standard: cellular communication. 2G, 3G, 4G, and 5G can afford reliable long- range connections easily. However, these networks drain batteries of small devices cursorily. For example, battery replacements or charging are determined. Despite this, such technologies are biased to the crucial issues such as power consumption, cost, and design intricacy when the devices command to cover extensive geographical zones and emphasize massive units of applications because of the necessity of critical multi-hop routing inconceivably uncivil surroundings. New principles for Low-Power Wide-Area Networks (LPWAN) are an efficient procedure to overcome various circumstances and explain the mechanisms for distinct IoT applications that include vast ranges of distances [3]. The Low-Power Wide-Area Networks (LPWAN) promotes the soaring of the internet of things market for different applications by allowing long-range communications at a low bit rate amid linked objects, such as sensors operating on a battery. There are a di-verse number of existing standards in the LPWAN domain that compete to increase business share among massive IoT technologies, such as LoRa® (Long Range) [4–6], SigFox [7], NarrowBand IoT (NB-IoT) [8], Weightless [9], WAVIoT [10], Wi-Fi HaLow [11], Ultra-narrow band (UNB) LTE-M, and others. However, the IoT industrial applications are cost-effective, wide-range, and power-efficient sensors and actuators. Furthermore, LoRaWAN protrudes by optimizing LPWAN for the battery lifetime, coverage, and cost. LoRa can be the resolution for this problem because it affords long-distance wireless connectivity, low-power consumption, and comparatively low price; the restriction is the low data transfer capacity [12]. Moreover, the significant features of LoRa over other different LPWAN technologies are open-source MAC protocol LoRaWAN specification, community support, and low-cost application accessibility [13]. LoRaWAN has distinct singularity, including chirp spread spectrum modulation technique, which implements Pure ALOHA protocol, and regulatory constraints on the radio duty cycle. The structural design of LoRaWAN networks shown in Fig. 1 may be a formidable task, particularly as the network scales up rapidly numerous devices. The link layer, LoRa, has some explicit radio-related parameters that can be modified, such as Spreading Factor (SF), Carrier Frequency (CF), Transmission Power (TP), Bandwidth (BW), and Coding Rate (CR). The previous parameters are mentioned possibly adjusted at a device and the network stage to improve the network operation, explicitly decreasing power consumption, increasing radio scope, and decreasing radio collision and error rates. Moreover, the massive demanding and developing appropriation of LoRa in IoT requisitions, equipment's models, and methods convenient to maintain and optimize its operation remain limited and inadequate. Each device in LoRaWAN must evaluate radio resources and their strict energy and capacity restrictions. It is unlikely that they will frequently communicate with the Network Server (NS). Furthermore, the NS will not manage this strategy in large

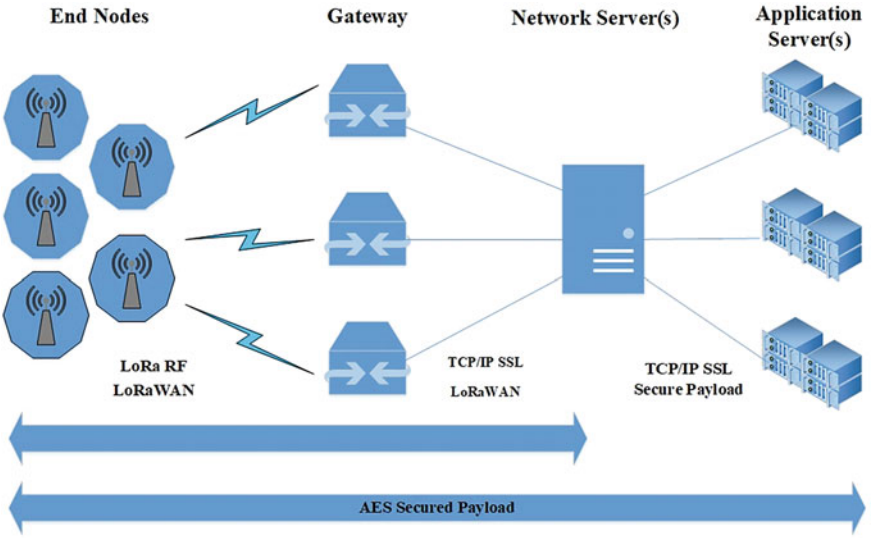


Fig. 1 The structural design of LoRaWAN networks

distribution due to severe limitations. The rest of the paper is organized as follows: Sect. 2 gives a brief overview of the most relevant works. In Sect. 3, we describe the problem statement. We develop our analytical model in Sect. 4, followed by simulation results. The conclusion is given in Sect. 5.

2 Related Work

The LoRaWAN network design faces critical circumstances in terms of network scalability and performance. These factors are fundamentally affected by the best preferable radio parameters and by environmental imperatives [14]. In this section, we summarize various researches' relevant works that defined a reasonable formal approach related to our proposed algorithms. By comparing the analysis, the performance and scalability of LoRaWAN networks, it is challenging to find parameter selection. Our community has made significant enrichment investigating the performance, link budget, resource management, and scalability for long-range networks. The authors in [15, 16] investigating the recently allocated learning for radio resource distribution in LoRaWAN had appealed to the Multi-Armed Bandit (MAB) dilemma. Every node is regarded as an intelligent delegate device to determine a transmitted value of SF and channel it to enhance the completion transmission rate [15] and the trade-off between energy efficiency and reliability by a sufficient remuneration process [16]. The quality of service of the LoRa is obtained by performing tests in [17] by producing and then monitoring the traffic between IoT end nodes and

LoRa IoT stations utilizing their proposed LoRa FABIAN protocol stack. In our recent research in [18], we present two machine learning algorithms for reducing the LoRaWAN devices power consumption. The two algorithms we deployed are the support vector regression and the deep neural network methods to prolong the battery life. The parameters configuration of the LoRaWAN device depends on the prediction of different metrics based on the user's demands. The author in [19] studies the functioning and indoor via obstacle diffusion of LoRa similar modification in the 2.4 GHz frequency band. However, the outcome of this research declares correspondence (because LoRaWAN functions mainly in sub-GHz bands), any of the observed effects. They performed a consequence that will be efficient also for decreasing the frequency bands. The work presented in [20] analyzes the efficiency of the LoRa concerning distinct settings of PHY and ecological circumstances. The investigation performed at the communication restricts the network flexible PHY structure and immense sending power to slower forms that increase the characteristic of the connection. That will lead to more agile PHY configurations that produces 100 times quicker sufficient bit rate than a more gradual precision. In our recent research in [21], we investigate the packet loss and collision scenarios. Moreover, we propose two algorithms to enhance the scalability of the network when the number of end nodes increases rapidly. The authors in [22] improve the Quality-of-Service (QoS) of LoRaWAN networks through a Mixed Integer Linear Programming (MILP) by fine-tuning specific radio parameters. Moreover, the authors determine the most optimal values for the Spreading Factor (SF) and Carrier Frequency (CF) radio parameters to decrease the packet interference rate and the energy consumption in LoRa networks through the channels traffic conditions. A brief survey of routing concerns in WSN simultaneously with an energy-efficient cluster-based proto col named LEACH with its vulnerabilities are presented in [23]. The authors also emphasize an enhanced version of LEACH protocol with its pros and cons compared to the Sec-LEACH protocol. Review research in [24] investigates the distinct environmental parameters like irrigation, monitoring, soil properties, and temperature for precision agriculture. Moreover, a comprehensive study is also performed on diverse crops which are covered using WSNs technologies. This study also highlights the different communication technologies and sensors available for Precision Agriculture (PA). The authors consider in [25] that a spectrum map is periodically determined by a network of sensors with cognitive radio capabilities. The LoRa network can employ a distinct set of frequencies in every geographical sub-zone. In addition, the result displays that the total throughput for the LoRaWAN network can be enhanced when the communication parameters in regards to frequencies and spreading factors are dynamically allocated.

3 Problem Definition

LoRa modulation technique empowers wide-range communication. As a matter of course, it raises the question of how a single LoRaWAN can serve diverse end devices

and gateways. LoRaWAN network assigns various standards of resources, materials, communication protocols, and encryption schemes to make a reliable wireless network in the extensive range with low-power consumption. Lora adopts CSS (chirp spread spectrum) modulation technology to achieve long- distance and low-power communication [26]. At the same time, it has the characteristics of anti-interference and anti-multipath. The main parameters of Lora communication are analyzed as follows: Spreading Factor (SF) illustrates the number of symbols sent per bit of information, takes a value from 7 to 12. Moreover, the higher the SF's value, the farther communication distance will be, and the higher the sensitivity is. Bandwidth (BW) refers to the spectrum occupied by each symbol. A larger BW provides a higher data rate, but a lower sensitivity. The coding rate is must be selected from one of four options: $\left[\frac{4}{5}, \frac{4}{6}, \frac{4}{7}, \text{ and } \frac{4}{8}\right]$. Every four relevant bits will be encoded by 5, 6, 7 or 8 transmission bits depending on its value. The lower the coding rate (the smallest is 4/8), the longer the time-on-air in transmission; therefore, sending a packet takes longer. As a result of the receiver's increased sensitivity, the user now has a higher link budget. On the other hand, slow data transmissions will influence battery consumption: increased time spent in transmission mode means more battery usage. The range of TP is between 4 and 20 dBm in steps of 1 dB. These specifications can be harmonized through the network design to improve the primary network performance, explicitly decreasing energy consumption, enhancing radio coverage, and minimizing radio interference and error rates. In addition, a duty cycle of 1% or 0, 1% is caused to restrictions on any end devices from transmitting information too often to make capacity for the other devices. Nonetheless, a significant concern linked to the LoRaWAN scalability is allocating the end devices across the scope of applications. That means many end devices, particularly those demanding higher throughput, have to be classified near the gateway. Also, in contrast to [23], we take into account packet loss caused by some random noise and collision in the unlicensed band with other LoRaWAN end device transmissions. In this work, we introduce a mathematical path loss model to analyze the influence of various parameters on the coverage area by adjusting the spread spectrum factor, bandwidth, and transmission power.

4 Math Model and Simulation Result

4.1 Mathematical Model

To enhance the LoRaWAN network performance as long as the optimization of radio parameters, we design the mathematical model of data transmission in a LoRaWAN network by extending our ideas published in [21]. Our performance problem for LoRaWAN specify values to the BW, SF, Pt parameters to enhance the link budget and manage the resources of the network. To assume the communication extent of LoRaWAN network, the equation path loss is given in the logarithmic domain as

follows [27]:

$$L_{PATH} = 10 * \log_{10}\left(\frac{4\pi df}{c}\right)^2 \tag{1}$$

where f represents the LoRa frequency, d is communication range, and c is the light speed. Accordingly, the link budget transfer path is represented as

$$L_{BUD} = \frac{P_T}{SR(SF, BW)} \tag{2}$$

where P_T and $SR(SF, BW)$ are transmitting power and receiver sensitivity, respectively, which relies on the spreading factor and bandwidth.

Thus, the receiving sensitivity is calculated as

$$SR(SF, BW) = 174 + 10 * \log10(BW) + NF + SNR \tag{3}$$

where NF and SNR are the receiver noise and Signal-to-Noise-Ratio, respectively. Furthermore, to calculate the Signal-to-Noise-Ratio (SNR):

$$SNR = \frac{E_B}{N_0} \tag{4}$$

$$E_B = P_R * T_{BIT} \tag{5}$$

where E_B is the energy per bit, N_0 is the noise power duration per bit, and P_R is the receiving power.

N_0 can be determined as

$$N_0 = NF * K * T \tag{6}$$

where K, T are the Kelvin constant.

Accordingly, the pair of operating frequency and spreading factor composes a logical channel. The symbol time (T_{sym}) of LoRa is given by the formula [29]:

$$T_{sym} = \frac{2^{SF}}{BW} s \tag{7}$$

SNR can be re-written as the following:

$$SNR = \frac{2^{SF} * P_R}{BW * NF * K * T} \tag{8}$$

Table 1 LoRa SX1272 characteristics

Transmission power (dBm)	Power consumption (mW)
20	412.5
17	297
13	92.4
7	95.4

Furthermore, to express the link budget is given as

$$L_{BUD} = \frac{P_T}{-174 + 10 * \log_{10}(BW) + NF + SNR} \tag{9}$$

Then, in order to get the evaluation of the maximum communication distance of the Lora link, we suppose that there is no antenna gain and set the path loss L_{PATH} to L_{BUD}, so it can be determined as

$$L_{PATH} = \frac{P_T}{-174 + 10 * \log_{10}(BW) + NF + \frac{2^{SF} * P_R}{BW * NF * K * T}} \tag{10}$$

Considering the communication range d

$$d = \left(10^{\frac{L_{PATH}}{10}}\right)^{1/2*c} \times 4\pi f \tag{11}$$

Table 1 shows various transmission power levels and corresponding current consumption that can be used with the LoRa SX1272 transceiver [28].

4.2 Simulation Results

End devices are geography disseminated in the system. Simulations are developed by Matlab utilizing diverse number values of SF (7, 8, 9, 10, 11, 12), and BW (125, 250, 500) kHz for each scenario. In the first scenario, we simulated with a slightly increased value of SF. Figure 2 illustrates the evaluation of increasing the range of transmission proportionally with the rising of SF values.

In the second and third scenarios, we suppose that the LoRa network includes one gateway and several end devices functioning in BW and SF's primary values. Thus, to further examine the performance improvement of our solution, Fig. 3 shows the obtained results in the second scenario by applying the proposed model using different BW values 120, 250, 500 kHz, respectively.

Furthermore, Fig. 4 shows the obtained results of the third scenario after comparing the transmitting power by applying the Eq. (11) concerning the spreading

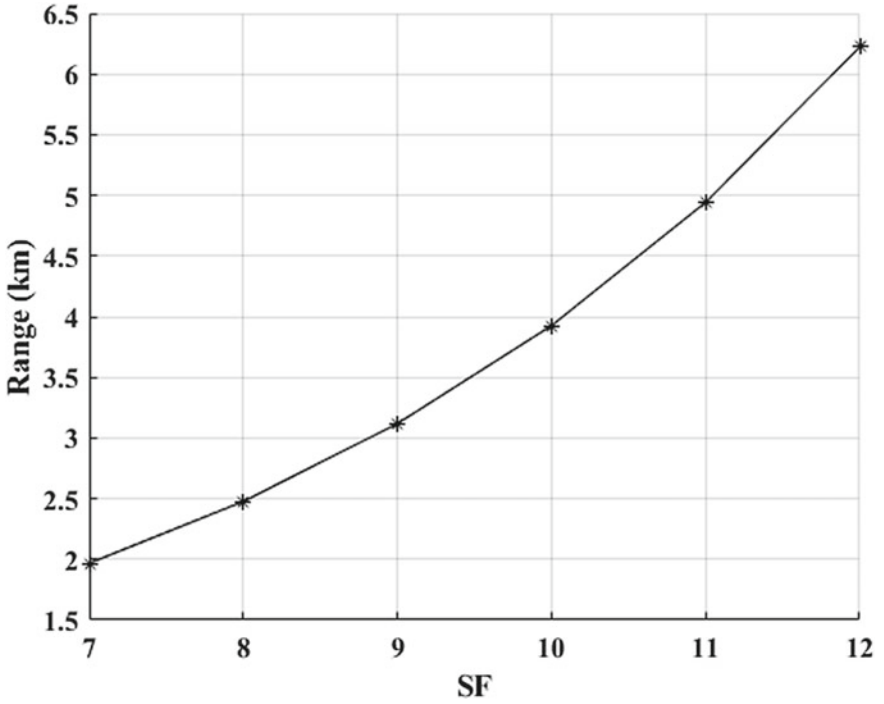


Fig. 2 The dependency of link budget on the comparison between the range and different spreading factors LoRa signals

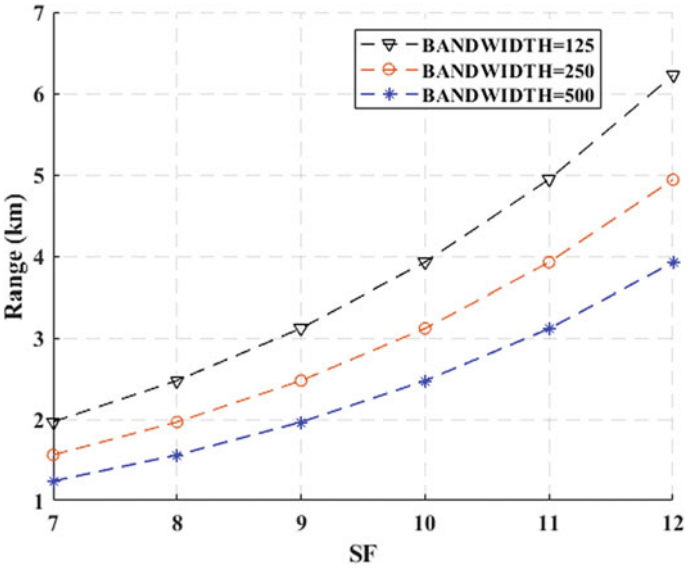


Fig. 3 Comparison of the range and BW values for LoRa signals with different spreading factors

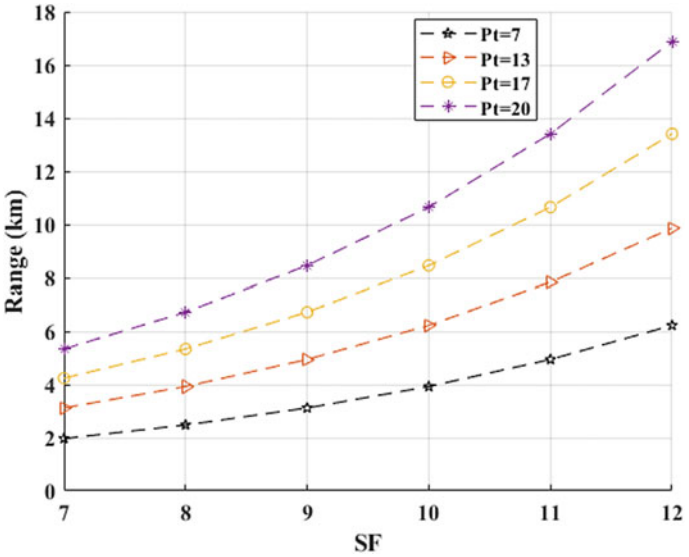


Fig. 4 Comparison of the range and transmission power values for LoRa signals with different spreading factors

factor values. Table 1 characterizes a comparison that provides more perceptions into the enhancement of the proposed distance spreading factor model.

5 Conclusion

The LPWAN is an impressive new technology, which has a prospective to enhance the wireless communication enabler for a diversity of IoT applications. In this paper, we mainly focused on one of the perspectives of LoRaWAN technologies, which have reached great interest in recent studies. In this paper, we analyzed and evaluated the resource allocation of LoRaWAN, and also measured the throughput of a LoRa link budget and improved the performance of LoRaWAN cell. We will develop the proposed optimization algorithm for future work to support various scenarios with extensive transmission range and many gateways, comparing and analyzing with other allocation methods. Moreover, to extend the optimization problems is to include the dynamic reconfiguration of the LoRa parameters. It may be interesting to analyze decentralized or distributed slicing control strategies to explain this challenge in extensive scale LoRa networks.

Acknowledgements This research is supported by Tempus Public Foundation, Stipendium Hungaricum Scholarship Programme and High Speed Networks Lab, Department of Telecommunications and Media Informatics, Budapest University of Technology. “This work was supported by the

Ericsson - BME 5G joint research and cooperation project, partly funded by the National Research, Development and Innovation Office, Hungary with project number 2018-1.3.1-VKE-2018-00005."

References

1. L. Atzori, A. Iera, G. Morabito, The internet of things: a survey. Comput. Netw. **54**(15), 2787–2805 (2010). https://doi.org/10.1016/j.comnet.2010.05.010
2. J.P. Bardyn, T. Melly, O. Seller, N. Sornin, IoT: The era of LPWAN is starting now. in *ESSCIRC Conference 2016: 42nd European Solid-State Circuits Conference 2016 Sep 12*. IEEE, (2016), pp. 25–30. https://doi.org/10.1109/ESSCIRC.2016.7598235
3. L. Angrisani, P. Arpaia, F. Bonavolonta, M. Conti, A. Liccardo, LoRa protocol performance assessment in critical noise conditions. in *2017 IEEE 3rd International Forum on Research and Technologies for Society and Industry (RTSI), Sep. 2017*. IEEE, (2017), pp. 1–5. https://doi.org/10.1109/RTSI.2017.8065952
4. Home page—LoRa Alliance®. https://lora-alliance.org/. Last accessed 03 May 2020
5. O. Georgiou, U. Raza, Low power wide area network analysis: can LoRa scale? IEEE Wirel. Commun. Lett. **6**(2), 162–165 (2017). https://doi.org/10.1109/LWC.2016.2647247
6. C. Goursaud, J.-M. Gorce, C. Goursaud, J.M. Gorce, Dedicated networks for IoT: PHY/MAC state of the art and challenges (2015). https://doi.org/10.4108/eai.26-10-2015.150597
7. Sigfox—the global communications service provider for the Internet of Things (IoT). https://www.sigfox.com/en. Last accessed 03 May 2020
8. L. Feltrin et al., Narrowband IoT: a survey on downlinkand uplink per- spectives. IEEE Wirel. Commun. **26**(1), 78–86 (2019). https://doi.org/10.1109/MWC.2019.1800020
9. Weightless Management Ltd. http://www.weightless.org/about/what-isweightless. Last accessed 03 May 2020
10. WAVIoT: low power wide area network for IoT and M2M. https://waviot.com/. Last accessed 03 May 2020
11. Wi-Fi HaLow—wi-fi alliance. https://www.wi-fi.org/discover-wi-fi/wi-fi-halow. Last accessed 03 May 2020
12. R.S. Sinha, Y. Wei, S.-H. Hwang, A survey on LPWA technology: LoRa and NB-IoT. ICT Express **3**(1), 14–21 (2017). https://doi.org/10.1016/j.icte.2017.03.004
13. LoRaWANTM 1.0.3 specification, (2018). https://lora-alliance.org/resource_hub/lorawan-specification-v1-0-3/. [Online]
14. U. Raza, P. Kulkarni, M. Sooriyabandara, Low power wide area networks: an overview. IEEE Commun. Surv. Tutor. **19**(2), 855–873 (2016). http://arxiv.org/abs/1606.07360. Last accessed 04 May 2020. [Online]
15. R. Bonnefoi, L. Besson, C. Moy, E. Kaufmann, J. Palicot, Multi-armed bandit learning in IoT networks: learning helps even in non-stationary set- tings. Lect. Notes Inst. Comput. Sci. Soc. Telecommun. Eng. LNICST **228**, 173–185 (2018). http://arxiv.org/abs/1807.00491. Last accessed 05 May 2020. [Online]
16. A. Azari, C. Cavdar, Self-organized low-power IoT networks: a distributed learning approach. in *2018 IEEE Global Communications Conference (GLOBECOM) 2018 Dec 9*. IEEE, (2018). https://doi.org/10.1109/GLOCOM.2018.8647894.
17. T. Petri´c, M. Goessens, L. Nuaymi, A. Pelov, L. Toutain, L. Toutain, Measurements, performance and analysis of LoRa FABIAN, a real-world implementation of LPWAN (2016). https://hal-imt.archives-ouvertes.fr/hal-01331966. Last accessed 06 May 2020. [Online]
18. H. Rajab, M. Ebrahim, T. Cinkler, Reducing power requirement of LPWA networks via machine learning. Pollack Periodica (2021). https://doi.org/10.1556/606.2020.00263
19. T. Wendt, F. Volk, E. Mackensen, A benchmark survey of long range (Lo- RaTM) spread-spectrum-communication at 2.45 GHz for safety applications. in *2015 IEEE 16th Annual*

Wireless and Microwave Technology Conference, WAMI- CON 2015. IEEE, (2015). https://doi.org/10.1109/WAMICON.2015.7120312
20. M. Cattani, C. A. Boano, K. Romer, An experimental evaluation of the reliability of lora long-range low-power wireless communication. J. Sens. Actuator Netw. **6**(2), (2017). https://doi.org/10.3390/jsan6020007
21. H. Rajab, T. Cinkler, T. Bouguera, IoT scheduling for higher throughput and lower transmission power. Wirel. Netw. (2020). https://doi.org/10.1007/s11276-020-02307-1
22. E. Sallum, N. Pereira, M. Alves, M. Santos, Improving quality-of-service in LoRa low-power wide-area networks through optimized radio resource management. J. Sens. Actuator Netw. **9**(1), (2020). https://doi.org/10.3390/jsan9010010
23. S. Tanwar, K. Thakkar, R. Thakor, P.K. Singh, M-tesla-based security assessment in wireless sensor network. Procedia Comput. Sci. **132**, 1154–1162 (2018). https://doi.org/10.1016/j.procs.2018.05.030
24. D. Thakur, Y. Kumar, A. Kumar et al., Applicability of wireless sensor networks in precision agriculture: a review. Wirel. Pers. Commun. **107**, 471–512 (2019). https://doi.org/10.1007/s11277-019-06285-2
25. G. El Helou, M. Ibrahim, R. El Tawil, Y. Mohanna, Radio resource optimization for spectrum-aware LoRa secondary system. in *2020 16th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob) (50308)*. IEEE, (2020), pp. 1–6. https://doi.org/10.1109/WiMob50308.2020.9253409
26. F. Adelantado, X. Vilajosana, P. Tuset-Peiro, B. Martinez, J. Melia-Segui, T. Watteyne, Understanding the Limits of LoRaWAN. IEEE Commun. Mag. **55**(9), 34–40 (2017). https://doi.org/10.1109/MCOM.2017.1600613
27. H. Linka, M. Rademacher, O. G. Aliu, K. Jonas, B. Rhein-sieg, S. Augustin, Path loss models for low-power wide-area networks: experimental results using LoRa. VDE ITG-Fachbericht Mobil kommunikation, no. May, 10–14 (2018)
28. Wireless and sensing products SX1272 development kit user guide. www.semtech.com. Last accessed 26 May 2020. [Online]
29. F. Mansourkiaie, L.S. Ismail, T.M. Elfouly, M.H. Ahmed, Maximizing lifetime in wireless sensor network for structural health monitoring with and without energy harvesting. IEEE Access **5**, 2383–2395 (2017). https://doi.org/10.1109/ACCESS.2017.2669020

On Security and Performance Requirements of Decentralized Resource Discovery in IoT

Mohammed B. M. Kamel, Peter Ligeti, and Christoph Reich

Abstract The resources in Internet of Things (IoT) are distributed among different physical geographic locations. In centralized resource discovery, the resources are registered in a centralized third-party server, and the clients can discover any resource by querying the centralized entity. In the decentralized resource discovery, the task of resource registration and discovery is distributed among many nodes in the system. Replacing a centralized entity with a distributed set of nodes requires that a system fulfils some security and performance requirements. In this paper, the centralized and decentralized resource discovery models are discussed. In addition, the properties of decentralized resource discovery are studied, and some of the fundamental and most important requirements for such models are discussed. Each of the fundamental requirements in decentralized resource discovery is analysed, and the possible approaches and their feasibility in IoT network are studied.

Keywords Resource discovery · Internet of things · P2P · Decentralized architecture

1 Introduction

The growing number of resources (sensors, service providers, etc.) in the Internet of Things (IoT) requires an efficient model of discovering those resources in the network. One possible solution is by applying the traditional client/server scheme, in

M. B. M. Kamel (✉) · P. Ligeti
Eotvos Lorand University, Budapest, Hungary
e-mail: mkamel@inf.elte.hu; mkamel@hs-furtwangen.de

P. Ligeti
e-mail: turul@cs.elte.hu

M. B. M. Kamel · C. Reich
Hochschule Furtwangen University, Furtwangen, Germany
e-mail: christoph.reich@hs-furtwangen.de

M. B. M. Kamel
University of Kufa, Najaf, Iraq

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_22

which all the resources are registered in a third-party entity and discovered through that entity. Registering all of the resources including their sensitive information such as location data to a third party entity might cause privacy and security issues. Additionally, the communication of IoT clients and resources that are on the edge of the network with the centralized entity (e.g. cloud) on the other edge of the network causes unpredictable latency [1]. In addition to the latency, the significant amount of used traffic for this process affects the system efficiency negatively. Due to these issues, a decentralized resource discovery can be a more efficient replacement to the centralized resource discovery model. However, there are some challenges and fundamental requirements that have to be addressed in such models to guarantee a secure and efficient resource discovery model.

The decentralized resource discovery [2] replaces the third party entity with multiple nodes in the network. To ensure an efficient model without a centralized entity that organizes the registration and discovery process, some requirements have to be fulfilled. As instance, if the centralized entity (e.g. cloud) has been replaced with distributed nodes (e.g. fog nodes), the model has to store the information related to the resources in the nodes that are close to the resources. This will reduce the latency that causes by sending the information to other edge of the network. Without this property, if the information of the resources have been stored in a far node in IoT, then the communication will be done with delay due to the long time it takes to reach that node. In this paper, we summarize some of the requirements to design a secure and efficient resource discovery model without relying on a trusted third party.

There are several researches that focus on studying different aspects and challenges in IoT. Rana et al. [3] performed a systematic survey focusing on the energy consumption and interoperability in different fields of IoT, including the resource management. In addition to interoperability, the authors have stated the energy-consuming tasks and components in IoT and studied the energy-efficient approaches and techniques that can be applied in different fields. Authors in [4] performed a survey to study various attacks and the used techniques for the detection of anomalies and attacks in IoT. They focused on machine learning and deep learning techniques in this field, and stated the number of use cases in IoT. Khail et al. [5] performed a survey to investigate the challenges in designing a resource discovery model for IoT and stated various proposed approaches. The main contributions of our paper are as follow:

- Discuss the security assumptions and various technologies in designing a decentralized resource discovery model.
- Investigate fundamental security and performance requirements in decentralized resource discovery models.
- Review the proposed solutions and techniques to address the discussed requirements.
- Identify open challenges and provide future recommendations.

The rest of this paper is organized as follows. The next section provides some security definitions and assumptions and summarizes the centralized and decentralized resource discovery models focusing on Distributed Hash Table (DHT) tech-

nology [6]. Section 3 discusses some of the requirements for decentralized resource discovery model and various proposed solutions. Finally, Section 4 presents our conclusions and the directions for future works.

2 Background

2.1 Security Definitions

Definition 1 (Access structure) Let $P = \{P_1, P_2, ...P_n\}$ be a set of n parties. An access structure is a collection Γ of non-empty subsets of P, i.e. $\Gamma \subseteq 2^{\{P_1, P_2, ...P_n\}}$. Any $U \in \Gamma$ is called authorized set, and any $X \notin \Gamma$ is called unauthorized set. We say Γ is monotonic access structure if for any M, N, if $M \in \Gamma$ and $M \subseteq N$, then $N \in \Gamma$.

Definition 2 (Bilinear pairing) Let G and G_1 be two cyclic groups of prime order p and q, respectively. Let G be an additive group of points of an elliptic curve over a finite field with order n and G_1 be multiplicative group of order n, and let $g \in G$ be a generator. A pairing is a map $e : G \times G \rightarrow G_1$, which satisfies the following properties:

a. Non-degeneracy: $\exists g \in G, e(g, g) \neq 1$;
b. Bilinearity: $\forall x, y \in Z_p, \forall g, h \in G, e(g^x, h^y) = e(g, h)^{xy}$;
c. Computability: $\forall g, h \in G$, there exists an efficient algorithm to compute $e(g, h)$.

In many cases, the security of pairing-based protocols is related to the following variants of the Diffie-Hellman problems:

Definition 3 (Decisional Diffie-Hellman (DDH) assumption) Consider a cyclic group G of order p, a random generator g and random $a, b, c \in Z_p^*$, the (g^a, g^b, g^{ab}) and (g^a, g^b, g^c) can be distinguished in probabilistic polynomial time with negligible probability only.

Definition 4 (Decisional Bilinear Diffie-Hellman (DBDH) assumption) Consider a cyclic group G of order p, a random generator g and random $a, b, c, d \in Z_p^*$, the tuples $(g^a, g^b, g^c, e(g, g)^{abc})$ and $(g^a, g^b, g^c, e(g, g)^d)$ can be distinguished in probabilistic polynomial time with negligible probability only.

2.2 Distributed Hash Table

DHT as a distributed storage technology is used to create a logical overlay of nodes on top of the underlay network. It is used in a number of decentralized resource

discovery models [2, 7] as an alternative to the trusted third party in the centralized resource discovery models. There are a number of implementations such as Pastry [8], Chord [9], Kademila [10] and Tapestry [11] that can be used to create a DHT overlay in a given environment. There is no centralized organizing entity that controls the joining and leaving processes of the nodes. Any node in DHT can be joined freely the created DHT overlay by using a collision-resistant one-way hash function that is used in DHT for both node allocation and data responsibility definition. Upon joining the DHT overlay, a new node uses the hash function to create its own seemingly unique identifier. This is done by feeding the hash function with some unique information such as the IP address of the node, and using the output of the function as the identifier of the node. The generated identifier indicates the logical position of the newly joined node in the overlay.

As in hash tables [12], the DHT is used to store and then retrieve the data in a distributed scheme. The stored data in the DHT consist of a pair of key/value parameters. The key parameter is generated by feeding an attribute of the stored data such as its name or type to the hash function. The value parameter includes the necessary information, such as its URI, to find the actual location of the required data. The same hash function can be used for both key generation in the key/value pair and the identifier generation of the participating nodes. This makes the mapping between the stored pair and the nodes works by matching the keys with the identifiers of the nodes in the overlay. If no exact identifier during matching process has been found, a closeness function will be used to determine the identifier of a close node in the DHT overlay that is responsible for storing the pair of key/value. Later, a similar process will be used to retrieve the stored value of a given key. It is noteworthy to mention that for n participating nodes in the overlay, it takes at most $O(log(n))$ steps to find a node that is responsible for storing a specific data based on the key parameter of the stored pairs. This feature makes the DHT scalable and efficient when it implements in large-scale networks.

2.3 Resource Discovery Models

IoT consists of a huge number of various resources such as devices or data. These resources are located in different geographic locations and can be accessed through their addresses. The clients in the system are the consumers of the IoT resources and discover the resources in order to get the address data of the required resources. The resource discovery [13] is a process of providing some information about the required resources by the consumers of the resources and getting the address data and further links of those required resources. The IoT resources have to be already registered in the system in order to be discoverable by the clients.

There are two main categories of resource discovery models. The first category depends on a centralized authority or a set of centralized authorities for registration and discovery, while the second category distributes the registration and discovery process among different participating nodes in the system. The discovery request

is issued by a client that determines the attributes of the required resources. The attributes can be their identifiers, their locations, their types and so on. The discovery request will be processed by the discovery system through the registration authority and the address data of the required resources will be returned back to the requested client.

3 Fundamental Requirements for Decentralized Resource Discovery

In this section, some of the fundamental and important requirements for decentralized resource discovery models will be described. Table 1 summarizes the discussed fundamental requirements. The requirements have been divided into two main categories: security requirements and performance requirements. The main discussed security requirements of a decentralized resource discovery model in IoT are discoverability, availability, authentication and privacy. The performance requirements have been divided into essential and secondary requirements. The discussed essential requirements are responsibility definition, management, visibility of resources and scalability. Considering the huge number of IoT devices and absence of a centralized organizing entity, these essential performance requirements are crucial in a decentralized resource discovery model. Some of the secondary requirements are multi-attributes discovery, range queries and location aware. While a decentralized resource discovery model might work without these features, but they are required to improve the performance of a proposed model. In the following sections, these requirements and some proposed solutions are described.

3.1 Responsibility Definition

In a centralized model of resource discovery, the responsibility of registering the resources is performed by a centralized trusted entity. Later on, to discover any resource, the query can be sent to the trusted entity or set of centralized trusted entities for discovering the required resource. Since in a decentralized resource discovery model, there is no single trusted entity that is responsible for registering and discovering the resources in the network, the task of registering and discovering the resources are distributed among many nodes in the system and the responsibility of each of the nodes has to be clearly defined. This means that by implementing the decentralized resource discovery model, each node in the distributed environment has the knowledge of the subset of resources that have to be registered by the node itself. On the other hand, the clients have the knowledge about the exact node that is responsible to store the access data of each subset of resources and direct the discovery request to the specific node based on the required resources.

Table 1 Important requirements of decentralized resource discovery models

Category	Requirement	Definition
Security	Discoverability	Resources can only be discovered by intended clients
	Availability	Resources remain discoverable even in case of partial failure of some nodes
	Authentication	Sender and request authentication during registration and discovery
	Privacy	Privacy of the data of resources and clients during registration/discovery
Performance (essential)	Responsibility	Define the responsibility of each of the nodes in the decentralized system
	Management	Provide interfaces to manage the resources in the system such as add, remove, search and update
	Visibility of resources	Ability of access all resources including those behind a NAT or firewall
	Scalability	Ability to adjust the registration and discovery processes as the system grows in term of number of nodes
Performance (secondary)	Multi-attributes	Resource discovery based on multi-attributes of the requested resources
	Range queries	In addition to exact value resource discovery, it can be done based on a given range of values
	Location aware	The nodes are positioned based on their physical locations, and the registration and discovery processes are location aware

By implementing DHT (Sect. 2.2) in resource discovery model, an overlay of nodes can be created. There are research models that used a single DHT layer [1, 2, 7, 14, 15] or a hierarchy of layers [16, 17] to create the structured overlay. Each node in the distributed system can be assigned with a d-bit binary identifier [2]. The responsibility of each of the nodes is defined as follows: Let $\mathcal{I}_d$ be a set of all possible sequence of d-bit binary digit (i.e. identifiers) and each node w in the distributed system has an identifier $id_w \in \mathcal{I}_d$ and each resource u has a parameter $tag_u \in \mathcal{I}_d$. Then the node w that is responsible for storing the information to access the resource u is defined as in Eq. 1:

$$res(u) = w, s.t.\ tag_u \approx id_w. \tag{1}$$

Each resource u is linked to a node w in the distributed system that its identifier $id_w \in \mathcal{I}_d$ is close or equal to its parameter tag_u, denoted with $id_w \approx tag_u$.

3.2 Discoverability Range

The resources in the system can be public or private resources. A public resource might be a temperature sensor mounted on a public building or a computing server that provides a public computing service. A private resource might be a camera in a private building or a printer in an organization. The resources (whether public or private) have to be discoverable by a specific or predefined set of clients in the network. As instance, a printer in an organization has to be discoverable only by the employees belong to that organization. The situation is the same with a private camera that has to be discoverable only by a predefined set of owners.

One possible solution is to create a model that divides the resources into two distinct layers, private and public layers [2]. The resources in the private layer should be discoverable by a subset of clients. In this model, the access address associated with any private resource is encrypted and it can only be discovered and accessed if the requested client has the relevant keys to decrypt the stored information. The information of private resources can be encrypted using symmetric encryption. The private resources send the shared keys to a predefined set of clients. These set of clients that have the shared keys are the only ones that are able to discover those private resources by decrypting the address data of the private resources using the preshared keys. The advantage of this model is that the symmetric encryption is used to encrypt the information of resources during the process of resource registration and discovery, which can be executed in the resource-constrained IoT devices. However, the disadvantage of this approach is that it requires the private resources to preshare the symmetric keys with the set of trusted clients. In addition, this model implements a binary level discoverability policy and there is no way to allow the clients to be able to discover the resources based on their attributes. The clients in this model are divided into two distinct groups, either having the key and are able to retrieve the address data of that private resource, or not.

Resource discovery models can introduce multi-layers of discoverability [7] instead of binary level model, utilizing the Attribute-Based Encryption (ABE) [18] to authorize the clients to discover the resources considering the attributes of each client. In traditional asymmetric encryption algorithms, the plaintext is encrypted using one of the keys in the key pair and the other key in the key pair is used to decrypt the ciphertext, i.e. the decryptor is the one that has access to the second key in the key pair. In contrast, the decryptor in ABE is a subset of clients based on an access structure (Definition 1) by introducing the concept of attributes. It uses the combinations of attributes of the clients as the public key to encrypt the address data of the resources during registration, while the private keys are calculated and assigned to the clients by the attribute authority based on the attributes of the client. Standing on the bilinear pairing techniques (Definition 2) and based on DBDH assumption (Definition 4), the ABE builds the various access structures to achieve fine-grained access control of data. In large-scaled distributed environments such as IoT network, the traditional ABE schemes with a single attribute authority might add a significant overhead as all registered resources and discovering clients have to access that

authority for key generation. In order to reduce the workload and attack risk of single authority, techniques such as multi-authority ABE [19] can be used. The required attributes of the clients in this model are defined by the authorized entities during the process of registering the resources in the system. A client proves its attributes to one of the attributes authorities and asks for the corresponding decryption key. A public resource can ask for zero attributes. In such case, all clients are able to discover it regardless of their attributes.

This model allows the resources to be registered with a predefined set of attributes that defines the access control policy, and the clients only by their inherent attributes are able to discover the resources in the network. Although this model replaces the binary level discoverability by multi-level and attribute-based discoverability, but due to the computationally heavy requirements of ABE, it can not be directly implemented in some of the IoT resource with limited computation power, and the trust has to be migrated to other members of the system.

3.3 System Availability

The availability of the resources in a decentralized resource discovery model can be attacked by letting the responsible nodes being inaccessible. This prevents the authorized entities to have access to the stored information and discover the required resources. In addition, some of the nodes can be temporarily being offline. Without taking this issue into consideration, the availability of the system can be failed during discovery process.

Replication can be used to ensure the availability of registered resources in the system [20]. In this case, the Eq. 1 can be replaced by mapping M

$$M : \mathcal{U} \rightarrow \mathcal{W}$$

in which $\mathcal{U}$ and $\mathcal{W}$ are the set of all resources and all nodes in the the distributed system, respectively.

$$M(u) = \{w : tag_u \approx id_w, \nexists w' \mid dst(id_{w'}, tag_u) < dst(id_w, tag_u)\}. \quad (2)$$

The dst can be any closeness function to compute the distances between any two nodes in the system. The cardinality of $M(u)$ depends on the replication factor rp parameter. The rp indicates the number of close nodes to w in the system that are responsible for storing the information of the previously mentioned resource u. It is chosen such that the members in any given subset of the set that includes all the nodes in the distributed system with cardinality rp are unlikely to fail (i.e. being inaccessible or offline) [2].

3.4 Authentication

In a decentralized resource discovery model, there are a number of control messages that manage the different processes in the system (Sect. 3.6). Adding, searching, removing or updating any of the resources in the system have to be done only after authenticating the requesting entity. This can be done through the use of various authentication protocols [21]. Although symmetric cryptography can be used for authentication, but it can not be implemented in large-scale scenarios as the number of required keys to be stored in each entity increases, and without proper integration of other protocols it might prevent the scalability of the system (Sect. 3.8). The digital signature based on asymmetric cryptography can be used to authenticate the entities in the system. Public Key Infrastructure (PKI) [22] creates the trust in the distributed public keys used during the authentication process. Since the traditional PKI requires heavy computation power that might not be available by all entities in the IoT network, the light-weight PKI [23] can be implemented in IoT.

During the design of the authentication protocol for IoT, the trust anchors and security assumptions of the different components of the system and their communications have to be defined. The key factors in authentication protocols in IoT that include resource-constrained devices is the added communication overhead and the required computational power to perform the authentication. Considering some of the IoT devices that are not able to perform some of the public key primitives due to heavy computational power and their resource-limited capability or the IoT devices that are energy constrained, adopting some of the authentication protocols that require heavy computation might not fit the IoT requirements.

3.5 Privacy

The privacy of the entities in the IoT network is one of the critical requirements that has to be guaranteed. A resource in the network is registered by storing its information including its address and properties in a specific node in the system. On the other hand, a client (i.e. issuer of a discovery process) provides some information about itself and the required resources during the discovery process. If the resource discovery model works based on the attributes of the clients, then the client has to provide its attributes during the discovery process. In addition to that, it has to provide the attributes of the required resources that might include information such as the location or the type of the resources. The resources and the clients in the resource discovery model should have control on the potential entities that might access their information during the registration and discovery process. Therefore, a resource discovery model has to reveal no information to the unauthorized entities in the network. This task is more challenging considering distributed responsibility of resource registration and discovery among many nodes in the network.

3.6 Management

The resources in IoT network are managed by an IoT entity. The IoT entity is any device in the IoT network with proper computational power that handles a resource and manages the resources that are connected to it. The resource discovery model has to provide defined interfaces for the authorized IoT entities to be able to add, update or delete resources in the network. Adding a resource will be done by registering it in a particular node in the IoT network. Later on, the authorized entities should be able to update the information related to specific resources through an interface in the resource discovery model. Updating can be done to address information, attributes of the resource or any related information. In addition to that, the resource discovery model has to allow the authorized entities to remove resources from the network. In the resource discovery models that utilize DHT technology, the messages provided by the DHT such as *put(key, value)* can be used for adding purposes. Other management functionalities can be modified or developed based on the specifications of the resource discovery model.

3.7 Visibility of Resources

During the resource discovery, discovering and accessing all resources in the network including those behind the Network Address Translator (NAT) are crucial. During discovery and access, if the NAT or firewall receives packets from a node in the network, and there is no NAT binding associated with the destination IP address and port, the packets will simply be dropped by the NAT. Using protocols such as PCP [24] allows the resources to control and open ports. In PCP proposal, a rendezvous server is used to distribute the addresses among other nodes. To remove the centralized rendezvous server, Address propagation [25], Distributed Address Table (DAT) [26] or blockchain [27] can be integrated as part of the resource discovery model to ensure the visibility of the resources by authorized clients.

3.8 Scalability

The scalability in a decentralized resource discovery describes its ability to adapt the different processes in the system including registration and discovery as the number of participating nodes grows. Resource discovery models that remove the centralized organizing entity and distribute the responsibility among many nodes in the system should continue to work without any degradation in their efficiency as the number of participating nodes increases. Structured p2p technologies such as DHT or Content-Addressable Network (CAN) [28] can be used to create an overlay that is self-configured and scalable without involving a centralized organizing entity.

3.9 Multi-attributes Discovery

The resources in IoT have their attributes that describe their different features such as their provided service, their locations, their identifiers and so on. In order to be able to discover the accurate required resources, more than one attribute are used during the discovery process. As an example, a client might be interested in discovering the available resources that offer a specific service in a specific location. In this case, if the resource discovery model does not support the multi-attributes discovery, the requesting entities will not be able to get the accurate resources based on the intended request. Multiple DHT queries [2] or profiling [15] can be used to add this feature to the model.

3.10 Range Queries

One of the important requirements of a resource discovery model is its support of range queries instead of an exact match for the properties of resources during the discovery process. The resource discovery can be based on a system that supports queries specifying lower and upper boundaries on a single or multiple attributes [29]. Hence, the resource discovery model that supports the range queries can handle a scale query enables the client applications to identify information of resources that have not been identified yet. Prefix Hash Tree (PHT) [30] can be used to support the range queries.

3.11 Location Aware

Considering the physical locations of nodes in the underlay network and mapping that on the overlay network helps improving the efficiency during resource registration and discovery. If the physical location in the underlay does not match the logical location in the overlay, two physically close nodes may be positioned to distant locations in the overlay and vice versa. Since the actual physical path of those nodes does not match their logical distance value an efficiency issue might be raised, which leads to operational inefficiency in applications running over it [31].

In the resource discovery model, the physical locations of nodes need to be considered during different processes. As instance, to reduce the delay during the registration, a resource has to be registered in a physically close node. The same applies during resource discovery process. Region-based DHT [32] creates multiple regions and register the resources in a node belongs to the same region considering the physical location of both resources and peers in the system [20]. Distributed Geographical Table (DGT) [33] can be used to build a layer to distribute the information

of resources depending on the physical locations of participating entities, and this information can be used to discover the required resources based on their physical locations [34].

4 Conclusions

In this paper, some of the fundamental and important requirements of a decentralized resource discovery model have been discussed. The requirements can be divided into two main categories: security requirements and performance requirements. The security requirements are those requirements that are related to the security and privacy of the system and its members. The performance requirements can be divided into essential and secondary requirements. The essential performance requirements are those that have to be included in the model, otherwise there could be a partial or overall failure in the system. The secondary performance requirements are important but failing to achieve them will only decrease the performance of the model. In the security requirements, the discoverability range ensures that each of the resources can be discovered only by the authorized clients. In addition to that, to ensure the availability of the system, a set of nodes can be defined to store a replica of the information of each of the registered resource in the system. The transmitted control messages between entities including the registration and discovery requests have to be authenticated. The privacy ensures that during the registration and discovery processes, no more than the allowed information about each of the members of the system are revealed.

The essential performance requirements are the responsibility definition of each of the nodes in the system, management, visibility of resources and scalability. In the decentralized resource discovery model, the scope of the responsibility of each of the nodes in the model has to be clearly defined. It should provide interfaces that allow the authorized entities to add, update and remove the resources in the system. The visibility of resources ensures that any discovered resource in the system can be reached at any given time, even if it is behind a NAT or firewall. The decentralized resource discovery model has to be scalable without involving a centralized organizing entity. In the secondary performance requirements, multi-attributes and range queries ensure a precise result of discovered resources to be returned to the clients. The location-aware resource registration and discovery ensures that the address data of each of the registered resources in the system are stored and discovered using a close node in the system. Although we have divided the requirements into two main categories, but they are related to each other. As instance, the management requirement can be discussed in term of security, or the management can affect other requirements such as scalability.

Regarding the direction of this research, in the future work, more requirements can be discussed, and the effect of each of the discussed requirements on each other and their relationships can be studied. Despite the extensive research on decentralized resource discovery models, some challenges need further research. As instance,

addressing all of the fundamental performance requirements in one model adds extra overhead that can negatively affect the system efficiency, comparing to centralized resource discovery models. On the other hand, and during addressing the security requirements, the limited computation power and the distributed nature of the IoT network need to be considered. Lack of centralized entity and inability of adopting cryptographic protocols that require heavy computations add more challenges in proposing an efficient, secure and privacy-preserving decentralized resource discovery model.

Acknowledgements The research was supported by Application Domain Specific Highly Reliable IT Solutions project, which has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the Thematic Excellence Programme TKP2020-NKA-06 (National Challenges Subprogramme) funding scheme, ÚNKP-20-3 New National Excellence Program of the Ministry for Innovation and Technology from the source of National Research, Development and Innovation Fund, SH programme and by the Ministry of Science, Research and the Arts Baden-Württemberg Germany.

References

1. B. Zhang, N. Mor, J. Kolb, D.S. Chan, K. Lutz, E. Allman, J. Wawrzynek, E. Lee, J. Kubiatowicz, The cloud is not enough: saving iot from the cloud, in *7th Workshop on Hot Topics in Cloud Computing (HotCloud 15)* (2015)
2. M.B. Kamel, B. Crispo, P. Ligeti, A decentralized and scalable model for resource discovery in iot network, in *2019 International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)* (IEEE, 2019), pp. 1–4
3. B. Rana, Y. Singh, P.K. Singh, A systematic survey on internet of things: energy efficiency and interoperability perspective. Trans. Emerg. Telecommun. Technol. (2020)
4. P. Malhotra, Y. Singh, P. Anand, D.K. Bangotra, P.K. Singh, W.-C. Hong, Internet of things: evolution, concerns and security challenges. Sensors **21**(5), 1809 (2021)
5. K. Khalil, K. Elgazzar, M. Seliem, M. Bayoumi, Resource discovery techniques in the internet of things: a review. Intern. Things 100293 (2020)
6. H. Zhang, Y. Wen, H. Xie, N. Yu, A survey on distributed hash table (dht): theory, platforms, and applications (2013)
7. M.B. Kamel, Y. Yan, P. Ligeti, C. Reich, A decentralized resource discovery using attribute based encryption for internet of things, in *2020 4th Cyber Security in Networking Conference (CSNet)* (IEEE, 2020), pp. 1–3
8. A. Rowstron, P. Druschel, Pastry: scalable, decentralized object location, and routing for large-scale peer-to-peer systems, in *IFIP/ACM International Conference on Distributed Systems Platforms and Open Distributed Processing* (Springer, 2001), pp. 329–350
9. I. Stoica, R. Morris, D. Karger, M.F. Kaashoek, H. Balakrishnan, Chord: a scalable peer-to-peer lookup service for internet applications. ACM SIGCOMM Comput. Commun. Rev. **31**(4), 149–160 (2001)
10. P. Maymounkov, D. Mazieres, Kademlia: A peer-to-peer information system based on the xor metric, in *International Workshop on Peer-to-Peer Systems* (Springer, 2002), pp. 53–65
11. B.Y. Zhao, L. Huang, J. Stribling, S.C. Rhea, A.D. Joseph, J.D. Kubiatowicz, Tapestry: a resilient global-scale overlay for service deployment. IEEE J. Sel. Areas Commun. **22**(1), 41–53 (2004)
12. W.D. Maurer, T.G. Lewis, Hash table methods. ACM Comput. Surv. (CSUR) **7**(1), 5–19 (1975)

13. S.K. Datta, R.P.F. Da Costa, C. Bonnet, Resource discovery in internet of things: current trends and future standardization aspects, in *2nd World Forum on Internet of Things (WF-IoT)* (IEEE, 2015), pp. 542–547
14. H. Wirtz, T. Heer, R. Hummen, K. Wehrle, Mesh-dht: a locality-based distributed look-up structure for wireless mesh networks, in *2012 IEEE International Conference on Communications (ICC)* (2012), pp. 653–658
15. S. Ali, S. Banerjea, M. Pandey, N. Tyagi, Wireless fog-mesh: a communication and computation infrastructure for iot based smart environments," in *Mobile, Secure, and Programmable Networking* (Springer International Publishing, Cham, 2019), pp. 322–338
16. H. Wirtz, T. Heer, M. Serror, K. Wehrle, Dht-based localized service discovery in wireless mesh networks, in *2012 IEEE 9th International Conference on Mobile Ad-Hoc and Sensor Systems (MASS 2012)* (IEEE, 2012), pp. 19–28
17. R. Mokadem, A. Hameurlain, A.M. Tjoa, "Resource discovery service while minimizing maintenance overhead in hierarchical dht systems. Int. J. Adapt. Resil. Autonom. Syst. (IJARAS) **3**(2), 1–17 (2012)
18. Y. Yan, M.B. Kamel, P. Ligeti, Attribute-based encryption in cloud computing environment, in *2020 International Conference on Computing, Electronics and Communications Engineering (iCCECE)* (IEEE, 2020), pp. 63–68
19. M. Chase, Multi-authority attribute based encryption, in *Theory of Cryptography Conference* (Springer, 2007), pp. 515–534
20. M.B.M. Kamel, P. Ligeti, C. Reich, Lamred: Location aware and privacy preserving multi layer resource discovery for iot. Acta Cybern. (2021)
21. M. El-Hajj, A. Fadlallah, M. Chamoun, A. Serhrouchni, A survey of internet of things (iot) authentication schemes. Sensors **19**(5), 1141 (2019)
22. J. Weise, Public key infrastructure overview, in *Sun BluePrints OnLine, August* (2001), pp. 1–27
23. N. Hong, A security framework for the internet of things based on public key infrastructure, in *Advanced Materials Research*, vol. 671 (Trans Tech Publications, 2013), pp. 3223–3226
24. D. Wing, S. Cheshire, M. Boucadair, R. Penno, P. Selkirk, Port control protocol (pcp), in *Internet Eng. Task Force, Fremont, CA, USA, RFC*, vol. 6887 (2013)
25. M.B.M. Kamel, P. Ligeti, A. Nagy, Improved approach of address propagation for f2f networks, in *Proceedings of the 2nd European Conference on Electrical Engineering and Computer Science* (IEEE, 2018), pp. 604–609
26. M.B.M. Kamel, P. Ligeti, A. Nagy, C. Reich, "Distributed address table (dat): a decentralized model for end-to-end communication in iot. J. Peer-to-Peer Netw. Appl. (2021)
27. E.F. Kfoury, J. Gomez, J. Crichigno, E. Bou-Harb, D. Khoury, Decentralized distribution of pcp mappings over blockchain for end-to-end secure direct communications. IEEE Access **7**, 110159–110173 (2019)
28. S. Ratnasamy, P. Francis, M. Handley, R. Karp, S. Shenker, A scalable content-addressable network, in *Proceedings of the 2001 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications* (2001), pp. 161–172
29. F. Paganelli, D. Parlanti, A dht-based discovery service for the internet of things. J. Comput. Netw. Commun. **2012** (2012)
30. S. Ramabhadran, S. Ratnasamy, J.M. Hellerstein, S. Shenker, Prefix hash tree: an indexing data structure over distributed hash tables, in *Proceedings of the 23rd ACM Symposium on Principles of Distributed Computing*, vol. 37 (2004)
31. E.K. Lua, J. Crowcroft, M. Pias, R. Sharma, S. Lim, A survey and comparison of peer-to-peer overlay network schemes. IEEE Commun. Surv. Tutor. **7**(2), 72–93 (2005)
32. M.B.M. Kamel, P. Ligeti, C. Reich, Region-based distributed hash table for fog computing infrastructure, in *13th Joint Conference on Mathematics and Informatics* (2020), pp. 82–83
33. M. Picone, M. Amoretti, F. Zanichelli, Geokad: a p2p distributed localization protocol, in *2010 8th IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)* (IEEE, 2010), pp. 800–803
34. S. Cirani, L. Davoli, G. Ferrari, R. Léone, P. Medagliani, M. Picone, L. Veltri, A scalable and self-configuring architecture for service discovery in the internet of things. IEEE Intern. Things J. **1**(5), 508–521 (2014)

EV Technology Trends & Placement of Electric Vehicle Charging Station: A Review

Tripti Kunj and Kirti Pal

Abstract Electric vehicles (EVs) have uncovered great enthusiasm from people in communal, the administration, and the foundation in current years. The pervasiveness of EVs is subsidized by their excellent ideal circum-positions diverging from ordinary inner-combustion vehicles including well-being, economy, and cleanness. EVs use power to control the vehicle, in this manner they are progressively steady and there would be no aroma of petroleum derivatives. The expense of the power is lower than that of petroleum derivatives, in this way EV is progressively efficient. In addition, petroleum derivatives are restricted assets and it takes excessively long for them to renew. In spite of that, power is increasingly ecological generous. The utilization of petroleum derivatives by vehicles has caused genuine air contamination, particularly in urban areas. With previously mentioned preferences, society has made a great exertion to investigate and develop EVs in the earlier years. However, with the high penetration of EVs load on the grid increases, and a problem arises for placement and sizing of the charging station. In this paper, various methods adopted for best placement and sizing of EVs are discussed in detail.

Keywords Electric Vehicle · Placement and sizing · Optimal placement

1 Introduction

Particularly in the transportation area, requests about electric vehicles are viably in progress as a push to diminish air toxins and EV's are acknowledged to transform into a significant means of transport by 2030. If EV supplant alive inner burning motors, substantially more power resolve be charging the EV, which will influence the Power System (PS). To manage this developing supply need, it's important to figure and dissect the charge example of electric vehicles (EV). Author utilizes Bayesian assumption strategies with complexity to estimate the 24-h charging arrangement & the outcome is compare with linear regression analysis (LRA) [1–3].

T. Kunj (✉) · K. Pal
Gautam Buddha University, Greater Noida, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_23

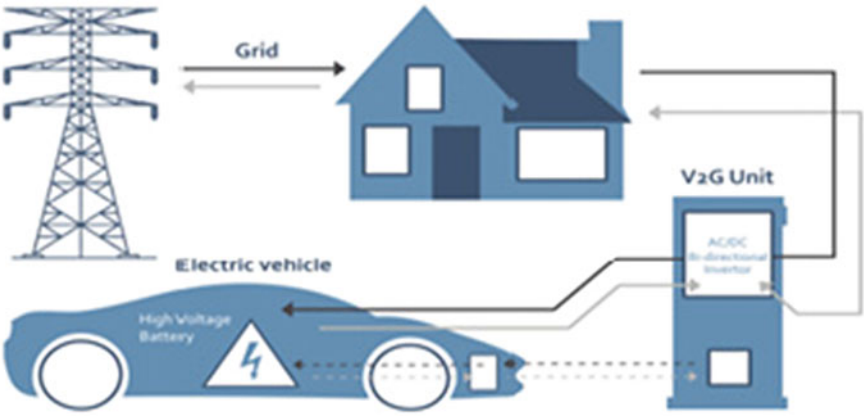


Fig. 1 EVs Interact with a Smart Power Electricity Network—Vehicle to Grid (V2G)

The matter of multi-organize site of EV charging stations of slow EV penetration rates are also discussed by author [1]. A nested multinomial logit model is utilized to break down the charging priority of the purchaser (EV buyer) and anticipate the accumulated charging demand at the charging stations. Through a progression of investigations utilizing the geographic what' more, statistic information from the city of San Pedro Area of Los Angeles, here show that the charging station arrangement is exceptionally reliable with the heat map of the traffic stream [4, 5].

Owing to the restricted electricity of batteries vehicle driver may experience uneasiness for extensive charging waiting time during their ride. In Fig. 1 EV charging management is discussed by the author. The author also discusses about to diminish drivers tour period and charging performance at the EV and CS sides [6].

In the paper [7] author shows a novel strategy for Local Government Authorities (LGAs) to characterize proper situations for the underlying arrangement of private on-road PEV (Plug-in Electric Vehicle) energizing set in urban zones to help expel boundaries to PEV endorsement. The strategy is real for use by LGAs with constrained monetary assets as it depends on basic Geographic Information System (GIS) examination of routinely accessible enumeration and stopping information to distinguish the spatial covers between zones where occupants are destined to be PEV clients and territories where there is a high dependence on private on-road places. The procedure has been actualized practically speaking to oversee a charging foundation establishment strategy for Southampton, UK, where 128 avenues (out of 1924 altogether) were proposed as proper areas [7, 8].

Here the author discusses about the collaborations of quatern developing technique for faster acceptance of EVs which is public flexibility, radio charging, vehicle-to-grid (V2G), and vehicle mechanization that have substantial abilities to improve keen and supportable built-up flexibility, and eventually expand built-up sustainability [9]. Here the author presents a social model of the public, revealed preferences for an

assortment of electric vehicles (EVs) while representing an idle (green) travel pattern erect and charging framework ascribes [10].

V-Charge suggests a explanation for charging self-sustaining EVs and effectively the use of scarce charging sources, therefore simplifying the lifestyles of the client and cumulative the viability of EVs. For the administration of parking masses and charging assets, V-Charge delivers an attendant again and a verbal exchange structure the paper only studied similar CSs, i.e., all CSs in the scheme had the similar conditions. In future researcher find out additional exciting awareness is to study the influence of varied CSs on the general performance [11–13].

Here the author optimally assigning source for public charging organization. The provisioning model is communicated as a numerical program with adjusting constraints and settled by an active set algorithm [14]. Partial least squares (PLS) way model methodology is used for Technique to arrange preferences by closeness to perfect result (TOPSIS) prototypical and bunch model, a sweeping urban status appraisal of charging establishment is finished ward on the preliminary investigation of 25 show metropolitan communities in China. The outcomes show that, a few urban areas are deed better in the four zones of strategy backing and venture, charging station structure and arranging, activity extension besides impact, activity administration and refuge. There are 6 pointers of the amount of charging heaps, amount of total indicting periods, vehicle types in administration, amount of combined charge for power, amount of preparing individuals and absolute venture that impact on city readiness [15].

The author presented a problem of smart charging for electric vehicle by computing routes that minimizing the cost and journey position, charge instance and the use of energy. The author resolved the problem with the help of evolutionary genetic algorithm. The conclusion shows that, proposed methods for electric vehicle have an optimal way out in a realistic moments & electric vehicle assigned to the smart charging station (CS) [16–18].

Here the author developed a static model of smart charging plug-in system of electric vehicle for public use, also reduce the price of electricity, destination and plug-in route choice of compact traffic area. This allocation model solved through active-set algorithm. This model is based on the mathematical program in equilibrium state [19].

2 Sizing and Placement of Charging Station

In the paper author decided FCS (Fast Charging Station) position, charger disposition at FCSs, and demand-to-FCS allocation subject to a reasonable price so as to exploit the demand coverage basically used the stochastic nature in the actual sum and timing of requests for service and the general service time at CS [20].

Here the author proposed a method for ideal areas of energizing stations are discovering, a flow refuelling Location model (FRLM) system is utilized. It manages areas for energizing stations to augment the stream that can go among source and

goal sets by refuelling. FRLM is extended by joining a probabilistic impression of as far as possible which differ contingent upon street conditions. Here build up a mixed integer nonlinear programming structure and propose a Benders-and-Price calculation by joining the Benders deterioration and section age to determine the arranged plan [21, 22].

Author [23] clarify that associating additionally charging stations in reasonable areas can discharge EV drivers extend uneasiness. To discover the areas of open charging stations, here recommended two optimization models for two various charging modes—quick and moderate charging, the objective is decreasing the absolute expense while supporting certain inclusion objective. Rather than utilizing discrete focuses here utilizing geometric articles to imply charging requests. Noticeably, to diligence the partial coverage problem (PCP) for systems stretching out the polygon overlay technique to part the requests out and about system [23, 24].

Here the author talked about a model dependent on Trip Success Ratio (TSR) so as to improve CS comfort for PEV drivers. Assortment of custom and distinctive driving practices are thought of, just as differing trip types (In-city, Highway). The administration scope of charging stations has been evaluated utilizing TSR with consideration of the vulnerability of excursion separations and the vulnerability in the Remaining Electric Range (RER) of PEVs and CS distribution stage [25].

Here the author suggests an enhanced numerical replica for detecting EVCS. Here consider the consecutive actions of the travellers, i.e., a human being with two main rests throughout the day should indict the vehicle in just solitary parking lots, therefore evading dual including the supply [26].

In the paper, the author proposed the principal assessment model which is Optimized Clustering with Statistical Based Local Model (OCSLM) that ponders the event recurrence, repeat, and length of technique inherent dormant events, similarly as system characteristics and relating essentialness data by consolidating these limits as opportunities for electric vehicle empowering inside an endless charging station distribution model [22].

Properties of Charging Station Placement (CSPL) CSPL issue are dissected and a calculation known as OCEAN is projected and process the ideal designation CS. Here a novel calculation is proposed to accelerate OCEAN which is called OCEAN-C [27].

Here the author introduced a method to evaluate the Load model for EV Fast charging station (EVFCS) considering an estimation-based burden displaying approach. Estimations of current and dynamic force are acquired from an EVFCS associated with IEEE-13 bus testing system, seeing diverse working conditions. These estimations are utilized to change the load model, limiting the bungle between the real load and the model reactions [28]. Ideal area arranging of EV Charging Station in Existing Distribution Network with steadiness conditions [29]. A Review on Latest EV Charging Technologies is discussed in this paper very beautifully [30].

3 The Impact on EV's Integration on Power Grid

Many authors explain different issues incorporate illustration of abundance supply from the grid, unbalance among request and provide in utility grid, electrical energy divergence in the grid, and so on. The author shows a review of the fast-charging stations for Electric Vehicle (EV) and their effects on leaving usefulness network alongside the answer for conquer this type of issue [31, 32]. Here the author produces a charge plan dependent on the accessible time allotments and load necessities to limit request vulnerability on matrix [33]. In the paper author offers a load management resolution for planning the charging of different PEVs in a smart grid framework. Unmanaged and easy-going PEV charging can develop power misfortunes, over-burdens and voltage varieties, which will influence the frameworks. In this way, a real time smart load management (RT-SLM) approach is proposed. The methodology lessens age cost by fusing time-differing market vitality costs and PEV proprietor favoured charging time regions dependent on need determination. An itemized brilliant matrix test plot is proposed and sets up the impacts and helps of proposed RT-SLM versus arbitrary charging. The proposed framework is a difference in the IEEE- 31 bus 23 kV distribution test framework aggregate with a few private LV 415 V systems based on genuine framework information of an area (in Western Australia). A sum of 22 LV feeders is proposed and provided from the HV main buses by means of 23 kV/415 V, 100 KVA distribution transformers (DT) [34].

In [35] the author discussed the harmful brunt of EVCS lots on the supply allocation system can't be deserted. Far above the ground charging lots of the speedy Charging position (SCP) outcomes in augmented high consignment order, condensed aloofness limits, electrical energy variability and consistency exertion.

A few authors portray a coordinated control arrangement produced for an EV battery grid edge. This paper plan EV load controllability, yet additionally its ability to give capacity to the matrix. EV control approaches will be improved so as to extend EV potential to offer extra types of assistance to the grid [36, 37].

Due to the EV, development in the interest for power to charge these vehicles could present generous difficulties to the electrical grid as far as extra burden due to unmanaged charge plans. So as to facilitate these issues, the alleging of the electrical vehicle's necessity adapted. The paper offerings improvement of a system intending to progressively regulator the charging of EVs to keep the best possible activity of the nearby appropriation operation and decrease the ecological impact [38, 39].

4 Adopted Technologies

4.1 Soft Computing Techniques

Additional hierarchical optimization method (HOM) is formed by the author [40], for distribution of DGs, CS, and protective devices. The best arrangement is researched dependent on the proposed HOM technique.

Here the author discusses about the use of Artificial intelligence (AI) to diminish the EV's Cost due and range anxiety. EVs prerequisite to be planned in such a way as to stay away from peak on the network that can affect in more energy costs and overwork confined allocation grids [41].

A multi-objective purpose is made and the most excellent explanation is intended via GA-PSO for cluster analysis of EV charging demand. Easy to contrivance; more suitable for placement problems but the main problem is taking extensive time to solve the location tricky [42].

In the paper author discusses about optimum allotment and sizing of Renewable Energy Sources (RES) and EVCS done concurrently with the help of GA-PSO. A multi-objective role is proposed to reduce the power sufferers, voltage variations, power delivering costs and car battery maintenance prices. The vagueness of input constraints is measured in the load flow which could also aid the supply network (DN) organizers to compose the most advantageous result [43]. In the paper author proposed the site problem for reducing the total social cost with the use of GA. An assessment record model is assembled dependent on terrestrial expense, building costs, street traffic flow, power network conditions, and power grid conditions by utilizing grey incidence and grey objective hypothesis, in which less information needed for the best arrangement [44].

4.2 V2G Technology

V2G is classified into two sorts Unidirectional V2G technology and Bidirectional V2G technology. In unidirectional V2G technology, single direction correspondence exists between a network administrator and EV to manage the charging pace of EV which can forestall over-burdening, system flimsiness, and voltage drop issues. In bidirectional V2G technology, an EV battery is seen as an energy stockpiling resource instead of simply a heap. This empowers a two-way stream of energy into the lattice as well as the other way around. This gives more noteworthy adaptability to the power grid framework as well as builds its unwavering quality and maintainability. Other than V2G two other lattice associated advancements are likewise developing in particular vehicle to home(V2H) and vehicle to vehicle(V2V). V2H gives a reinforcement to home electric apparatuses and home environmentally friendly power sources.

In [45] an advancement model is talked about to boost the benefit for the profit of grid operator. A detailed risk analysis for grid operator study is given in [46, 47].

4.3 Modelling and Simulation of City

In the paper author discussed about a spatial Markov chain model for typical charging load of EVs in metropolises. Geospatial maps were utilized to appraise the charging profile or combination of profiles of the charging stations dependent on the close by building types [48].

In the paper a stochastic go-ahead route interruption resupplying site problem with column delay for electric taxi CS, these structures are important to analyses battery-powered electric shared-ride taxi working in a linked, federal letter way. The perfect and explanation technique are built on a bi-level, imitation–optimization network that collects an upper-level multiple-server sharing perfect with queuing stay [49, 50].

The author presented regarding charging station for electric vehicle in smart cities area. The author reviewed about challenge arises in charging station and battery depletion. To overcome this problem through design model of a system. The author focused on reduces the frequency of wait time and trip time of electric vehicles with the help of traffic simulator [51].

5 Objective Functions and Constraints of Charging Station Placement Problem

CS placement is a multipart optimization issue worried about placement and estimating of CS. A portion of the basic objective functions and limitations of this optimization issue is recorded in Table 1 constraints like Current limit, Voltage limit, charging theme limit, and budget are soft constraints. And, Power edge, Charging request balance are hard constraints.

6 EV Technology Trends for Future Charging Station

Charging technology

One intriguing pattern to watch with regards to the EV market will be upgrades in charging innovation. According to EV buyers, they are right now constrained in their ability principally because of their range-per-charge and charging time. In the event that straightforwardly contrasted with the internal combustion engine (ICE) vehicle market, EVs are not exactly as helpful for significant distance travel, despite the fact

Table 1 List of objective functions and constraints of Charging Station Placement Problem

Objective functions	Constraints
Rate	Current limit
EV flow	Voltage limit
Power loss	Active power edge
Voltage stability	Reactive power edge
Reliability	No. of CS placed at a particular position
Waiting time	Number of chargers placed at the charging stations
Accessibility index	Charging Request Balance
Charging time	Power stability equation
Total Harmonic Distortion	Budget

that they are considerably more ecological inviting. It just isn't generally advantageous to need to revive each battery in 2–3 hours and stick around 30 min to get back out and about for the vast majority. To conquer this issue, EV producers should step up the accessibility of fast charging stations in the world over. A few organizations, similar to Tesla, have just made some incredible work with their worldwide rollout of supercharging stations. They likewise uncovered a versatile charging station not long ago. Porsche additionally reported a year ago that they have built up a supercharging arrangement that chops the run after an ideal opportunity to 15 min. Other vehicle makers are additionally taking action accordingly as well.

Solar-powered car

Another interesting technology pattern to envision will resolve around sunlight based controlled vehicles. Many manufacturers are as of now progressing in the direction of making sun-oriented vehicles a reality very soon. By permitting the vehicle's battery to be charged progressing, this could be a flat-out distinct advantage for the EV market. There are now some fascinating models as of now experiencing testing and 2020 could see some of them closer to production. Others are endeavoring to create sunlight based controlled hybrids too. Whenever accomplished, some estimate that the worldwide market for sunlight based controlled EVs could be the request for $1 Billion out of 2020. In any case, similarly all things considered with each new innovation, any models near the creation are inconceivably costly. One model, The Light year One, could interfere with an astounding $107,000 a unit. The key contests of deceitful such leading-edge vehicles, battery range, electrical schemes, solar array strategy, physical resources [52] (Fig. 2).

Electric planes

Various different organizations are as of now creating smaller electric airplanes that ought to be accessible in the upcoming years, as indicated by their assessments. An organization, the Israeli firm Aviation uncovered a model of a fully electric airplane.

Fig. 2 Solar powered electric vehicle

They guarantee that it will be the first on the world and have to the option to put across 9 travelers for 1,000 km. It is predictable to be on the rampage in 2022 [53]. Eviation has set 2019 as a take-off date to start testing completely electric flights and 2022 as when standard customers can purchase a ticket for an all-electric provincial worker flight. Uber's electric taxi eVTOL is supposed to start demoing in Dallas and Los Angeles by 2020 and be economically accessible by 2023. Taking a gander at this all the more realistically, a genuine electric passenger flight flying loaded with passengers from New York to London is 20 years off. What's more, that is being generous. As MagniX's Ganzarski said, "so many needs to occur" to get to that point. In the event that these flights work out, they could be a distinct advantage, especially for the climate. Up to that point, it's an ideal opportunity to work on carbon impression here on the ground. The gas accepting vehicle probably won't produce as much as that plane outing, yet it's still a ton however the significant test is the heaviness of the plane. New topologies of the power electronics might be necessary to have a substantial weight effect in this space [54] (Fig. 3).

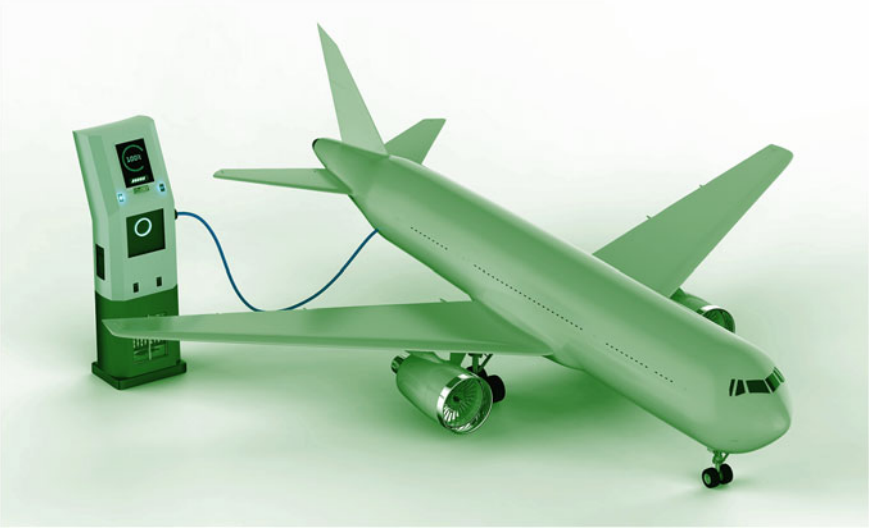


Fig. 3 Electric planes design for future

7 Conclusion

This paper presents a comprehensive review of literature on optimal EVCS arrangement and estimating issues and different trends identified with it. Different improvement strategies that handled the EVCS situation and measuring issues are illustrated and basically examined with their advantages and inconveniences. This work will give the most applicable and critical data about the existing studies. Here also discuss the various impacts of EVs integration on grid & there are various adopted technology presented, V2G technology and in future solar powered charging station will be installed and some researcher design the electric planes also.

References

1. E. Sortomme, Combined bidding of regulation and spinning reserves for unidirectional vehicle-to-grid. IEEE PES ISGT, Innovative Smart Grid Technologies, pp. 16–20 (2012)
2. M.A. Fasugba, P.T. Krein, Cost benefits and vehicle-to-grid regulation services of unidirectional charging of electric vehicles. Proceedings of the IEEE ECCE 2011: Energy Conversion Congress and Exposition, Sep 17–22. pp. 827–34 (2011)
3. J. Gallardo-Lozano, M.I. Milanés-Montero, M.A. Guerrero-Martínez, E.R. Cadaval, Electric vehicle battery charger for smart grids. Electr. Power Syst. Res **90**, 18–29 (2012)
4. K. Verma, K.B. Singh, D.T. Shahani, Grid to vehicle and vehicle to grid energy transfer using single-phase bidirectional AC–DC converter and bidirectional DC–DC converter, Proceedings of the International Conference on Energy, Automation and Signal, ICEAS, Dec 28–30. pp.1–5 (2011)
5. J.G. Pinto, V. Monteiro, H. Goncalves, B. Exposto, D. Pedrosa, C. Couto, et al., Bidirectional battery charger with grid-to-vehicle, vehicle-to-grid and vehicle-to-home technologies, in Proceedings of the IEEE IECON: 39th Annual Conference of the Industrial Electronics Society, Nov 10–13. pp. 5934–39 (2013)
6. C. Liu, K.T. Chau, D. Wu, S. Gao, Opportunities and challenges of vehicle-to-home, vehicle-to-vehicle, and vehicle-to-grid technologies. Proc. IEEE **101**(11), 2409–27 (2013)
7. R. Pode, Battery charging stations for home lighting in Mekong region countries. Renew. Sustain Energy Rev. **44**, 543–560 (2015)
8. N. Rotering, M. Ilic, Optimal charge control of plug-in hybrid electric vehicles in deregulated electricity markets. IEEE Trans. Power Syst. **26**(3), 1021–1029 (2011)
9. A. Foley, B. Tyther, P. Calnanet, Impacts of electric vehicle charging Electricity market operations. Appl. Energy **101**, 93–102 (2013)
10. M.R. Sarker, Y.D. Vorkin, M.A. Ortega-Vazquez, Optimal participation of an electric vehicle aggregator in day-ahead energy and reserve markets. IEEE Trans Power Syst. **31**(5), 3506–3515 (2012)
11. https://www.smartgrid.gov/the_smart_grid/plugin_electric_vehicles.html
12. C. Luo, V. Gupta, Placements of EV charging station-Balancing benefits among multiple entities (2018)
13. C. Yue, W. Tong, K. Omprakash, M. Geyong, An EV charging management system concerning drivers trip duration and mobility uncertainty. IEEE Trans. Syst. Man. Cybern. Syst. **51**, 2168–2216 (2016)
14. Da. Han, L.M.S. Kim, J.H. Roh, J.P. Yang, J.B. Park, Forecasting of Electric Vehicles Charging Pattern Using Bayesians method with the Convolution. IFAC, pp. 413–418 (2019)
15. J. Gil-Aguirre, S. Perez-Londono, J. Mora-Florez, A measurement-based load modelling methodology for electric vehicle fast charging stations. Electr. Power Syst. Res. (2019)

16. M. Grotea, J. Prestona, T. Cherretta, N. Tuck, Locating residential on-street electric vehicle charging infrastructure: a practical methodology. Trans. Res. **74**, 15–27 (2019)
17. M. Taiebat, M. Xu, Synergies of four emerging technologies for accelerated adoption of electric vehicles: shared mobility, wireless charging, vehicle-to-grid, and vehicle automation. J. Cleaner Prod. 794–797 (2019)
18. F. Nazaria, A.K. Mohammadian, T. Stephens, Modeling electric vehicle adoption considering a latent travel pattern construct and charging infrastructure. Trans. Res. **72**, 65–82 (2019)
19. K. Seddig, P. Jochem, W. Fichtner, Two-stage stochastic optimization for cost-minimal charging of electric vehicles at public charging stations with photovoltaics. Appl. Energy **242**, 769–781 (2019)
20. J. Timpner, L. Wolf, Design and evaluation of charging station scheduling strategies for electric vehicles. IEEE Trans. Intell. Trans. Syst. **15** (2014)
21. J. Tan, W.H. Lin, A stochastic flow capturing location and allocation model for siting electric vehicle charging stations. IEEE 17th International Conference on Intelligent Transportation Systems (2014)
22. F. Wua, D. He, Y. Yin, Y. Guan, Optimal deployment of public charging stations for plug-in hybrid electric vehicles. Trans. Res. **47**, 87–101 (2013)
23. Y. Xiong, J. Gan, B. An, C. Miao, A.L.C. Bazzan, Optimal Electric Vehicle Charging Station Placement. Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (2015)
24. N. Wang, Y. Liu, G. Fu, City readiness system assessment of electric vehicle charging infrastructure. 2014 IEEE 17th International Conference on Intelligent Transportation Systems (2014)
25. R.J. Rei, F.J. Soares, P.M.R. Almeida, J.A. Lopes, Grid interactive charging control for plug-in electric vehicles. 13th International IEEE Annual Conference on Intelligent Transportation Systems Madeira Island (2010)
26. F. Alesiani, N. Maslekar, Optimization of Charging Stops for Fleet of Electric Vehicles: A Genetic Approach
27. Z. Li, A. Khajepour, J. Song, A comprehensive review of the key technologies for pure electric vehicles. Energy **182**, 824–839 (2019)
28. T. Kunj, K. Pal, Optimal location planning of EV charging station in existing distribution network with stability condition. 7th International IEEE Conference on Signal Processing and Integrated Networks. ASET, Amity University, Noida, pp. 1060–1065 (2020)
29. V. Chauhan, T. Kunj, K. Pal, A Review on latest EV charging technologies. Int. J. Adv. Sci. Technol. **29**(7), 9903–9910 (2020)
30. K. Hayakawa, H.G. Enrico, G.S. Stein, T. Shiga, Online mechanisms for charging electric vehicles in settings with varying marginal electricity costs. Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (2015)
31. A. Hess, F. Malandrino, B. Moritz, R.C. Casetti, K.A. Hummel, J.M.B. Ordinas, Optimal Deployment of Charging Stations for Electric Vehicular Networks
32. Y.J. Jang, Survey of the operation and system study on wireless charging electric vehicle systems. Trans. Res. **95**, 844–866 (2018)
33. C. Bordina, A. Tomasgard, SMACS MODEL a stochastic multihorizon approach for charging sites management, operations, design, and expansion under limited capacity conditions. J. Energy Storage **26** (2019)
34. A.A. Alhazmi, A.M. Haytham, M.A. Magdy, Optimal allocation for electric vehicle charging stations using Trip Success Ratio. Electr. Power Energy Syst. **91**, 101–116 (2017)
35. J. Cavadas, G.H.D. Almeida, J. Gouveia, A MIP model for locating slow-charging stations for electric vehicles in urban areas accounting for driver tours. Trans. Res. **75**, pp.188–201 (2015)
36. P. Fekete, Lim, S. Martin, K. Kuhn, N. Wright, Improved energy supply for non-road electric vehicles by occasional charging station location modelling. Energy, pp.1033–1040 (2016)
37. K. Huang, P. Kanaroglou, X. Zhang The design of electric vehicle charging network. Trans. Res. **49**, 1–17 (2016)

38. J. Jung, Y.J. Joseph, R. Jayakrishnan, J.Y. Park, Stochastic dynamic itinerary interception refueling location problem with queue delay for electric taxi charging stations. Trans. Res. **40**, 123–142 (2014)
39. C. Lee, J. Han, Benders-and-Price approach for electric vehicle charging station location problem under probabilistic travel range. Trans. Res., 1–23(2017)
40. T. Kunj, K. Pal, Impact of EVs in distribution network.IEEE-COER-2019, International conference on Artificial Intelligence and Applications, College of Engineering Roorkee (COER), 20–21 November (2019)
41. M.R. Mozafar, M.H. Moradi, M.H. Amini, A Simultaneous Approach for Optimal Allocation of Renewable Energy Sources and Charging Stations based on Improved GA-PSO Algorithm
42. X. Ren, H. Zhang, R. Hu, Y. Qiu, Location of electric vehicle charging stations: a perspective using the grey decision-making model. Energy **173**, 548–553 (2019)
43. M. Shepero, J. Munkhammar, Spatial Markov chain model for electric vehicle charging in cities using geographical information system (GIS) data. Appl. Energy **231**, 1089–1099 (2018)
44. W. Yang, A user-choice model for locating congested fast charging stations. Trans. Res. **110**, 189–213 (2018)
45. M.R. Khalid, M.S. Alam, A. Sarwar, M.S.J.A. Asghar, A Comprehensive review on electric vehicles charging infrastructures and their impacts on power-quality of the utility grid. E-transportation (2019)
46. R. Faria, P. Moura, J. Delgado, A.T.D. Almeida, Managing the charging of electrical vehicles. Impacts on the Electrical Grid and on the Environment (2010)
47. A.M. Hariri, M.M. Hejazi, H.H. Dezaki, Reliability optimization of smart grid based on optimal allocation of protective devices, distributed energy resources, and electric vehicle plug-in hybrid electric vehicle charging stations. J. Power Sources **436** (2019)
48. E.S. Rigas, S.D. Ramchurn, N. Bassiliades, Managing electric vehicles in the smart grid using artificial intelligence: a survey. IEEE Trans. Intell. Trans. Syst. (2014)
49. S. Deilami, A.S. Masoum, P.S. Moses, M.A.S. Masoum, Real time coordination of plug-in electric vehicle charging in smart grids to minimize power losses and improve voltage profile. IEEE Trans. Smart Grid **2**(3) (2011)
50. S.M. Kandil, E.Z. Farag, M.F. Shaaban, M.Z.E. Sharafy, A combined resource allocation framework for PEVs charging stations. Renewable energy resources and distributed energy storage systems
51. J.P. Liu, T. Zhang, J. Zhu, T. Ma, Allocation Optimization of Electric Vehicle Charging Station (EVCS) Considering with Charging Satisfaction and Distributed Renewables Integration
52. S. Deb, K. Tammi, K. Kalita, P. Mahanta, Impact of electric vehicle charging station load on distribution network
53. https://interestingengineering.com
54. Global EV Outlook 2021. https://www.iea.org/reports/global-ev-outlook-2021/policies-to-promote-electric-vehicle-deployment#abstract

Design of Multiband Pattern Reconfigurable Antenna Loaded with Circular Split Ring Resonators

Umhara Rasool Khan, Abdul Basit, Javaid A. Sheikh, G. M. Bhat, and Suhaib Ahmed

Abstract This paper presents a reconfigurable antenna that offers multiband operation and pattern reconfigurability. Geometry of the patch is modified by integrating PIN diodes in the top layer of the design. The ground is loaded with four circular split ring resonators (C-SRR) for multi-band operation. Pattern reconfigurability is achieved by controlling the current in the radiating patches of the design. The antenna operates at the frequency of 4.3 GHz, 4.5 GHz, 5.5 GHz and 5.7 GHz in different states of operations (depending upon the switching condition of the PIN diodes) where substantial gain, beam tilt and bandwidth is achieved. The gain of the prototype achieved varies up to 3.1 dBi and the main beam direction in these states vary from $\theta = -30°$ to $\theta = +50º$ providing a better beam steering capability and thereof allowing for efficient use of spectrum/power.

Keywords Reconfigurability · Circular Split Ring Resonator(C-SRR) · Beam-tilt · Spectrum

1 Introduction

Wireless communication system along with artificial intelligence, data analytics, internet of things saw tremendous growth in the last decade with no sign of slowdown in sight. Nevertheless, interoperable devices are demand of the time and it is necessary requirement that the wireless devices need to be compact, lightweight, cheap and also be capable of working at different frequencies efficiently. Most of the antennas radiate in a fixed direction which is a serious drawback in certain applications like satellite communication system, radar system, on/off body antenna designs for biomedical applications etc. To overcome this, different reconfigurable antennas have been proposed till date. In addition to this, reconfigurable antennas

U. R. Khan (✉) · A. Basit · J. A. Sheikh · G. M. Bhat
Department of Electronics & IT, University of Kashmir, Srinagar, India

S. Ahmed
Department of ECE, BGSB University, Rajouri, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_24

provide optimum solution to detrimental effects that are faced by the communication system in multipath fading environment where we conventionally employ smart antenna systems that are expensive, bulky and less efficient. After studying the concept of reconfigurability in antennas, it is concluded that in comparison to multi antenna systems, reconfigurability minimizes the complexity of hardware, cuts back the number of resonating elements, reduces the design time and cost.

Different designs have been proposed on and off by researchers to address these concerns. In [1], an equilateral triangular patch antenna is proposed that offers dual band switchable frequency of operation by employing the PIN diodes. A U Slot patch antenna for wireless headset application is reported in [2] wherein main beam is steered at the frequency of operation in Bluetooth band (2.4–2.48 GHz) by employing PIN diodes. A reconfigurable multiband PIN diode-based antenna for WLAN/WiMAX is presented in [3]. In [4], PIN diodes in conjunction with chip inductors have been used for accomplishing frequency reconfiguration and miniaturization simultaneously in the proposed antenna design. In [5], a hybrid feed network is employed to excite the odd and even modes of the metamaterial-based loop antenna for achieving the pattern diversity effectively. Also, optical switches have been reported to be used for obtaining reconfigurable antennas [6–8]. In [9], top patch substrate of antenna is manually rotated to achieve pattern reconfigurability with an omni-directional or a directional pattern. A dual band reconfigurable MIMO antenna is proposed in [10]. In [10], dual feed MIMO configuration allows to achieve dual band of operation and enabling reconfiguration of the structure at the desired band of operation by employing SRR in the ground plane. A parasitic array antenna is proposed in [11] that offers reconfigurability in both frequency band and radiation pattern. By utilizing RF PIN diodes, proposed antenna operates at two different frequencies (2.38 and 2.68 GHz) with six modes at each frequency. In [12] a circular patch antenna with dual mode operation is proposed. An antenna showing a clearer picture of beam tilt in [12] is proposed for pattern diversity applications. The antenna generates two radiation patterns, which are excited using an L-shaped probe and a coplanar waveguide (CPW) network.

1.1 Motivation, Objective, and Contribution of the Paper

The interoperability of emerging wireless devices and the systems as a whole is a new norm and frequency reconfigurability in antennas in particular is a must for using them in diverse applications. Frequency reconfigurability in an antenna reduces the bandwidth requirement of spread spectrum signals as there is no need of covering all the operating frequencies simultaneously and thereof the performance of the design shall be better without any ramifications on its size and complexity [13, 14]. Other than that frequency reconfigurability shall also ensure that the prospective antenna is used in variety of applications and can be integrated with ease in modern wireless system [15]. The simultaneous reconfiguration of the frequency and pattern with

wideband operation of an antenna design is an interesting and challenging area to be explored which provides the main motivation behind our work.

The objective of this article is to propose an antenna design that offers reconfigurability in the frequency and the radiation pattern simultaneously. In addition to this, efforts have been made to achieve better bandwidth for the frequencies of interest.

This work proposes a single layer antenna design wherein the grounded is loaded with four circular split ring resonators and the three radiating patches at top are integrated with each other by four electronic switches. As briefly discussed in introduction, several authors have realized the frequency reconfigurability in the antenna design by varying its effective length. This has been executed either by adding or removing the portion of its length by utilizing different methods. PIN diodes have been employed owing to its low cost, relatively high-power handling capability, faster switching speed, good isolation, low insertion loss in comparison to RF MEMS and are also easier to fabricate for optimal performance. To control the state of the switches, dc bias is applied externally to the PIN diodes.

The proposed antenna possesses a simple geometry, a single feeding point and above all there is no need of supplementary impedance matching network for the design. The design is simple but at the same time its performance is cogent that forced us to propose this model. PIN diodes used in the slots of the antenna control the distribution of the current and thereof achieving the desired goals.

The detailed design of the prospective antenna is discussed in Sect. 2. The simulation results are presented in Sect. 3 and performance of the antenna is investigated in Sect. 4. Section 5 concludes the paper.

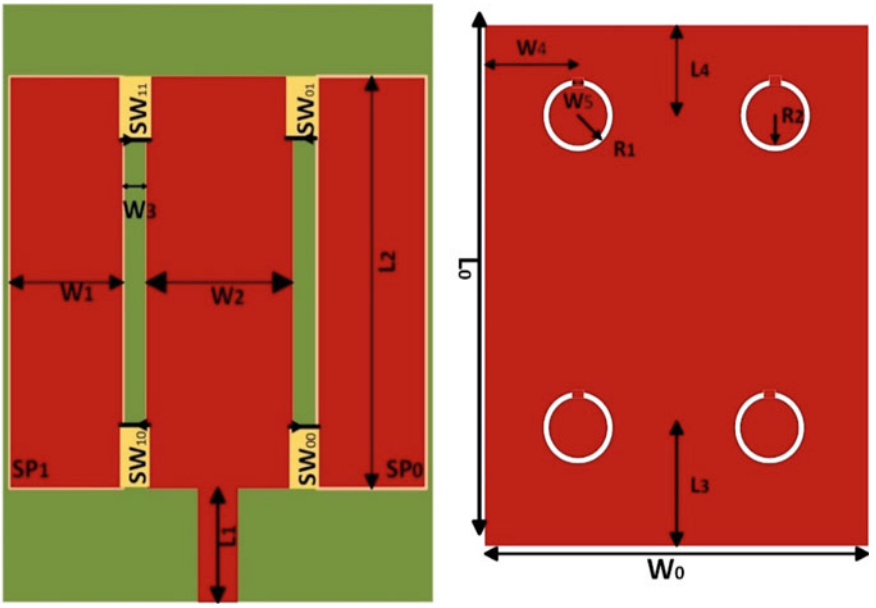
2 Design Framework

The geometrical layout of the prototype is illustrated in Fig. 1.

Figure 1a illustrates that the top layer comprises of three radiating patches (main patch, subsidiary patches attributed as SP_0, SP_1) that play an important role in current distribution and hence in determining the direction of main beam. Four PIN diodes, SW_{00}, $SW_{01,}$ $SW_{10,}$ SW_{11}, are embedded in the two vertical slots of the radiating patch. The proposed antenna is modelled on FR-4 epoxy with dielectric constant of 4.4, loss tangent of 0.02 and height of 1.6 mm. Circular split ring resonators four in number are loaded in the ground as is evident from Fig. 1b. These are placed exactly underneath the vertical split of the radiating patch. These resonators are responsible for multiband operation of the design.

Table 1 shows the detailed summary of various parameters of the geometrical layout of the prototype. The optimised size of the antenna is $0.37\ \lambda_0 \times 0.72\ \lambda_0$ and has the potential to be used in array design.

To develop an acumen about the phenomena of asymmetrical distribution created by RF PIN diodes, a diode (SMP 1345) is modelled as a resistor (1.5 Ω) in series with the packaging inductance of 0.45 nH during ON condition. When the diode is in OFF state, it is modelled as a parallel combination of high value resistor (500



(a) (b)

Fig. 1 **a** Top view and **b** bottom view of the prospective antenna

Table 1 Detailed dimensions of the prototype

Parameter	Value (in mm)	Parameter	Value (in mm)
Lo	46	Wo	39.6
L1	11.2	W1	7.1
L2	30	W2	10.4
L3	14	W3	2.75
L4	8	W4	11
R1	5	W5	1.2
R2	4.3		

kΩ) and a capacitor of capacitance 0.2 pF in conjunction with a series packaging inductance of 0.45 nH. As the diodes are utilized in the scheme, it is imperative to decouple the RF and DC supply lines so as to suppress the unnecessary interaction of the supplies. This is achieved by employing a 20pF blocking capacitor.

Table 2 defines the states of operation of the design, where the attributes of the antenna are examined and are compared. In this work only those states have been defined in which significant change in the under-evaluation parameters of the antenna is noted.

Table 2 Detailed summary of states of operation of the prospective antenna

State	SW_{00}	SW_{01}	SW_{10}	SW_{11}
0	OFF	OFF	OFF	OFF
1	ON	OFF	OFF	OFF
2	OFF	OFF	ON	OFF
3	ON	ON	ON	OFF
4	ON	OFF	ON	ON
5	ON	ON	OFF	OFF
6	OFF	OFF	ON	ON
7	ON	OFF	ON	OFF
8	ON	ON	ON	ON

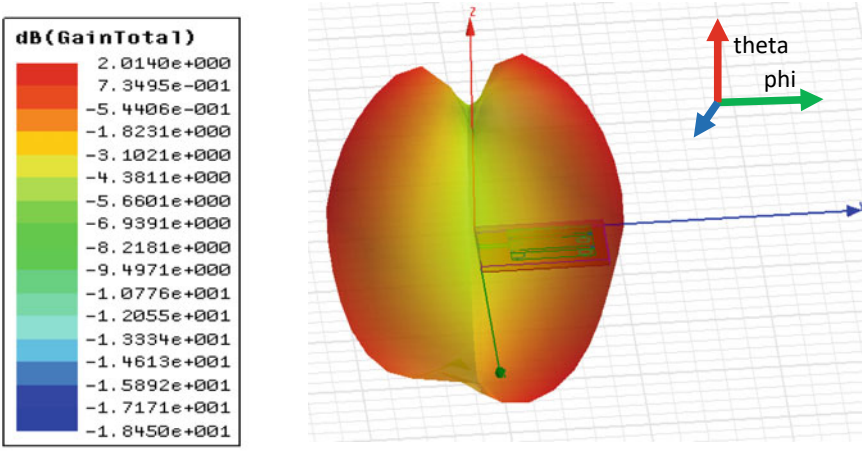
3 Result and Discussion

High frequency structure simulator (HFSS) v 15 is used for designing and simulating the prospective antenna design. In state 0, the antenna operates at two frequencies of 4.3, 4.5 GHz with a return loss of −13.29, −15.26 dB and the current is limited only to the main patch. In state 1, wherein SW_{00} is in ON mode and rest of the three switches are in OFF state, the prototype resonates at 4.3 GHz and 5.3 GHz with a return loss of −15.36 dB and −12.06 dB respectively. Figure 2, illustrates how the current is distributed in the main and the subsidiary radiating patch SP_0 that are joined via SW_0. Figure 3, shows the current distribution in the radiating patches when switch SW_{10} is in ON state. In this state, current distribution is not only limited to the main patch but extends to the subsidiary patch SP_1 also. Prototype resonates at 2.9 GHz, 4.3 GHz and 5.3 GHz in State 2 with a return loss of −11.81 dB, −13.6 dB and −15.91 dB. When all the four switches are in ON condition, the structures resonate at two frequencies viz., 3.7 GHz and 5.5 GHz with remarkable achieved gain at 5.5 GHz only. In nut shell, the main patch drives the parasitic metal strips that lies adjacent to it via the electronic switches leading to the change in current distribution and hence tilting beam in different directions. The multiband operation of the design is achieved due to the resonators that are loaded on the ground.

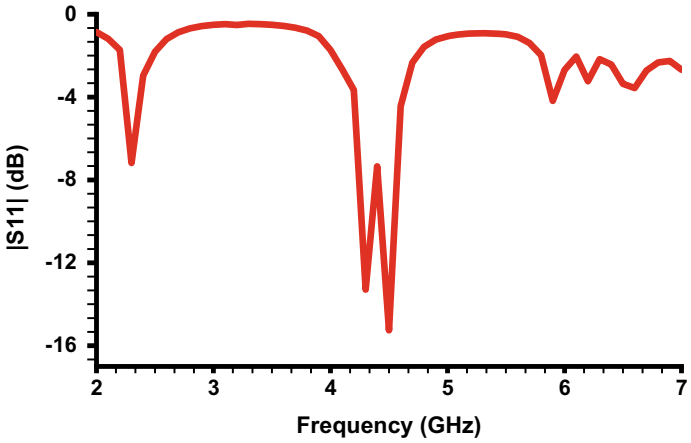
The ground is loaded with circular split ring resonators that leads to shifting of the resonating bands to lower frequencies as is validated by the following equation.

$$f_l = \frac{c}{C_s\sqrt{\varepsilon_r}}$$

where C_s is the circumference of the slot and ε_r is the effective relative permittivity of the substrate. In order to lower the frequency of operation, either the circumference of the slot is increased or the permittivity of the substrate needs to be enhanced. Embedding the slots also helps in miniaturization of the antenna. Figure 2 illustrates the 3D radiation pattern achieved in state 0. It is interpreted from the Fig. 2a that the



(a)



(b)

Fig. 2 For state 0 **a** 3-D radiated beam at 4.5 GHz **b** Return loss versus frequency plot

main beam is directed at $\theta = +50°$ at resonant frequency 4.5 GHz and the antenna shows two resonant frequencies at 4.3 GHz and 4.5 GHz in this state of operation as is conspicuous from Fig. 2b.

Conventionally pattern diversity is achieved by varying the phase of the excitation source. However, this can also be achieved by varying the current path in different modes of operation. The basic idea behind this method is that the path difference created in different states leads to the phase difference and thereof the variation in beam direction as well. Figure 3 provides the insight about the process of current steering in the radiating patches by controlling the states of operation in the design. In Fig. 4 it is apparent that the beam is steered by 20º in State 1 and 2 in the prospective antenna design.

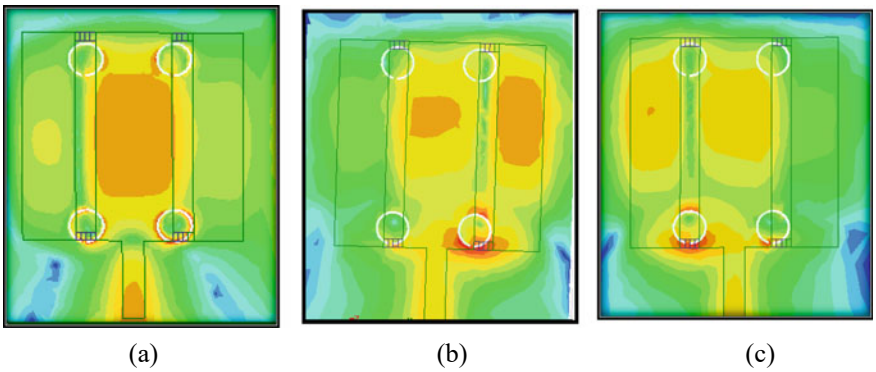


Fig. 3 Current distribution in **a** State 0 **b** state 1 **c** state 2

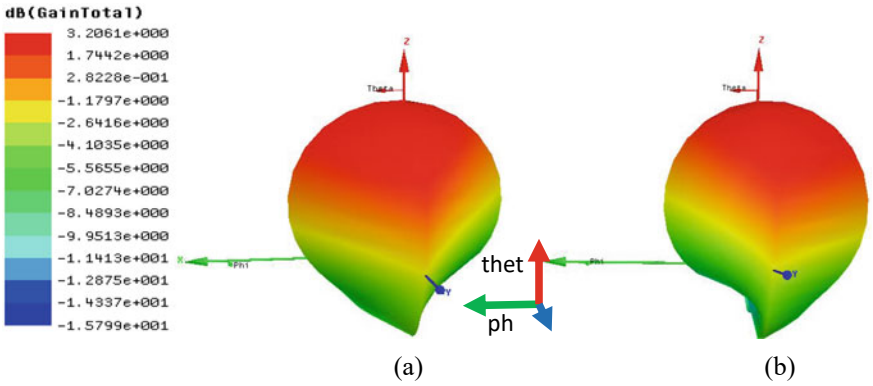


Fig. 4 3D radiated beam in **a** state1 **b** state 2 at 4.3 GHz

For resonant frequency of 4.3 GHz, the beam steers from +50º to −10° in state 4. Table 3 is self-explanatory about the impedance matching of the designed antenna at different resonant frequencies. The antenna is showing minimum losses in State 7 which infers that it is perfectly matched in this state. It is worthwhile to mention that the gain of the antenna above 0 dB is desired for most of the applications. Therefore, in the states where the gain is less than 0 dB, such states haven't been considered for performance analysis in this study owing to the fact that the design in the said states can't be utilized for any applications until and unless the design is used in arrays where its gain can be improved.

Furthermore, Fig. 3 provides the understanding about the complimentary operation of the design in state 1 and 2 but it is clearly elaborated by the Fig. 5 where the design shows resonances at more or less at same frequencies but with a beam tilt Fig. 4 which validates the claim of pattern reconfigurability.

The frequencies at which the design resonates in state 3 and 5 is depicted in Fig. 6. The antenna resonates at different frequencies in these states of operation and exhibits beam tilt at these frequencies offering interoperability.

Table 3 Detailed summary of resonant frequencies of an antenna operating in different states of operation

State	Resonant Frequency (GHz)	Return Loss(dB)
0	4.3, 4.5	−13.29, −15.26
1	4.3, 5.3	−15.36, −12.06
2	2.9, 4.3, 5.3	−11.81, −13.6, −15.91
3	3.5, 4.7, 5.7	−16.42, −22.85, −11.05
4	3.5, 4.3, 5.7	−15.44, −14.9, −10.48
5	3.5, 4.3, 5.7	−27.82, −16.45, −13.27
6	3.3, 4.3, 4.4	−12.9, −22.37, −11.24
7	5.1	−30.015
8	3.7, 5.5	−10.6–23.47

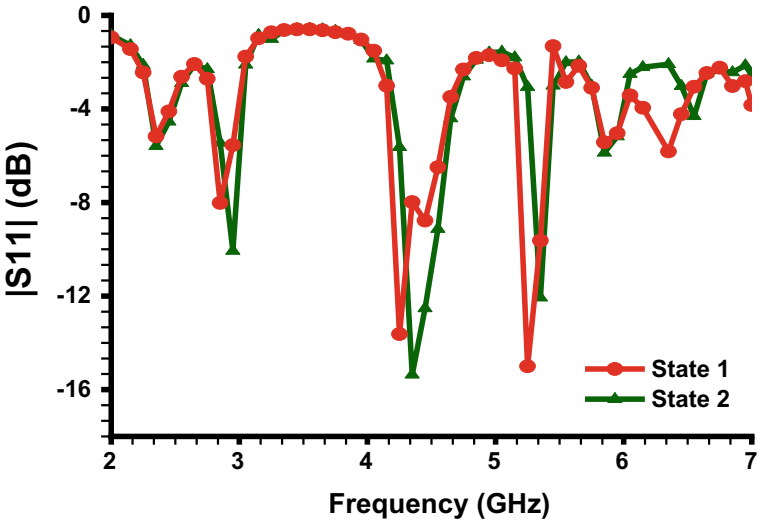


Fig. 5 Return loss versus frequency plot for state 1 and 2

As already noted above the Table 4 provides the summarized details of the different parameters of the antenna in different states of operation. It depicts how the beam direction and gain is varied in E(Y–Z), H(X–Z) planes in different states of the proposed antenna. It is implicit that the main beam is steered from −20° to + 20° for ∅ = 90° in state 1 and 2. In state 5 and 6, the direction of main beam varies from −30° to + 40°. It is conspicuous from the summarized details that the elevation angle varies from −30 ° to +50° over whole azimuthal range with substantial gain achieved simultaneously. It is not only found in these two states but Table 4 portrays a clear picture of this phenomena. It is appropriate to mention that the proposed design works at different frequencies in different set of operation; thereby allowing for multiband operation of the device and enhancing the versatility of the antenna.

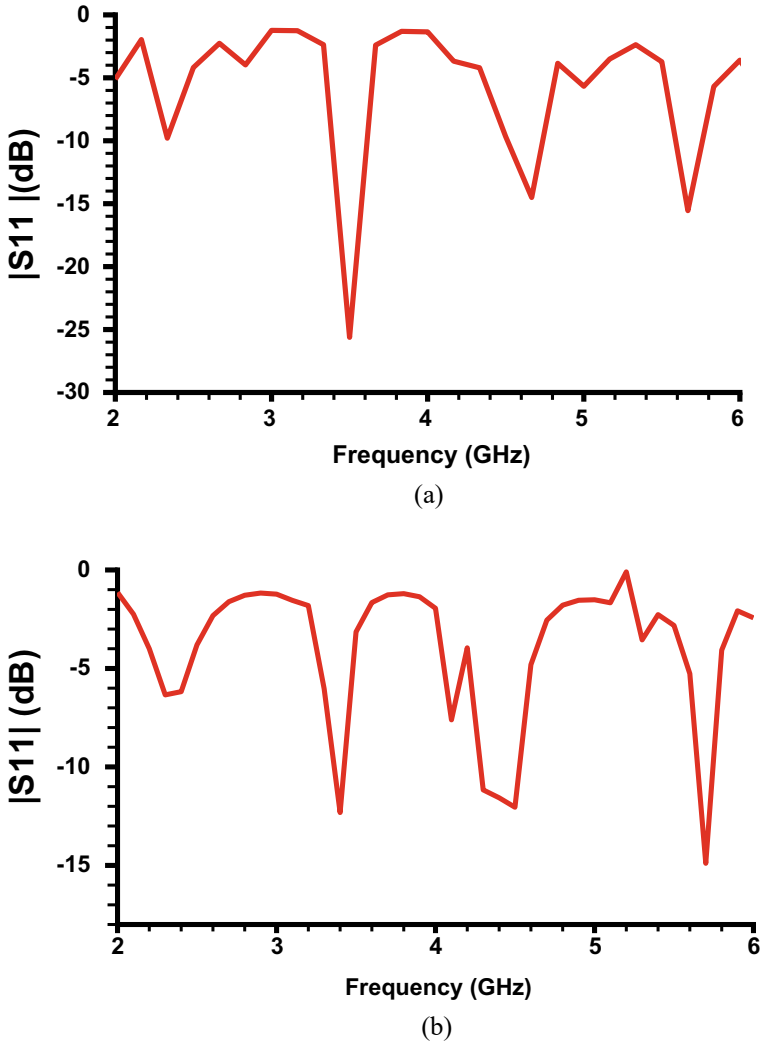


Fig. 6 Return loss versus frequency plot for **a** state 3 **b** state 5

According to 802.11b/g and 802.11a standards, the bands used for WLAN are (2.400–2.484) GHz, (5.15–5.35) GHz, and (5.725–5.825) GHz. Also, a 4.3 GHz band can be used for aeronautical radio navigation and thereof making our design a potential choice for the above applications.

Table 4 Summary of the gain achieved in different states of operation alongwith the main beam direction

State	Resonant frequency (GHz)	Total gain (dB)	Main beam direction (in degrees)		−10 dB bandwidth
			Ø	θ	
0	4.3, 4.5	0.9, 2.0	180, 90	−10, +50	1.01, 2.41
1	4.3, 5.3	3.2, −0.1	90,90	+ 20, −20	5.1, 1.01
2	2.9, 4.3, 5.3	−4.1, 3, 2.0	90, 90, 90	20, 20, −20	1, 1.5, 1.03
3	3.5, 4.7, 5.7	−2.9, 1.1, 0.3	90,90,180	−30, + 50, −40	4.4, 3.31, 2.48
4	3.5, 4.3, 5.7	−3.6, 1.8, 0.4	90, 90, 0	30, 50, 40	1.1, 4.7, 4.1
5	3.4, 4.3, 5.7	−5.1,3.1, 0.7	90, 180, 180	−50, −10, −30	4.78, 5.85, 1.5
6	3.3, 4.3, 4.4	−5.7, 1.2, 1.6	90, 0, 90	60, −10, 40	2.5, 3.2, 1.9
7	5.1	−4.1	0	0	3.3
8	3.7, 5.5	−5.6, 2.1	0, 90	−30, −20	3.79, 4.37

4 Performance Comparison

The performance of the proposed design is evaluated by computing the various parameters such as number of switching elements, type of reconfigurability, fractional bandwidth, size of the antenna and gain achieved by the design in various states of operations. The performance comparison with existing design in literature is presented in Table 5.

It is quite evident that the proposed design outweighs the other design owing to its miniaturised size, simple design, its multiband operation and at the same time achieving the frequency reconfigurability.

Table 5 Comparison of Proposed reconfigurable antenna with the existing designs

Reference	Centre frequency (GHz)	Size of the antenna	Gain (dBi)	Switching elements	Reconfigurability type	FBW
[16]	2.3	0.84 $\lambda_0 \times$ 0.69 λ_0	1.71	1-PIN diode	Pattern	1.3
[17]	4.5, 4.8, 5.2, 5.8	0.75 $\lambda_0 \times$ 0.75 λ_0	3.8	4-PIN diodes	Frequency and Pattern	2.2–3.5
[18]	2.4, 5	0.72 $\lambda_0 \times$ 0.72 λ_0	6.8	Mechanical	Pattern	52.5
[19]	1.85, 2.45	-	2.96	8-PIN diodes	Pattern	4.8,9.7
Proposed	4.3, 4.5, 5.3, 5.5, 5.7	0.37 $\lambda_0 \times$ 0.72 λ_0	1–3.9	4 PIN diodes	Frequency and Pattern	1.9–5.85

5 Conclusion

In this work, a successful endeavour has been made to design a compact multiband reconfigurable antenna. The proposed prototype consists of a ground loaded with four C-SRR for multiband operation and the top layer-patch is incorporated with vertical slits. PIN diodes are embedded in these vertical slits of the patch that controls the movement of current. PIN diode ensures the asymmetrical movement of the current that allows for the pattern reconfiguration. This design can find its potential use in different wireless applications owing to the exhibition of interoperability. The beam steering capability can be validated after fabricating the antenna and performing the measurements.

References

1. Y. Sung, A switchable microstrip patch antenna for dual frequency operation. ETRI J. **30**(4), 603–605 (2008)
2. S. Kang, C.W. Jung, Reconfigurable beam steering using U-slot patch antenna with high gain and low SAR for wireless headset applications. Microwave Opt. Technol. Lett. **57**(3), 542–547 (2015)
3. Y.I. Abdulraheem, G.A. Oguntala, A.S. Abdullah, H.J. Mohammed, R.A. Ali, R.A. Abd-Alhameed, J.M. Noras, Design of frequency reconfigurable multiband compact antenna using two PIN diodes for WLAN/WiMAX applications. IET Microwave Antennas Propagation **11**(8), 1098–1105 (2017)
4. M.-I. Lal, T.-Y. Wu, J.C. Hsieh, C.-H. Wang, S.-K. Jing, Design of reconfigurable antennas based on an L-shaped slot and PIN diodes for compact wireless devices. IET Microwaves Antennas Propag. **3**(1), 47–54 (2009)
5. Y. Zhang, K. Wei, Z. Zhang, Y. Li, Z. Feng, A compact dual-mode metamaterial-based loop antenna for pattern diversity. IEEE Antennas Wirel. Propag. Lett. **14**, 394–397 (2014)
6. G.P. Jin, D.L. Zhang, R.L. Li, Optically controlled reconfigurable antenna for cognitive radio applications. Electron. Lett. **47**(17), 948–950 (2010)
7. Y. Tawk, A.R. Albrecht, S. Hemmady, G. Balakrishnan, C.G. Christodoulou, Optically pumped frequency reconfigurable antenna design. IEEE Antennas Wirel. Propag. Lett. **9**, 280–283 (2010)
8. Y. Tawk, J. Costantine, S. Hemmady, G. Balakrishnan, K. Avery, C.G. Christodoulou, Demonstration of a cognitive radio front end using an optically pumped reconfigurable antenna system (OPRAS). IEEE Trans. Antennas Propag. **60**(2), 1075–1083 (2011)
9. P.K. Li, C.J. You, H.F. Yu, Y.J. Cheng, Mechanically pattern reconfigurable dual-band antenna with omnidirectional/directional pattern for 2.4/5GHz WLAN application. Microwave Opt. Technol. Lett. **59**(10), 2526–2531 (2011)
10. A.J.S. Annamary, N. Velmurugan, Dual feed pattern reconfigurable antenna for wireless applications. Wireless Pers. Commun. **99**(2), 829–837 (2018)
11. T. Guo, W. Leng, A. Wang, J. Li, Q. Zhang, A novel planar parasitic array antenna with frequency- and pattern-reconfigurable characteristics. IEEE Antennas Wirel. Propag. Lett. **13**, 1569–1572 (2014)
12. Y-D. Yan, Y-C. Jiao, C. Zhang, A novel dual-mode patch antenna with pattern diversity and beam-tilting for aircraft applications. Int. J. RF Microwave Comput. Aided Eng. **30**(4), e22136 (2020)

13. L. Ke, Z. Wang, Degrees of Freedom Regions of Two-User MIMO Z and Full Interference Channels: The Benefit of Reconfigurable Antennas. IEEE Trans. Inf. Theory **58**(6), 3766–3779 (2012)
14. J.M. Hannula, J. Holopainen, V. Viikari, Concept for Frequency Reconfigurable Antenna Based on Distributed Transceivers. IEEE Antennas Wirel. Propag. Lett. **16**, 764–767 (2016)
15. H.A. Majid, M.K.A. Rahim, M.R. Hamid, N.A. Murad, M.F. Ismail, Frequency-Reconfigurable Microstrip Patch-Slot Antenna. IEEE Antennas Wirel. Propag. Lett. **12**, 218–220 (2013)
16. I. Lim, T. Yun, S. Lim, Low-profile pattern-reconfigurable antenna with vertical and horizontal shorting lines in grounded CPW technology. IEEE Antennas Wirel. Propag. Lett. **13**, 1589–1592 (2014)
17. Y.P. Selvam, M. Kanagasabai, M.G.N. Alsath, S. Velan, S. Kingsly, S. Subbaraj, Y.V.R. Rao, A low-profile frequency- and pattern-reconfigurable antenna. IEEE Antennas Wirel. Propag. Lett. **16**, 3047–3050 (2017)
18. A.A. Palsokar, S.L. Lahudkar, Compact asymmetric fractal frequency and pattern reconfigurable monopole antenna. AEU-Int. J. Electron. Commun. **111**, 152915 (2019)
19. Z.L. Lu, X.X. Yang, G.N. Tan, A multidirectional pattern reconfigurable patch antenna with CSRR on the ground. IEEE Antennas Wirel. Propag. Lett. **16**, 416–419 (2016)

Optimal Thermal Coordination Dispatch for Demand Side Management

Nidhi and Kirti Pal

Abstract Now a day's demand side management is essential requirement as the sources are limited but load demand increases day by day. Two types of demand management are most popular now days, home energy management and industrial energy management. In Home energy management it is suggested to shift the load of home appliances from peak hour to off peak hour. However industrial load cannot be shifted from peak load day hours to off peak hours like in evening or night. Industrial load should be managed in day time only. In this paper an approach to manage industrial peak hour demand is presented by using solar power plant (SPP). This paper also proposes a novel method to identify the optimal location of SPP by analyzing voltage stability index. The index is calculated under maximum loading condition by using continuous power flow program (CPF). The paper explains the need of demand side management by comparing the saving of energy and cost at different penetration level of SPP. The proposed method is implemented in an IEEE 39 bus system in MATLAB environment.

Keywords Demand side management · Solar power plant · Continuous power flow · Voltage stability index · Optimal power flow

1 Introduction

As the technology has developed a revolutionized change in lifestyle has been observed. In today's lifestyle human uses many electronic gadgets [1]. The demand of electricity is always high not in personal houses but also in industries. This demand is fulfilled by conventional and non conventional energy resources. In non conventional energy resources, wind and solar energy are most popular now days. The worldwide expansion of renewable sources plants providing some of the great advantages like the reduction of carbon emissions, heterogeneity of supply, less dependence on foreign fuels, economic enlargement and jobs in manufacturing and installation of renewable

Nidhi · K. Pal (✉)
Gautam Buddha University, Greater Noida, India
e-mail: kirti.pal@gbu.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_25

sources plants [2, 3]. The major problem associated with these plants is their irregular or variable nature which sometimes makes the energy dispatching difficult. For this reason, the needs of new instruments are required for maintaining the balance between demand and supply.

Energy management is required in both generation and demand side. Energy is managed in generation side by using unit commitment method. In [4, 5] author explained how the energy is managed in generation side using solar energy by unit commitment method. Energy management in distribution network using different control technologies are discussed in [6, 7]. Home energy management (HEM) with solar energy was introduced in year 1979 [8]. Now a day's HEM schemes involve intelligent algorithms to track and manage the uses of home appliances [9, 10]. A fuzzy based controller for energy scheduling and charging/ discharging battery power for HEM is discussed in [11]. An optimization technique like artificial bee colony algorithm is also used for scheduling home appliances [12].

For the balance between demand and supply, a very encouraging tool comes into the picture i.e. Demand Side Management which provide flexibility in the economic and efficient operation of the power supply at the demand side [13]. There are a lots of studies that assimilate the DSM for economic estimation of demand response with the help from mathematical model [14]. The main aim of DSM is to find the equitable value of the demand response in alleviate with irregularity of the renewable energy such as use of solar energy are discussed in [15, 16] while combination of solar with thermal energy is discussed in [17].

This paper uses solar energy for industrial demand management and divided in 5 sections with introduction in Sect. 1. The methodologies used for demand side management of industrial load, working of continuous power flow program, working of solar power plant, voltage stability index used for optimal location of SPP are discussed in Sect. 2. The proposed methodology is implemented on an IEEE 39 bus system. The maximum loading condition for all load buses of 39 bus system is determined by CPF program. Then voltage stability index is calculated at each load bus with maximum loading. The load bus at which the index value is high is taken as most unstable bus. The results for optimal location of SPP using voltage stability index is explained in detail in Sect. 3. The industrial load reaches its peak value at day time and the output of SPP is also high at day time. Therefore the SPP used for peak hour industrial demand management with different penetration level are discussed in detail in Sect. 4. Conclusion of proposed work and result of implementation is discussed in Sect. 5.

2 Used Methodology

2.1 Demand Side Management

The changing pattern of usage of electrical energy in today's scenario has led to think of some new advanced technique to be applied to electric power utilities and that is DEMAND SIDE MANAGEMENT generally known as DSM [18].

DSM is a program that is driven by most of the utilities now. It is a program which focuses on planning, implementing and monitoring of activities of electric power usage and generally motivates the consumer to change their electric power usage pattern so that a good load profile can be obtained by the utilities and that will be producing power in an optimized way. The main objective of DSM is to encourage consumers to shift their loads to the other day of time or at off peak hours so that power could be optimized from the utility side as well as the consumer side that will help the consumer in reducing their electricity bill.

The load shape objective defined in DSM is shown in Fig. 1.

The figure basically consists of two main parameters that are valley filling and peak clipping. The base load considered at valley filling and peak clipping shows maximum loading limit. Load shifting which is the figure above shows the difference between maximum loading and base loading [19]. Load shifting objective will be determined by Continuous power flow program.

The equations defined in load shifting objective are:

$$P_{d,new}(t) = P_{d,mL}(t) + \sum_{N=1}^{L} P_{d,m,L}(t) \tag{1}$$

$$\mathrm{P}_{d,down} = \sum_{\mathrm{N=1}}^{\mathrm{L}} \mathrm{P_d} \tag{2}$$

$$P_{d,up}(t) = P_{d,new} - P_{d,down} \tag{3}$$

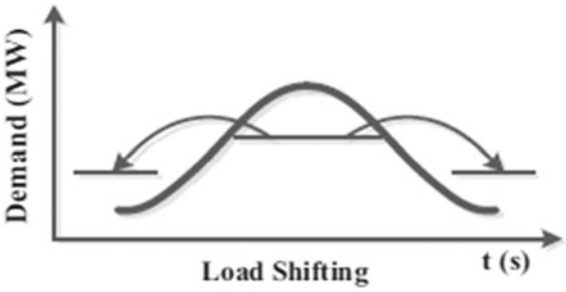


Fig. 1 Load shape objective of DSM

where:

$P_{d,new}(t)$= total load demand with DSM during (t) interval.
$P_{d,down}$= demand of individual load at base condition.
$P_{d,up}(t)$ = load shifting factor.
L = total number of individual loads.
$P_{d,mL}(t)$ = maximum loading limit of load during (t) interval.

2.2 Continuous Power Flow (CPF)

Continuous power flow method is used for solving most of the load flow equations in power system which identify the critical locations i.e. the voltage stability limit. With applying DSM program, the voltage stability comes into the picture so this need to be calculated for DSM.

The load flow analysis is done by considering Jacobian matrix which is the most significant part of load flow analysis. From jacobian matrix, eigen values are obtained which can be used to find the voltage indices and helps to find the stability point. The jacobian matrix of load flow equations become singular at the voltage stability limit that is the matrix will give the minimum value at the point of voltage collapse. With the help of "CPF" this problem can be fixed and will give the required solution to problem.

Continuation power flow program gives successive load flow solutions according to the load scenario. It consists of prediction and correction steps to reach the required critical point. From prediction and correction steps, a tangential vector is considered for obtaining critical point (Fig. 2).

The load flow equations used in "CPF" are again formulated in a different way by placing some parameters into the load flow equations that are represented as follows:

The injected powers can be written for the i^{th} bus of 'n' bus system is as follows:

$$P_i = P_{Gi} - P_{Di} \tag{4}$$

$$Q_i = Q_{Gi} - Q_{Di} \tag{5}$$

where: the subscripts G and D denote generation and load demand respectively on the related bus.

In order to reproduce a change in load, a load parameter 'W' is introduced in demand powers P_{Di} and Q_{Di} the following equations are:

$$P_{Di} = P_{Dio} + W(P_{\Delta base}) \tag{6}$$

$$Q_{Di} = Q_{Dio} + W(Q_{\Delta base}) \tag{7}$$

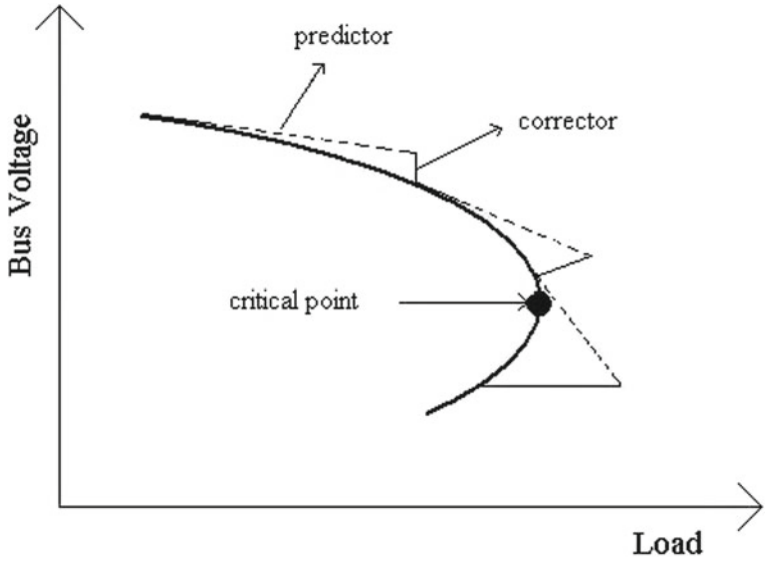


Fig. 2 Prediction-correction steps

where: P_{Dio} and Q_{Dio} are original load demands on ith bus whereas $P_{\Delta base}$ and $Q_{\Delta base}$ are given quantities of powers chosen to scale W appropriately. After substituting new demand powers from Eqs. 6 and 7 to 4 and 5, the new set of equations obtained can be represented as:

$$F(\Theta, V, W) = 0 \tag{8}$$

where ɵ denotes the vector of bus voltage angles and V denotes the vector of bus voltage magnitudes. The base solution for (W = 0) is found via continuous power flow, and then continuation and parameterization processes are applied.

Flow Chart of CPF:

The load bus with highest voltage sensitivity factor or which is more prone to changes can be thought as the most sensitive bus in a system. A feeble bus will be taken as more sensitive to load changes. In other words, the load connected to this bus is affected more than other loads in case of an unexpected load increase and will make power system unstable (Fig. 3).

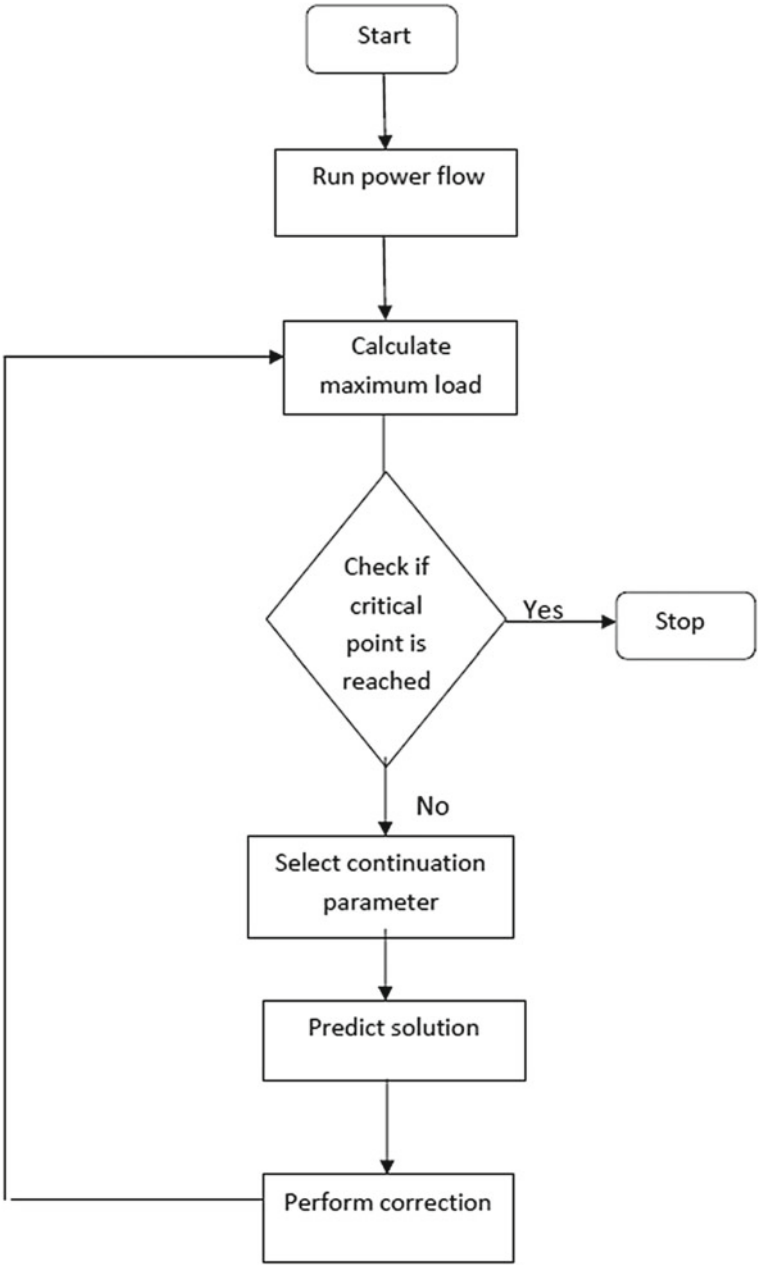


Fig. 3 Flowchart of continuous power flow program

2.3 Solar Power Plant (SPP)

In the direct conversion, the photovoltaic system or PV is utilized to change solar radiation into electrical energy and the plants used for this purpose are solar parks. The power acquire from this process is not continual and depends on the solar radiation and the atmospheric conditions outside. The output power of photovoltaic unit intended for solar radiation s is:

$$P_{PV_o}(s) = N * FF * V_y * I_y \tag{9}$$

where:

$$FF = \frac{V_{MPP} * I_{MPP}}{V_{OC} * I_{SC}} \tag{10}$$

$$V_y = V_{OC} - K_V * T_{cy} \tag{11}$$

$$I_y = s\left[I_{sc} + K_i * \left(T_{cy} - 25\right)\right] \tag{12}$$

$$T_{cy} = T_A + s\left(\frac{N_{OT} - 20}{0.8}\right) \tag{13}$$

where: N = number of PV modules.
T_{cy},T_A= cell temperature and ambient temperature (°C).
K_i,K_V = temperature coefficients of current and voltage (A/°C and V/°C).
N_{OT}= nominal operating temperature (°C).
FF = fill factor.
V_{OC} and I_{SC} = open circuit voltage and short circuit current.
V_{MPP} and I_{MPP} = maximum power point of voltage and current.

2.4 Voltage Stability Index

The voltage stability index is necessary for providing the information about the tolerable limits of the buses. If any bus is crossing the tolerable limit assigned to that particular bus then it can further create the disturbance in the system i.e. with little change or variation in load it will be the most affected bus in the system and may be responsible for the voltage instability. The reason behind the voltage instability is the reactive power in the system.

Reactive power plays a very important role in the voltage stability. It is not only needed for the active power generation but for the voltage stability too, if it is not available in the required amount in the system then it will affect the voltage stability

of the buses. With the help of voltage index, the critical locations will be detected for the installation of SPP model.

This index helps in determining the voltage violence from its predefined value.

The voltage stability index is given by

$$V_{si} = \sum_{i=1}^{N_B} \frac{1}{2n} (V_{i,b} - V_{i,db,mll})^{2n} \tag{14}$$

where V_{si} = voltage stability index.

N_B= Total number of buses.

$V_{i,b}$ = Voltage at bus i in base case.

$V_{i,db,mll}$ = Voltage at bus i when load bus db is at maximum loading limit.

n= exponent of penalty function

3 Optimal Locations for SPP Using CPF Method

The objective of the paper is to manage industrial day time peak hour demand using solar energy. To know the optimal location of solar power plant (SPP) here the voltage stability index is used. As the demand increases and power is supplied to fulfill this demand voltage vary largely. To know the maximum demand at load bus CPF program is used here under MATLAB environment. An IEEE-39 bus system with 20 load buses is used for implementing the proposed work.

This study is divided in 20 intervals as an IEEE 39 bus system is having 20 load buses. In each interval only one load bus demand is maximized using CPF program to know the maximum load of that particular bus. The voltage stability index is calculated for each interval. The result of CPF and corresponding voltage stability index is presented in Table1. From table it is easy to understand that at interval 6,8,17 and 18 voltage stability index is high with load bus 9, 15, 27 and 28 respectively. The high index value represents the instability at that bus. These load buses are selected as optimal location to install SPP. However bus number 39 has maximum active load and bus number 9 has the maximum reactive load according to Table 1.

4 Peak Hour Demand Side Management Using SPP

DSM mainly aims for encouraging the consumer to shift their loads to off peak hours like the weekend time or at the night time so that a better load profile could be maintained from the utility side and this could be reflected in reducing their electricity bills in return of this. However this kind of load management is possible only in residential load. In case of industrial load it is not possible to shift the load according to availability of power supply.

Table 1 Load bus with maximum demand and Voltage sensitivity index

Interval	Load Bus no	Load under MLL condition		V_{si}	Max. Demand (MW)	Load shifting factor	Load at base case	
		Active load	Reactive load				Active load	Reactive load
1	1	**1545.27**	699.80	0.3039	7701.9	1447.67	97.6	44.2
2	3	**1900.52**	14.17	0.2611	7832.7	1578.47	322	2.4
3	4	**1934.70**	711.97	0.3327	7688.9	1434.67	500	184
4	7	**1592.76**	572.25	0.2907	7613.2	1358.97	233.8	84
5	8	**1872.98**	633.66	0.2896	5427.4	826.83	522	176.6
6	9	**1530.59**	15,682.7	**1.0106**	7778.3	1524.07	6.5	66.6
7	12	**78.73**	812.27	0.1952	6324.4	70.17	8.53	88
8	15	**1725.17**	824.85	**0.4161**	7659.4	1405.17	320	153
9	16	**1912.53**	187.77	0.2798	7837.8	1583.57	329	32.3
10	18	**1685.43**	320.02	0.3211	7781.7	1527.47	158	30
11	20	**2199.16**	333.11	0.3006	7773.4	1519.17	680	103
12	21	**1793.94**	752.93	0.3569	7774.2	1519.97	274	115
13	23	**1817.00**	621.08	0.2975	7823.7	1569.47	247.5	84.6
14	24	**1928.87**	576.29	0.2291	7874.5	1620.27	308.6	92.2
15	25	**1759.67**	370.79	0.2628	7789.9	1535.67	224	47.2
16	26	**1670.32**	204.28	0.3057	7785.5	1531.27	139	17
17	27	**1759.99**	472.88	**0.3859**	7733.2	1478.97	281	75.5
18	28	**1619.68**	217.01	**0.3862**	7667.9	1413.67	206	27.6
19	29	**1757.28**	166.74	0.3308	7728.0	1473.77	283.5	26.9
20	39	**2565.08**	580.86	0.2591	7715.3	1461.07	1104	250

In this paper industrial load is considered to present the idea of DSM using SPP. To present the industrial load, each load bus profile is changed. The total new load of the system is 6727.23 MW. The corresponding load for each bus is mentioned in Table 2. To show the effectiveness of the proposed DSM three cases are considered:

Case 1: 50% of total increased load is shared by SPP.
Case 2: 80% of total increased load is shared by SPP.
Case 3: 100% of total increased load is shared by SPP

SPPs are located at bus number 9, 15, 27 and 28 as per the location identified in Sect. 3.

For comparison the cost is also calculated at 0% penetration of SPP. In this case the peak load power is supplied only through conventional generators. For each penetration of SPP model the cost has been calculated. To calculate the cost optimal power flow (OPF) program is used. The OPF program is run for all three cases and also for 0% penetration of SPP where the power is supplied only through conventional

Table 2 Cost analysis with different SPP penetration level

Load bus No	Active peak load (MW)	Cost ($/MWh) with 0% penetration	Cost ($/MWh) with 50% penetration (case1)	Cost ($/MWh) with 80% penetration (case2)	Cost ($/MWh) with 100% penetration (case3)
1	136.64	17.504	14.789	14.084	13.733
3	450.80	41.952	23.503	16.896	14.683
4	650.00	37.272	21.850	16.367	14.512
7	303.94	34.408	20.850	16.057	14.417
8	522.00	34.066	20.743	16.039	14.426
9	10.40	27.013	18.289	15.263	14.173
12	13.65	35.310	21.072	16.013	14.296
15	320.00	35.579	21.101	16.015	14.284
16	361.90	35.123	20.888	15.895	14.194
18	173.80	37.718	21.879	16.298	14.411
20	680.00	34.692	20.662	15.721	14.037
21	274.00	34.899	20.775	15.808	14.117
23	247.50	34.546	20.584	15.661	13.984
24	308.60	35.111	20.883	15.892	14.191
25	246.40	12.575	12.575	12.876	12.842
26	152.90	23.732	16.700	14.373	13.510
27	281.00	29.066	18.632	15.099	13.856
28	206.00	23.471	16.584	14.316	13.471
29	283.50	23.240	16.462	14.231	13.395
39	1104.00	22.288	16.609	14.737	14.013
TOTAL	**6727.23**	**609.565**	**387.43**	**307.641**	**280.545**

generators. From Table 2 it can be analyzed that by increasing the penetration level of SPP with conventional generators the costs value is showing a drastic change i.e. the costs is decreases as the penetration level of SPP increases which depicts the advantage of installation of renewable sources into the system.

As here only industrial load is considered which increases power demand only in day time. This increased industrial day load can be easily supplied through SPP. Different SPP penetration levels are used here to show the power saving and energy saving as shown in Fig. 4.

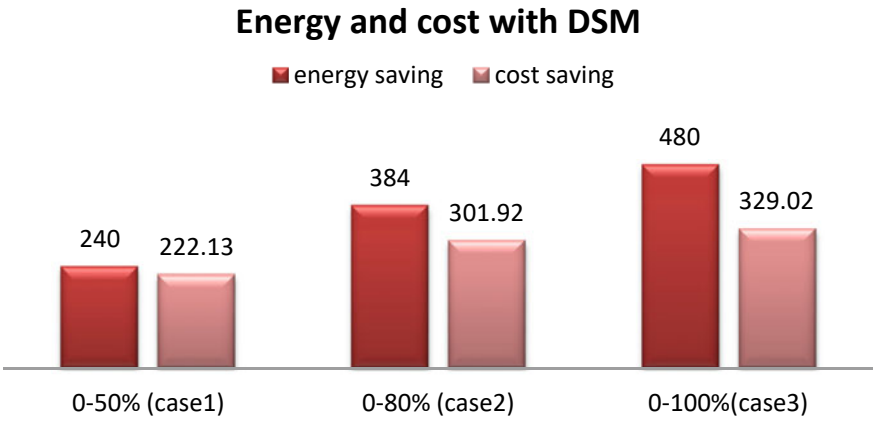


Fig. 4 Energy and cost saving with DSM

5 Conclusion

Demand side management is a very effective method with the aim to manage the electric sources which reflects in the consumers' electricity 'bill' and also adopts the method of involvement of renewable sources installation into the power system as the conventional sources will be depleting after some time but the renewable sources comes with the advantage that they are available in abundance and also has inexhaustible nature.

This paper presents the cost saving and energy saving approach for industrial peak hour day time demand energy management. To supply the peak hour demand SPP is proposed to save the thermal energy and cost of generation. Different levels of penetration of SPP are used which show that as the penetration level of SPP increases cost and energy used by conventional generator reduces. This result also validates the location of SPP as the thermal energy and cost both are reduces. Here four SPP locations have been identified under maximum loading condition of load buses. These locations are identified as most unstable load buses due to maximum load calculated by using CPF.

References

1. B. Mahapatra, A. Nayyar, Home energy management system (HEMS): concept, architecture, infrastructure, challenges and energy management schemes. Energy Syst. 1–27 (2019). https://doi.org/10.1007/s12667-019-00364-w
2. I. Khan, Importance of GHG emissions assessment in the electricity grid expansion towards a low-carbon future: a time-varying carbon intensity approach. J. Clean. Prod. **196**, 1587–1599 (2018). https://doi.org/10.1016/j.jclepro.2018.06.162

3. I. Khan, Power generation expansion plan and sustainability in a developing country: a multi-criteria decision analysis. J. Clean Prod **220**, 707–720 (2019). https://doi.org/10.1016/J.JCLEPRO.2019.02.161
4. A. Agarwal, K. Pal, Optimal unit commitment for secure operation of solar energy integrated smart grid. Recent Innovations in Computing, Lecture Notes in Electrical Engineering, January 2021, vol 701. Springer, Singapore, pp 707–718
5. A. Agarwal, K. Pal, Optimization of unit commitment problem using genetic algorithm. Int. J. Syst. Dyn. Appl. (IJSDA) **10**(3), 21–37 (2021)
6. R. A. Fuhrmann et al., Demand side management strategy for distribution networks volt/var control: a FCS-Model predictive control approach. J. Control Autom. Electr. Syst. **31**(6), 1499–1507 (2020)
7. L. Li, Coordination between smart distribution networks and multi-microgrids considering demand side management: a trilevel framework. Omega, 102, 102326 (2021)
8. R.L. Moen, Solar energy management system, in 18th IEEE Conference on Decision and Control including the Symposium on Adaptive Processes, pp. 917–919 (1979)
9. A.R. Boynuegri, B. Yagcitekin, M. Baysal, A. Karakas, M. Uzunoglu, Energy management algorithm for smart home with renewable energy sources, in 4th International Conference on Power Engineering, Energy and Electrical Drives, pp. 1753–1758. IEEE (2013)
10. Z. Garroussi, et al., A hybrid non-dominated sorting genetic algorithm for a multi-objective demand-side management problem in a smart building. Int. J. Electr. Comput. Eng. (2088–8708) **10**(1) (2020)
11. S. Zhou, Z. Wu, J. Li, X.P. Zhang, Real-time energy control approach for smart home energy management system. Elect. Power Compon. Syst. **42**(3–4), 315–326 (2014)
12. Y. Zhang, P. Zeng, C. Zang, Optimization algorithm for home energy management system based on artificial bee colony in smart grid, in 2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), pp. 734–740. IEEE (2015)
13. L. Zhengmao et al. Optimal stochastic deployment of heterogeneous energy storage in a residential multienergy microgrid with demand-side management. IEEE Trans. Industr. Inform. **17**(2), 991–1004 (2020).
14. J. Saebi, M.H. Javidi, Economic evaluation of demand response in power systems with high wind power penetration. J. Renew. Sust. Energy.**6**, 033141 (2014)
15. D. Kristin, M. Latorre Jesus, O. Luis, R. Andres, Demand response in an isolated system with high wind integration. IEEE Trans. Power Syst.**27**(1), 20–9 (2012)
16. Y. Ashkan, I. Herbert Ho-Ching, F. Tyrone, T. Hieu, An approach for wind power integration using demand side resources. IEEE Trans. Sust. Energy **4**(4), 917–24 (2013)
17. Z. Wei, P. Yu, S. Hui, Optimal wind–thermal coordination dispatch based on risk reserve constraints. Euro. Trans. Electr. Power **21**, 740–56 (2011)
18. M.H. Alham, M. Elshahed, D.K. Ibrahim, E. El Din Abo El Zahab, Optimal operation of power system incorporating wind energy with demand side management. Ain Shams Eng. J. **8**, 1–7 (2017)
19. H. Jun, B. Eng, M. Eng., The development, implementation, and application of demand side management and control (DSM+c) algorithm for integrating microgeneration system within built environment. Ph.D. Thesis. University of Strathclyde; March 2009

Optimal Routing in Wireless Sensor Networks: A Review

Jasleen Kaur, Punam Rattan, Brahm Prakash Dahiya, and Reenu

Abstract Optimal route selection is very important inside the wireless sensor network which affects the network energy. It is very important to use energy efficient manner to send packets from sensor nodes to base station. Network life-time can be incremented through Residual Energy. Many literature surveys have been conducted regarding energy efficient in WSNS. But there is still a need for a methodical review (MR) to detect valid problems in WSN to select route. In this paper, the energy efficient techniques have been reviewed based on selection criteria and quality validation. Here many algorithms are described which are based on load balancing, transmission cost and energy consumption. Through the Systematic Review, it is found that energy consumption is a huge issue in WSNs. The paper also shows drawback of existing research which consume more energy.

Keywords WSN · Routing · Energy efficiency · SLR

1 Introduction

In recent past, WSNs have become a great source of interest to research and scientific community. Though WSNs are different from traditional wireless networks, and therefore pose many challenges to solve such as limited battery energy that restricted the network lifetime etc. WSN consists of hundreds or thousands of small wireless sensor nodes equipped with data computing and communicating capabilities. A sensor node senses the physical environment, and therefore the information obtained is converted into data packets then transmitted to BS. Energy constraint is one among the foremost significant issues for WSNs, because the battery of every sensor node can't be easily replaced or recharged in harsh environment [1, 2]. Therefore, all embedded batteries of deployed sensor nodes need to be managed, so as

J. Kaur · P. Rattan · Reenu
Computer Application C T University, Ludhiana, Punjab, India

B. P. Dahiya (✉)
Computer Science, BLM Girls College, Nawanshahr, Punjab, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_26

to prolong the network lifetime. WSNs are prejudiced by several challenging influences as précised below. It is projected that wireless sensors will play a superior and superior role in our everyday's' life in the near future of new types of applications. It is connected with number of devices and send/receive processed data to/from other sensing nodes. We are well on the way towards a hundred of billions-wireless sensor are available in market [3]. This development is ongoing the wireless sensors becoming an ever-present component of our life in homes, vehicles, traffic, healthcare, education, production and monitoring & controlling our daily activities. It is an important element in this advancement that participate by low-cost hardware and resource-efficient solutions in software development [4, 5].

Sensor nodes that are used in WSNs are dissimilar from normal sensors because it can sense or monitor the physical quantity, process and provide the long distance communication function. The complete architecture of a senor node is shown in Fig. 1. The brief explanation about the sensor node architecture is as follows.

Calculation Subsystem (CS): The function of CS is extremely complicated because it considers node functional control and processing. It's closely associated with other subsystems. The energy consumption of the whole nodes are often suffering from the continuous different active status, function and performance also because the exchange of various conditions and shortly.

Communication Subsystem: It includes baseband processing part and the radio frequency (RF) communication part. The subsystem energy consumption is affected by data rate transmission power, modulation mode and the frequency of the wireless transceiver device. Due to physical characteristics of data communication, the energy consumption during receiving time is same as even with the idle time.

Sensor Subsystem: Sensor subsystem comprises a collection of transform device and sensors nodes. The energy depletion of sensor subsystem is the result of converting the physical signal into an electrical signal and then converted into digital signal with help of A/D converter.

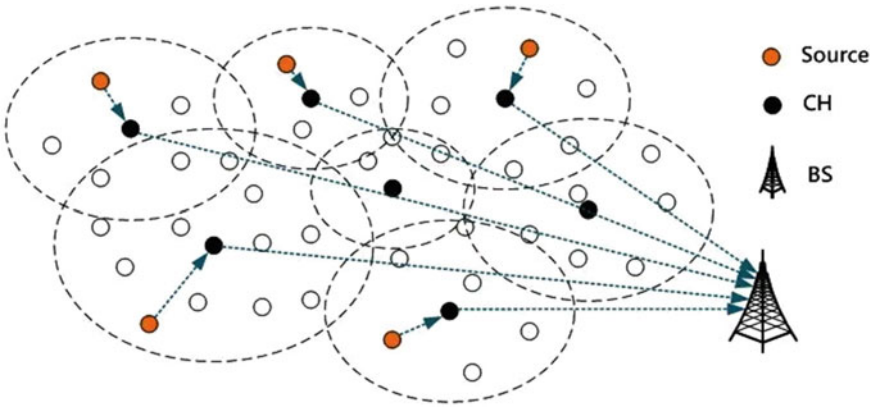


Fig. 1 Routing in WSNs

Energy Supply Subsystem: The power unit includes battery source to drive all the units in the system. These sources also depend upon the usage of the node in a specific application. Large part of the energy is wasted in communication rather than in processing the information. In some applications, sensors are equipped with specific units for example; the global positioning system uses to track location information in WSN. Therefore, sensor nodes movement is required to capture the location information. There units have small in size, low production cost and less power consumption [6, 7].

1.1 Network Design Constraints

Network capabilities and performance depends on the features or characteristics of the sensor network. Following are features of sensor networks designing [8–15].

Small Node Size: The primary design objective is to minimize the node size, because of a large number of sensors installed in harsh environment. So the objective is to keep the minimum cost and also energy saving sensor node deployment

Low Node Cost: Another objective is to have the most cost effective sensor network. It becomes more important to control the cost because the sensor nodes die due to environmental effects and it is hard to replace them.

Low Energy Consumption: Another important objective in network design is to make it energy efficient. As sensor nodes operate on batteries and it is hard to replace the batteries, the network needs to be energy efficient to prolong the whole network lifespan.

Scalability: The sensor network scalability is very important. Since the sensors are in order of thousand or million, the network design should be accessible to different sizes of the network. Network topology changes happen with a node failure, join and move. Thus network protocol should be adaptive and cooperative in topology changes.

Reliability: The emphasis is on the smooth transmission of data without any errors and time varying channels. So wireless protocol design is to ensure that.

Data transmission is well controlled to make it error free and reliable data delivery.

Fault Tolerance: WSNS nodes are affected by environmental changes. So sensor nodes should have fault tolerance self-testing, self-repairing, self-calibrating and self-recovering capabilities.

Channel Utilization: Because of restricted bandwidth resources, sensor network should ensure channel utilization in a way to make maximum bandwidth.

The requirements; based on delivery latency and packet losses. Some examples are like in fire monitoring and in scientific research, the packet delay and packet loss can't be afforded. So for designing the application, understanding the parameters is very important.

There are many ways in which the protocol will be beneficial, but consider some following metrics:

Ease of Deployment: The sensor network may contain hundreds or thousands of sensor nodes in harsh environment.

Network lifetime: WSNs lifetime is based on nodes alive, first nodes dies and last nodes dies. Sensor nodes lifespan play an important role to improve WSNs lifetime.

Delay: Data from the sensor have to be fast and reliable. In some applications, data is time sensitive and delays cannot be afforded.

Quality: The results of sensor data should be very close to reality. Quality is measured by the amount of data received at BS either actual or aggregate data.

Remaining dissertation is separated in following manner: In section two we address SLR; here all the processes related to research become the part of discussion. In next section which is the third section review of written material is submitted in relation to examined papers. In next section which is the fourth section conclusion of our dissertation is submitted.

2 Literature Review

This section yields the related work that has been done so far on the hot topic of energy efficient clustering algorithms used in WSNs. In this protocol variable technologies of group based transmission protocol are determined in the form which is explained below [15–30].

2.1 TEEN Protocol

Inside the wireless sensor network this type of protocol exists in the form of those transmission protocols which is highly energy efficient. It is implemented for reducing energy usage for the improvement of constancy and network life span. Inability of expired nodes to reach concluding point and inability of consumer to take information out of network becomes the weaknesses of this protocol. Because of this reason nodes fail to communicate in the company of group-head. For this additional transmission protocol TSEP is bring in to use. Such type of protocol contains various stages of diverse nodes; advance stage, middle stage and usual stage. Power consumption is the main reason behind the use of teen protocol. This protocol becomes favorable in support of time sensing serious request. It becomes the most important issue of TEEN.

2.2 APTEEN Protocol

Number of mixed transmission protocols is already available in support of energy usage and network lifespan. Here, a new transmission protocol which becomes

famous with the name of APTEEN is introduced for the very first time. It is a dissertation in which [26] researcher make a comparison in the middle of TEEN, LEACH and APTEEN protocol in support of energy usage and network lifespan. In order to send data in a periodical manner this protocol utilized the flexible and realistic network. Here "Security models for hierarchical clustered wireless sensor network" [27] submitted safety arrangements, safe transmission plans and methods of key management. Safe hierarchical safety transmission protocols are used in support of safety and address models of safe transmission, key management and group formation model.

In order to organize keys a method is introduced in [28]. Inside the wireless sensor networks, a group of key is assigned to all the cluster head and cluster keys are assigned to all clusters in the Base Station. Here, base station becomes an immovable point. Here mobile phones of user are communicated. Inside the network, for the purpose of communication in the middle of various nodes such types of keys are employed. For ensuring safety, information is transmitted with the help of encoding for improving network safety.

2.3 Pegasis

It exists in the form of protocol which transmitted packets of each and every node in the direction of base station in a sequential manner [29]. For enhancing network life span, uses local information calculation is used by LEACH protocol for sending information in the direction of sink-out of group. Its suitability in support of like network and its unsuitability in support of unlike becomes the most significant concern. SEP becomes a protocol which will provide advance and usual node. In comparison to usual node energy of other node is high.

2.4 Energy Efficient Distributing Clustering

Inside the wireless sensor networks, usage of energy in an improved manner becomes the most important objective is to improve the energy efficiency. It is a dissertation [30] where two methods are proposed by researcher for the minimization of end to end delay and energy usage. Primarily, the distribution grouping method, in support of energy usage and end to end delay is proposed. It is a method in which groups select excellent group head. After that, inter grouping transmission method in support of end to end delay. It is a method in which a latest cost function of energy is used. Lots of practical explanations are already available in support of energy usage and delay. At the same time, these explanations fail to become perfect in support of long timecoverage. Inside the wireless sensor networks, grouping methods are used for the purpose of energy usage and end to end delay due to their efficiency. Tree of groups are combined on the basis of those distributed scheduling method which

are energy efficient [31]. Here, each and every information which came out of the entire sensing device is combined by the author. Here, delay becomes minimum in those situations where combined packet reaches to the sink node. More than one transmission method is employed in support of group. Out of which, the transmission method which is bring in to use in support of energy efficient chain from the clusters to the sink becomes helpful in 3 hop groups [32].

In HEED [30] based on intersection, remaining usage of energy and group head limits are selected occasionally. DEAR exist in the form of additional transmission protocol [33]. Since DEAR is a multipath routing protocol, remember delay, energy usage, and efficiency. The researcher identifies two problems linked to energy effective routing in Energy Delay Optimization in an asynchronous Sensor Network with Several Gateway [2]. The first step is to build the tree that will inform you how to route efficiently. The second issue is how to assign wakeup frequencies to various transmission trees. For the resolution of primary problem a perfect explanation is provided by the author. The perfect resolution in support of second problem is the implementation of polynomial time method because they are the NP related problems. Inside the wireless sensor networks balance in ecological transmission occurs for energy delay. Inside the wireless sensor networks, at the time of CH election those optimization problems become more important which possess multiple goals [34]. It integrates information of corresponding sensor nodes and transmitted this information in the direction of sink. In order to dissipate energy hardware radio, l-bit received information. In a packet whose length is constant group heads are implemented for getting information out of its members. For expanding group head member's fuses of group are used.

2.5 LEACH Protocol

Inside the wireless sensor networks WSN, consumption of energy in an efficient way becomes the most significant problem. Preserve valuable sensor nodes in the result of energy efficient. Its contribution in network life span is very significant. Inside the wireless sensor networks sensing device depends upon battery, in which energy remain consistent when it is available in imperfect amount. LEACH is the transmission protocol which is implemented here. It becomes the protocol which is utilized in support of energy usage and for the improvement of network life span. Inside the wireless sensor network various methods are used for the designing of group based wireless sensor networks. For the transmission of information in the direction of base station parallel data and events are utilized in an exclusive manner. Group head is employed for reducing energy usage. It is a dissertation where a group head LEACH protocol has been proposed by them. This protocol becomes useful in support of energy usage and for the improvement of network life span. When group head send a request message in its direction it established inside the transmission table. For fixing each and every node, group will remain up-to date as per TDMA list. Use of grouping transmission protocols inside the network, transport and data link

layer improves the network effectiveness and tackles the "hot spot" related issues. Inside the wireless sensor leach is the primary organized transmission protocol. In leach node of energetic group head and native cluster sets becomes the most significant attributes. It uses stochastic method [35]. In leach minimum lifetime of the nodes and the maximum volume of data mains to occupied with deceased nodes. In this network extendibility related issues fails to remain appropriate. Leach-C exists in its modified form where information related to remaining energy is transmitted by nodes in the direction of base stationing the beginning of all rounds. In situations where the base station is located in the vicinity of sensor nodes it fails to achieve better efficiency in comparison to leach. This will become the main weakness of LEACH-C. The clustering algorithm's primary goal is to reduce the amount of energy used by nodes. The nodes' conditions are to distribute the load over a fixed time frame [36]. This schemes demand high energy inside the clusters and to relay the aggregate data to the BS, so a sensor node can pick the cluster head with the required signal power. The leach procedure has the downside of not being shared in the cluster head.

The Leach-C protocol is used for this [37]. The low energy adaptive clustering hierarchy (leach) is used in the sensor network to minimize energy usage, maximize network lifespan, and split the average energy node. The Leach algorithm is divided into posterior and forward numbers, which are then allocated for the stable phase and cluster forming. This paper's key issue is clustering focused on hierarchical structures. As seen in Table 1, routing protocols are used for energy cluster scalability, scalability, distribution delay, and load balancing performance.

Hybrid Energy Efficient Distributed clustering (HEED) is clustering technique. It worked basis on residual energy in the election of CH. It is extended version of LEACH. It is used to minimize the communication overhead. But it consumed high energy for local communication between CH and base station. The Power Efficient Gathering in Sensor Information Systems (PEGASIS) is consumed lower energy during CH election. Simulation results showed that PEGASIS perform better than LEACH for various sized networks. But, the problem with PEGASIS is that during the leader selection for each round, the remaining energy is not taken into the account and there is only one sensor for connecting to the base station [38–40].

Table 1 Cluster based transmission protocols

Scheme name	Power consumption	Adaptively of group	Adaptively	Delivery delay	Distribution of load
LEECH	Very low	Average	Extremely less	Extremely less	Average
HEED	Average	Good	Average	Average	Average
TEEN	Very High	Good	Bad	Less	Excellent
PEGASIS	Low	Bad	Extremely less	Very long	Average

In other hand Energy Delay Index for Trade-off (EDIT), this is a routing algorithm. In this both objectives (energy and delay) are optimized. A large number of communication and energy nodes are deployed in WSNs. Once these are deployed in an atmosphere, there is a difficulty to replace and recharge the batteries of WSNs nodes. Therefore, load balancing must be done for communication between the WSNs nodes [41, 42]. A Genetic algorithm is used to increase the WSNs lifetime, by utilizing auction data of node i.e. its local battery power, topology strength and intelligent CH selection [43]. An Ant Colony Optimization (ACO) is a clustering technique, which is based on swarm intelligence. In this LEACH protocol is applied on ACO algorithm in WSNs, to reduced energy consumption [44]. A Stable Election Protocol (SEP) developed to improve the lifetime of WSNs. In this protocol the nodes which have the maximum energy among all of the accessible nodes. Therefore, those particular clusters are selected as a CHs [45].

Other side Cluster Chain Weight Metrics (CCWM) is used to select the limited number of CH based on weight metrics. Node has highest metrics value is selected as a CH. The simulation result CCWM is better than LEACH by achieving higher energy efficiency; performed better load distribution and improved network lifetime [46]. A genetic algorithm proposed solution, to elect the optimal CH using load balancing technique and reduced energy wastage as compare to LEACH [47]. A self-configurable clustering mechanism (SCCH) is used to elect CH. It replaced disordered nodes by selected more intelligent node as a CH [48]. A new algorithm introduced i.e. Regional Energy Aware Clustering with Isolated Nodes (REAC-IN). It is a new local energy aware clustering method. It is used isolate nodes for WSNs and CH elected basis on nodes weight [49].

An improved Orphan LEACH (O-LEACH), which provided coverage of the whole network with a minimum number of O-LEACH nodes and has a very high connectivity rates. Based on the simulation results O-LEACH performed better than LEACH in terms of coverage, connectivity rate, energy and scalability [50]. A Fuzzy logic based clustering algorithm proposed to improve life spam WSNs. In which Super-CH (SCH) is elected among the CH, who can only send the information to the mobile BS by choosing suitable fuzzy descriptors. The proposed protocol result better than the LEACH protocol in terms of the first node dies, half node alive and better lifetime [51]. The dynamic CH selection method for WSNs is used to solve the problem of the unreasonable cluster head selection. The network lifetime is increased by 50%, higher than LEACH [52]. PSO algorithm which is one of the popular optimization algorithms is applied for routing in WSN [53] and evolutionary based routing protocol are often found in [54]. Very recently, Harmony Search Algorithm (HSA), which is one among the optimization approaches used as meta-heuristic optimization method, is employed as clustering based routing protocol for energy efficient WSN. ABC algorithm is recently applied to pick the CH selection for energy reduction in WSN and subsequently, hybrid ABC algorithm is additionally applied for routing to increase WSN lifetime.

Most of the above routing protocols required global positioning system in WSN, because initially the situation of every node should be known to execute these protocols. the utilization of GPS system requires additional hardware that increases the

dimensions of nodes and price of the system. ACO algorithm to look the several tracks after cluster formation. But this algorithm doesn't consider inter and intra cluster delay and prior information of pheromone to seek out the paths. It is planned to increase the localization correctness for unknown sensor nodes that counted distance between network nodes and unknown nodes.

Additionally, the fitness function of NS-IPSO might be possibly enhanced with the consideration of the sensor node partition and anchor node distribution.

This protocol is compared in the company of protocols like ETSSE protocol and TSE protocol and it is evaluated in the company of SE protocol. There are various stages in grouping process. These stages are election of group head, formation of group, and the transmission of information in the direction of sink. The election of CH is done on the basis of possibilities. When the selection of cluster head is done, nodes intellect data in the direction of group head and after that it is referred in the direction of sink. In E2HRC grouping method, node position remains safe in those situations where group head operates the entire function [17].

Inside the wireless sensor networks, detection of complex events exists in the form of major problem. Number of methods is already available for the transmission of information and for the detection of event. Out of these methods on is routing tree. In the context of energy, this tree fails to become perfect. Therefore, a tree which becomes famous with the name of event detection is introduced by the author. The basic intention behind the introduction of this tree is to make the consumption of energy effective and for tackling the capacity of transmitted data [52]. The events which are detected freely with the help of sensors becomes famous in the form of original events and the events which are detected on the basis of various correlated sub-events becomes famous in the form of hybrid events. The primary stage is the stage where formation of EDT takes place and the stage in which complex events are identified in support of equivalent data become the second stage. [53, 54]. One of its major advantage is that when the input node is small and fails to become obvious then the cost of transmission of input node grows.

Inside the wireless sensor networks, useful protocols are implemented in support of trust reputation system. The coordination in the middle of nodes and candidate election concentrates on OR protocol [55]. EXOR becomes the novel perspective of transmission. It becomes possible for EXOR to work in a successful and resourceful manner [56]. Inside the wireless sensor networks [57–59]; safety exists in the form of most important problem. The bias reason behind this is the transmission of data in the direction of those routes which not developed [61–67]. Inside the wireless sensor networks, active transmission arrangements are introduced in support of that social network which is established on the basis of trust (Table 2).

2.6 Safe Energy Efficient Transmission Method

Inside the wireless sensor networks shortest route in the middle of nodes is selected with the help of routing. Energy emerges in the form of biggest problem since last

Table 2 Assessment table related to group based transmission protocols

Scheme	Effect on power consumption	Weaknesses
EAHBEPP [9]	For the achievement of power consumption they use Holistic method	In this, dissertatisn holistic method is used. Its mill weakness is tbetime which it takes forth? executisn of difficult calculations [44]
SCERP [3]	Consumes more energy when It is far away from the BS	Black hole attack [68]
EECEEP [5]	In comparison to another present day protocol energy become h high efficient	Here, node of group head is elected in a random manner. At the same time, election of node marandom manner is not secure because it becomes possible that a node is Sybil mode [69]
EECBRA [4]	In comparison to LEACH protocol the efficiency of th is algorithm is four energy joule	Multiple channel attenuation becomes the main weakness of this dissertation because it decreases the system effectiveness [70]
EEKCP [10]	It has been demonstrated by output that the energy of submitted mehod is good	The drawback of k-means is that it only changes the symmetrical data and k means accpts to deal with a circular cluters and that each cluster has randmoly equal numbers of opinition [71]
SEP-E [11]	It is a protocol in which energy improves up to thirty percent	SEP protocol fails to maintain alternation in energy [48]
MHR [12]	It nearly use eighty five percent energy	Use of similar links which tails to provide unwanted m mathamatical signals. Additional weaknass is the trail Emission of key before transmission of original information [74]
DHCR protocol [13]	Energy consumption e improved up to ten percent	Black hoe attach [68], Fabrication attack [75]
CH-Leach [14]	Energy consumption is increased up to eighty five percent	Each and every variable is similar and it is spread in circular form. If one o fib am miss, tbEnthEkmEan fails [71] Sybil attack [10]
REEHCBR protocol [15]	In comparison to existing ire nearly hundred percent better	Black hoi? attack [68]
HA [16]	Expired nodes reduced up to eighty eight percent wiifc the help of better methods	Black bola attack [68]
E2HKCHA [17]	In comparison to real RPL the method is one point seven percent better	Election of node in a random UH IS not secure because it becomes possible that a node is Sybil node [69]

few decades. The data which is transmitted in the middle of nodes is prevented from attacks with the help of transmission protocol. Within "RAEED-EA: A formally Analyzed Energy Efficient WSN routing Protocol, nodes are equipped with routing protocol of Robust Formally Analyzed Protocol for WSN (RAEED)". In the latest form of "Energy Aware (RAEED-EA)", transmission related decision is taken on the basis of output and remaining energy. For the evaluation of effectiveness, transmission protocols are assessed in a very formal way by determining network life span. One of the techniques is formal verification that is used to check the paths and to measure minimum and maximum consumption of energy. It provides security against Daniel of service attack [70–72].

Once the sensing data is collected it is transferred with the help of clusters. Once the data is accumulated out of nodes it is transmitted in the direction of base station with the help of group head (GH). It becomes possible to separate GH into two sections for tackling the burden of group head. Out of these two sections one is RCH and the other is DCH. RCH tackle the path of base station and nodes. In DCH accumulated and compact data is referred in the direction of RCH. In situations where the base station energy level is good the capability of transmission is also good. At the time of group formation, group head is selected by the base station for forwarding the accumulated data in the direction of base station. For the accumulation of information DCH is bring in to use. The main disadvantage of data accumulation is that it becomes complicated in situations where end to end privacy in the middle of system and the sink is expected [73]. RAEED exist in the form of that transmission protocol which is utilized in support of safety in opposition to denial of service attacks. Here, a better form of RAEED is submitted i.e. RAEED-LB which operates in different stages [74–76].

These stages are Key Setup (KS), Route Setup (RS), and Data Forwarding (DF). In order to encrypt data primary stage is used. For this purpose, pair and group key is interchanged in the company of each other. Information is exchanged by nodes with the help of pair keys. Information are encrypted with the help of group keys. In the second phase know their neighbors and next level of node. Here, a level is assigned to all nodes and it depends its hop distance from the sink. In the last stage the adjacent node which has maximum output is selected. Appropriate alterations are done inside the last stage for the purpose of node selection. The allocation of burden in RAEED-LB protocol verifies that information always remain in specified limit. The distribution of load in a balanced way becomes the most significant problem. Inside the system, because of method when the quantity of improved nodes is large it will make a huge impact on efficiency. A list of publications are discussed in Table 3 in a brief manner. All of these publications offered safety. Load balancing and energy efficient path from multiple different approaches.

RAEED-LB is modified by the author of publication [23]. Distribution of load in a balanced manner becomes the most significant concern inside the transmission of wireless sensor networks. In network when large amounts of nodes are improved due to the cause of procedure it will affect the performance and too small the quality of services as well [73]. EDT is used by the author of publication [55]. It brings

Table 3 Assessment table related to safe power consumption transmission method

Method	Consequences of power consumption	Attacks
OR metric [21]	Ratio of Power lies in between five to twenty percent	It becomes a dissertation in which opportunistic transmission is utilized. Inclusion of forward set election and prioritization becomes its main weakness [60]
HCDM [22]	Effectiveness of Energy improved up to sixty five percent	For the accumulation of information DCH is bring in to use. The main disadvantage of data accumulation is that it become complicated in situations where end to end privacy in the middle of system and the sink is expected[73]
TPLBSOR [24]	Sixty five percent to energy is used	Drawbackofopportunisticroutingisthatitcontainsforwardsetselection and prioritization [60]
EXOR [25]	Not mentioned	Cryptanalysis attacks, input based attacks and so on [76]
RAEED-LB [23]	In five hundred units it improves thirty five percent in comparison to RAEED	
EAHBRP [9]	For making power effective Holistic method is employed	
SEP-E [11]	The energy in SEPE protocol is approximately thirty percent more	
MHR [12]	Nearly eighty five percent of energy is used	
CH-Leach [14]	Energy is improved up to eighty five percent	Sybil attack[69]
EECFPCHSP [79]	Energy becomes effective in support of numbers rounds	Not mentioned
REEHCBRP [15]	In comparison to present day protocol it is nearly hundred percent better	Black hole attack[68]
TPLBSOR [24]	Energy is used up to sixty five percent	–
SCERP [3]	SCREP protocol consumes more energy when it far-away from the BS	Black hole attack
EECBRP [5]	In comparison to other present day protocol it uses energy in a highly effective manner	Sybil attack[69]

(continued)

Table 3 (continued)

Method	Consequences of power consumption	Attacks
ILC [16]	Its power consumption is good for overcoming the in active nodes near about eighty eight percent	Sybil attack[69]
EECBRA [4]	In comparison to LEACH protocol the efficiency of proposed method is improved by four joule	–
EEKCP [10]	It has been demonstrated by outputs that the power of submitted plan is good	–
NTARP [19]	In comparison to already existed TRPM is eleven percent safe	–
DHCRP [13]	In comparison to another protocols it is approximately ten percent better	Black hole attack[68],Fabrication attack[75]
EXOR [25]	In comparison to present day protocol the efficiency of proposed method is improved by twenty five percent	Cryptanalysis attacks, attacks which are based on input and many more [76]

in to use for making consumption of power effective and for tackling the different information. For detecting alternation which takes place inside the wireless sensor networks reality is bring in to use. [52]. For interchanging of transmission related data packets of Hello are bring in to use.

Information's are transmitted by all nodes in the direction of sink node with the help of shortest route. It will create flooding in complete system. In [16], modification is done in leach in such a way that the effectiveness of its power improved when the amount of inactive nodes reduced to near about eighty eight percent. In situations where head node becomes inactive it encounters black-hole attack [68].

2.7 Safety Assessment

It is a dissertation which is performed on the basis of already available written material. Here, it is considered that it becomes possible to analyses few publication in the company of safety view point. Occurrence of various attacks is still possible. Attacks which take place on various methods are discussed in table seven in an absolute manner. For avoiding such type of attacks no prevention is needed. HCDM method is bringing in to use by the author of publication [22]. DCH method is bringing in to use for the purpose of data accumulation. The main disadvantage of data accumulation is that it becomes complicated in situations where end to end privacy in the middle of system and the sink is expected [73]. The use of DES makes this approach suffer from key search attacks. CH- Leach protocol is bring in to use by the author of other publications [14]. Here, it is assumed by k-means that variance of each and every variable is similar and it is spread in circular form. If one of them miss, then the k mean fails [71] Sybil attack [69]. In table eight various publications are compared. It is table which clearly specifies which publication is completely depends upon power consumption, which one is based on safe transmission and which one is based on both.

3 Conclusion and Future Work

The improvement of science and technology is increasing day by day and WSN has become very popular with it. In this way, sensors are used in different environment. The energy of the sensor is limited, for which it is very important to optimize the sensor energy. Energy efficiency is a very important issue. Many algorithms have been introduced that work on energy efficiency. This paper describes methodical literature review to create research questions. The energy efficient techniques have been reviewed based on selection criteria and quality validation. Here, many algorithms are described which are based on load balancing, transmission cost and energy consumption. After this survey, we have found out the drawback of all the existing

algorithms. After this Literature review, it will be easier to propose energy efficient algorithm in the future to extend life time of WSN.

References

1. R. Sandhya, N. Sengottaiyan, S-SEECH secured- scalable energy efficient clustering hierarchy protocol for wireless sensor network, in International Conference on Data Mining and Advanced Computing (SAPIENCE), Ernakulam, pp. 306–309 (2016)
2. R. Cohen, B. Kapchits, Energy-delay optimization in an asynchronous sensor network with multiple gateways, in 8th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks, Salt Lake City, pp. 98–106 (2011)
3. V. Bahale, M. Pund, Performance evaluation of SCERP: a cluster based network lifetime improvement protocol for WSN. Int. Res. J. Eng. Tech. **04**, 2462–2469 (2016)
4. K. Awan, A. Ali, F. Aadil, K. Qureshi, Energy efficient cluster based routing algorithm for wireless sensors networks. International Conference on Advancements in Computational Sciences (ICACS), Lahore, pp. 1–7 (2018)
5. J. Shen, A. Wang, C. Wang, P. Hung, C. Lai, An efficient centroid-based routing protocol for energy management in WSN-assisted IoT. IEEE Access **5**, 18469–18479 (2017)
6. W. Zhu, J. Cao, M. Raynal, Energy-efficient composite event detection in wireless sensor networks. IEEE Commun. Lett. **22**(1), 177–180 (2018)
7. C. Suescun, M. Cardei, Event-based clustering for composite event detection in wireless sensors networks. IEEE 35th International Performance Computing and Communications Conference (IPCCC), Las Vegas, pp. 1–8 (2016)
8. B. Kitchenham, O. Brereton, D. Dudgen, M. Turner, J. Bailey, S. Linkman, Systematic literature review in software engineering—a systematic literature review. Inf. Softw. Technol. **51**(1), 7–15 (2009)
9. K. Biswas, V. Muthukkumarasamy, E. Sithirasenan, K. Singh, An energy aware heuristic-based routing protocol in Wireless Sensor Networks. 17th International Conference on Computer and Information Technology (ICCIT), Dhaka, pp 53–58 (2014)
10. M. Razzaq, D. Ningombam, S. Shin, Energy efficient K-means clustering-based routing protocol for WSN using optimal packet size, in International Conference on Information Networking (ICOIN), Chiang Mai, pp. 632–635 (2018)
11. L. Devi, A. Rao, Optimization of energy in wire- less sensor networks using clustering techniques, in Communication and Electronics Systems (ICCES), Coimbatore, pp. 1–4 (2016)
12. T. Huynh, A. Dinh, C. Tran, Delay-constrained energy-efficient cluster-based multi-hop routing in wireless sensor networks. J. Commun. Netw. **18**(4), 580–588 (2016)
13. Y. Istwal, S. Verma, Dual cluster head routing protocol in WSN. 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, pp 1–6 (2017)
14. W. Abushiba, P. Johnson, S. Alharthi, C. Wright, An energy efficient and adaptive clustering for wireless sensor network (CH-leach) using leach protocol, in 13th International Computer Engineering Conference (ICENCO), Cairo, pp. 50–54 (2017)
15. P. Khandor, A. Chand, Reactive energy efficient hetero- geneous clustering based routing protocol for WSN. International Conference on Wireless Communication, Signal Processing and Networking (WiSPNET), Chennai, pp. 22–24 (2017)
16. W. Wu, N. Xiong, C. Wu, Improved clustering algorithm based on energy consumption in wireless sensor networks. IET Netw. **6**(3), 47–53 (2017)
17. W. Zhang, L. Li, G. Han, L. Zhang, E2HRC: an energy-efficient heterogeneous ring clustering routing protocol for wireless sensor networks. IEEE Access **5**, 1702–1713 (2017)
18. B. Sun, D. Li, A comprehensive trust aware routing protocol with multi-attributes for WSN. IEEE Access **6**, 4725–4741 (2018)

19. A. Khan, K. Saghar, R. Ahmad, A. Kiani, RAEED-EA: a formally analyzed energy efficient WSN routing protocol. 13th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, pp 346–349 (2016)
20. N. Kumar, Y. Singh, An energy efficient and trust management based opportunistic routing metric for wireless sensor net- works, in Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), Waknaghat, pp. 611–616 (2016)
21. V. Bawage, D. Mehetre Energy efficient Secured Routing model for wireless sensor networks. International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), Pune, pp. 865–869 (2016)
22. N. Khan, K. Saghar, R. Ahmad, A. Kiani, Achieving energy efficiency through load balancing: A comparison through formal verification of two WSN routing protocols, in 13th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, pp. 350–354 (2016)
23. N. Kumar, Y. Singh, Trust and packet load balancing based secure opportunistic routing protocol for WSN, in 4thInternational Conference in Signal Processing, Computing and Control (ISPCC), Solan, pp. 463–467 (2017)
24. B. AnandaKrishna, N. Madhuri, M. Rao, B. VijaySekar, Implementation of a novel cryptographic algorithm in Wireless Sensor Networks. Conference on Signal Processing And Communication Engineering Systems (SPACES), Vijayawada, pp. 149–153 (2018)
25. R. Vadlamudi, S. Umar, A review of APTEEN in wireless sensor networks. Int. J. Sci. Eng. Comput. Technol. **3**(9), 306 (2013)
26. K. Sharma, M. Ghose, Security model for hierarchical clustered wireless sensor networks. Int. J. Comput. Sci. Secur. IJCSS **5**(1), 85–97 (2011)
27. A. Ahmed, M. Younis, A survey on clustering algorithms for wireless sensor. Comput. Commun. **30**, 2826–2841 (2007)
28. S. Lindsey, C. Raghavendra, PEGASIS: Power- efficient gathering in sensor information systems, In Proceedings, IEEE Aerospace Conference, Big Sky, pp. 1–6 (2002)
29. O. Younis, S. Fahmy, HEED: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks. IEEE J. Mag. **3**(4), 366–379 (2004)
30. Y. Li, L. Guo, S. Prasad, An Energy-Efficient Distributed Algorithm for Minimum-Latency Aggregation Scheduling in Wireless Sensor Networks. IEEE 30th International Conference on Distributed Computing Systems, Genova, pp. 827–836 (2010)
31. T. Huynh, C. Hong, An Energy Delay Efficient Multi-Hop Routing Scheme for WSN transactions on information and systems (2006)
32. B. Shi, W. Zhang, G. Xue, J. Tang, C. Wang, DEAR: delay-bounded energy-constrained adaptive routing in wireless sensor networks. Proceedings IEEE INFOCOM, Orlando, pp. 1593–1160 (2012)
33. H. Ammari, On the energy-delay trade-off in geographic forwarding in always-on wireless sensor networks: a multi-objective optimization problem. Sci. Direct. **57**(9), 1913–1935 (2013)
34. J. Shen, A. Wang, C. Wang, Y. Ren, J. Wang, Performance comparison of typical and improved LEACH protocols in wireless sensor network. First International Conference on Computational Intelligence Theory, Systems and Applications (CCITSA), Yilan, pp. 187–192 (2015)
35. M. Handy, M. Haase, D. Timmermann, Low energy adaptive clustering hierarchy with deterministic cluster-head selection, in 4th International Workshop on Mobile and Wireless Communications Network, Stockholm, pp. 368–372 (2002)
36. W. Abushiba, P. Johnson, S. Alharthi, W. Wright, An energy efficient and adaptive clustering for wireless sensor network (CH-leach) using leach protocol, in 13th International Computer Engineering Conference (ICENCO), Cairo, pp. 50–54 (2017)
37. J. Zheng, A. Jamalipour, Energy efficiency and power control, in Wireless sensor network a networking perspective (eds), pp. 307–337, Wiley-IEEE Press (2009)
38. C. Ioannou, V. Vassiliou, C. Sergiou, an intrusion detection system for wireless sensor networks, in 24th International Conference on Telecommunications (ICT), Limassol, pp. 1–5 (2017)
39. A. Diop, Qi., Q. Wang, S. Hussain, An advanced survey on secure energy-efficient hierarchical routing protocols in wireless sensor networks. Int. J. Comp. Sci. **10**(1), 1–11 (2013)

40. U. Patel, S. Modi, B. Suma, A survey: MAC layer protocol for wireless sensor networks. Int. J. Emerge Technol. Adv. Eng. **3**(9), 203–211 (2013)
41. A. Patle, N. Gupta, Vulnerabilities, attack effect and different security scheme in WSN: a survey, in International Conference on ICT in Business Industry & Government (ICTBIG), pp. 1–6 (2016)
42. A. Pathan, L. Hung, H. Chong, Security in wireless sensor networks: issues and challenges. 8thInternational Conference Advanced Communication Technology, Phoenix Park, 2006, pp. 1048 (2016)
43. S. Avancha, A holistic approach to secure sensor networks (2005)
44. Algorithms and Protocols for Wireless Sensor Networks | Wiley Online Book (2008)
45. G. Smaragdakis, I. Matta, A. Bestavros, SEP: a stable election protocol for clustered heterogeneous wireless sensor networks. Boston University Computer Science Department (2004)
46. F. Aderohunmu, J. Deng, An enhanced stable election protocol for WSN, Department of Information Science, New Zealand (2009)
47. N. Javaid, S. Mohammad, K. Latif, U. Qasim, Z. Khan, M. Khan, HEER: Hybrid energy efficient reactive protocol for wireless sensor networks, in Electronics, Communications and Photonics Conference (SIECPC), Saudi International, pp. 1–4 (2013)
48. T. Sharma, B. Kumar, G. Tomar, Performance Comparison of LEACH, SEP and DEEC Protocol in Wireless Sensor Network (2012)
49. L. Mahajan, N. Sharma, Improving the stable period of WSN using dynamic Stable leach Election Protocol, in International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), pp. 393–400 (2014)
50. X. Zheng, Y. Feng, J. Yang, Z. Lu, Energy efficient and fault tolerant routing protocol for mobile wireless sensor networks, in International Conference on Information and Communications Technologies (ICT), pp. 1–5 (2014)
51. Y. Li, C. Ai, C. Vu, Y. Pan, R. Beyah, Delay-bounded and energy-efficient composite event monitoring in heterogeneous wireless sensor networks. IEEE Trans. Parallel Distrib. Syst. **21**(9), pp. 1373–1385 (2010)
52. K. Romer, F. Mattern, Event-based systems for detecting real-world states with sensor networks: a critical analysis. Proceedings of the Intelligent Sensors, Sensor Networks and Information Processing Conference, Melbourne, pp. 389–395 (2004)
53. C. Suescun, M. Cardei, Event-based clustering for composite event detection in wireless sensors networks, in IEEE 35th International Performance Computing and Communications Conference (IPCCC), Las Vegas, pp. 1–8 (2016)
54. K. Liu, N. Abu-Ghazaleh, K. Kang, Location verification and trust management for resilient geographic routing. J. Parallel Distrib. Comput. **67**(2), 215–228 (2007)
55. S. Biswas, R. Morris, Ex-OR: opportunistic multi-hop routing for wireless networks, in Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, New York, NY, USA, 35(4), pp. 133–144 (2005)
56. V. Mohindru, Y. Singh, Efficient approach for securing message communication in wireless sensor networks from node clone attack. Indian J. Sci. Technol. (9), 32 (2016)
57. P. Gong, T. Chen, Q. Xu, ETARP: an energy efficient trust-aware routing protocol for wireless sensor networks. J. Sens. **7**, 1–10 (2015)
58. A. Ahmed, K. Bakar, I. Channa, A. Khan, A secure routing protocol with trust and energy awareness for wireless sensor network. Mob. Netw. Appl. **21**(2), 272–285 (2016)
59. H. Liu, B. Zhang, H. Mouftah, X. Shen, J. Ma, Opportunistic routing for wireless ad hoc and sensor networks: Present and future directions. IEEE Commun. Mag. **47**(12) (2009)
60. S. Das, Filters, wrappers and a boosting-based hybrid for feature selection. In Icml **1**, 74–81 (2001)
61. Z. Zhang, X. Chu, Risk prioritization and failure mode and effects analysis under certainty, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, PR China (2011)
62. H. Chen, Z. Han, Z. Fu, *Quantitative trustworthy evaluation scheme for trust routing scheme in wireless sensor network* (IEEE Trustcom/BigDataSE/ISPA, Helsinki, Finland, 2015), pp. 20–24

63. M. Zhang, R. Zheng, Q. Wu, W. Wei, X. Bai, H. Zhao, B-iTRS: a bio-inspired trusted routing scheme for wireless sensor networks. J. Sens. (5), 1–8 (2015)
64. Z. Chen, M. He, W. Liang, K. Chen, Trust-aware and low energy consumption security topology protocol of wireless sensor network. J. Sens. **6**(201), 1–10 (2015)
65. Z. Hu, Y. Bie, H. Zhao, Trusted tree- based trust management scheme for secure routing in wireless sensor networks. Int. J. Distrib. (2 015)
66. M. Yarvis, N. Kushalnagar, H. Singh, A. Rangarajan, Y. Liu, S. Singh, Exploiting heterogeneity in sensor networks, in INFOCOM 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE, vol. 2, pp. 878–890
67. C. Karloff, D. Wagner, Secure routing in wireless sensor networks: attacks and countermeasures. Proceedings of the First IEEE International Workshop on Sensor Network Protocols and Applications, Anchorage, pp. 113–127 (2003)
68. M. Jan, P. Nanda, X. He, R. Liu, A Sybil attack detection scheme for a centralized clustering-based hierarchical net- work. Trustcom/BigDataSE/ISPA IEEE **1**, 318–325 (2015)
69. V. Pamula, A. Tripartite, Performance of selection combiner over Nakagami fading channel with Laplace Cochannel Interference, IEICE Commun. Express, pp. 94 (2018)
70. I. Dhillon, Y. Guan, B. Kulis, Drawback of k-mean, Proceedings of the tenth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining, New York, NY, US (2004)
71. K. Saghar, Formal modelling and analysis of denial of services attacks in wireless sensor networks, Ph.D. Thesis, Northumbria University (2010)
72. A. Sari, E. Caglar, Load balancing algorithms and protocols to enhance quality of service and performance in data of wsn, in Security and Resilience in Intelligent Data-Centric Systems and Communication Networks, Elsevier, pp. 143–178 (2018)
73. J. Li, A symmetric cryptographic algorithm in wsn security. Int. J. Online Eng. IJOE **13**(11), 102–110 (2017)
74. P. Mohanty, S. Panigrahi, N. Sarma, Security issues in wireless sensor network data gathering protocols: a survey. J. Theoret. Appl. Inform. Technol. Tezpur (2010)
75. J. Kelsey, B. Schneier, D. Wagner, *Hall* (C. Cryptanalytic attacks on pseudorandom number generators, In International Workshop on Fast Software Encryption, 1998), pp. 168–188
76. Y. Liu, Q. Wu, T. Zhao, Y. Tie, F. Bai, M. Jin, An improved energy-efficient routing protocol for wireless sensor networks. Sensors **19**(20), 4579 (2019)

Perturbation by Sybil Attack in Clustering for Open IVC Networks (COIN) Protocol—A Protocol in Cluster-Based Routing Category for Infrastructure-Less VANETs

Nishtha, Devaashish Sharma, and Manu Sood

Abstract The infrastructure-less VANETs, for the execution of myriad applications associated with moving vehicles, involve diverse routing protocols grouped into five different categories to fulfill their distinguished requirements. One such category of protocols is the cluster-based routing protocols category, in which the lack of centralized infrastructure for management in infrastructure-less VANETs is overcome by a single selected vehicle known as cluster head in a group of vehicles in a cluster. This cluster head is accountable for managing vehicle-related functions within as well as outside the cluster. Regardless, the infrastructure-less VANET lacks a dedicated infrastructure for controlling the network and is prone to several security attacks. Sybil attack is one such disastrous attack where a malicious vehicle can launch one or more superficial or virtual vehicles called Sybil vehicles that appear as legitimate vehicles to all other vehicles engaged in VANET. Sybil attack disturbs various communication mechanisms and separately affects the inherent routing mechanisms by launching Sybil vehicle(s) at some strategically sensitive places as per the prescribed norms of the routing protocol to manage/control prime functioning of the VANET. Clustering for Open IVC Networks (COIN) is a prominent protocol in the cluster-based routing protocols category tailored for infrastructure-less VANETs. This paper examines whether the Sybil attack can perturb the operation of the COIN protocol. Our experimental design and setup supported by different simulations have confirmed that the Sybil attack can consistently disrupt the routing mechanism in COIN, allowing various malicious activities in the VANET.

Keywords Sybil attack · Cluster-based routing protocols · COIN

Nishtha (✉)
Department of Computer Science, RGGDC, Chaura Maidan, Shimla, India

D. Sharma
Department of Computer Science and Engineering, NSUT, New Delhi, India

M. Sood
Department of Computer Science, Himachal Pradesh University, Shimla, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_27

1 Introduction

Vehicular Ad hoc Network (VANET) is a temporary setup that uses wireless interconnection among traveling vehicles within their radio range to render vehicular safety, accident prevention, traffic monitoring, and other infotainment services to vehicle users [1–4]. VANETs may be deployed using three different architectures: one is an infrastructure-based architecture where infrastructures such as roadside units (RSUs) are exclusively present for managing vehicles in their neighborhood. The vehicles in this architecture invariably follow the Vehicle-to-Infrastructure (V2I) communication mode. Second is the infrastructure-less architecture deficient in dedicated custom-built infrastructure for network control. In this architecture, packet forwarding from a sending vehicle to the target vehicle(s) in a different radio range is settled, utilizing a peer-to-peer approach where the in-between vehicles simultaneously act as host and router [5]. Sending vehicles in a one-to-one manner forward the packets to intermediate vehicles until their delivery to the target vehicle. Vehicles utilize a unique Vehicle-to-Vehicle (V2V) mode of communication in this architecture. The third is the hybrid architecture that merges V2V and V2I modes of communication concurrently [4].

Among these three architectures of VANET, the pure ad hoc architecture does not require expensive additional infrastructure, may be deployed in unexpected scenarios such as places of natural disaster, hostile locations, war-fields/ enemy-areas or interior regions of a country such as deserts, forests, mountains [3, 5].

One additional feature is the periodic broadcast of beacons by every participating vehicle in VANET. Beacons are short messages to broadcast status information and help in establishing cooperative and safety applications. These beacons are exchanged among vehicles within their transmission range and are called one-hop or local broadcast [6, 7].

With the development and pervasiveness of VANET, a plethora of vehicular applications with different service demands is emerging for infrastructure-less VANET [6–8]. To meet the objectives of secure, reliable, private, minimum communication delay, and diverse application scenarios, distinguished categories of routing protocols are framed for infrastructure-less VANETs [1, 6–8]. One such category of routing protocols is the cluster-based routing protocols category among the others: topology-based routing protocols category, position-based routing protocols category, broadcast-based routing protocols category, and geocast-based routing protocols category [1, 4, 8]. In all protocols in the cluster-based routing protocols category, vehicles are combined into groups known as clusters. Every routing protocol in this category uses a specific strategy to select a single vehicle known as cluster head (CH). This vehicle acts as a cluster leader in its cluster. The cluster head maintains a cluster table comprising the addresses and geographic locations of all cluster members and gateways. It maintains a Cluster Neighbor Table holding details of adjacent clusters also [4, 9–12]. So, the cluster head is responsible for route maintenance in all inter-cluster communications and intra-cluster management functions. In all these routing protocols, the geographic location of every vehicle is determined

by Geographic Positioning System (GPS). It is achieved because the GPS unit is placed in the On-Board Unit (OBU) of every member vehicle in VANET [6].

In this manner, the cluster-based category of routing protocols attempts to cover up the absence of centralized infrastructure in infrastructure-less VANETs. However, the lack of dedicated infrastructure persists in all routing protocol categories covering the cluster-based routing protocols category tailored for infrastructure-less VANETs.

It makes infrastructure-less VANETs vulnerable to security attacks wherein Sybil attack is one attack, particularly prevalent in peer-to-peer networks without centralized management [2, 3, 13–16]. In this attack, a malicious vehicle entered in VANET publicizes multiple vehicles with several IDs at different locations known as Sybil vehicles, nonexistent in VANET. But, this develops a false impression of the existence of more neighboring vehicles than actually present to all other vehicles in VANET [14, 16]. Launching, demolishing, and entire communication through the Sybil vehicles, in reality, is controlled by the malicious vehicle. Thus, the malicious vehicle may fabricate these Sybil vehicles to disrupt either a part of VANET or the complete VANET [13]. Sybil attack affects the real traffic scenarios leading to catastrophic situations like traffic jams, accidents, additionally manipulates various mechanisms like misbehavior detection, fair allocation of network resources, and routing mechanisms in VANET [13, 17].

Routing protocols are the most vital in all communication networks like VANET, where a network relies on the successful execution of its applications. These applications further depend upon the successful accomplishment of the employed routing protocol [1, 4, 8, 18]. These protocols in VANET facilitate the forwarding of information from one sending vehicle to the target vehicle(s) within a specified time constraint, with characteristics of ultra-reliable and low latency transmission [7]. Any disarrangement in these protocols may destroy the goal behind VANETs deployment. However, with the Sybil attack, the Sybil vehicle positioned within the route stipulated by the routing protocol may disrupt the used protocol in numerous ways. While forwarding the packets from source to target vehicle, the malicious vehicle may truncate, modify, drop, or add erroneous routing information in the packets collected through the in-between Sybil vehicle behaving as the packet forwarding vehicle. Firstly, the malicious vehicle may insert erroneous routing information, and consequently, the data packets will follow a different route and will be delivered to a vehicle, not the target vehicle, or delivered with delayed delivery of the target vehicle. Secondly, this vehicle may alter the routing information. The addition of erroneous data by truncating the data packets or modifying the data packets will provide incorrect information to the target vehicle. Thirdly, it may repeat old routing information, and replaying old information will repeatedly deliver the same information to the target vehicle. Additionally, it may drop the complete packets, and this will forbid any communication to occur in VANET [13, 19].

In this paper, our prime focus is on the cluster-based routing protocols category designed for infrastructure-less VANET. Significant benefits of the cluster-based routing protocols category compared to other categories like a topology-based routing protocols category include that protocols in this category are efficient in accommodating continuous vehicular growth in VANET. These cluster-based routing protocols

category helps in increasing scalability, searching shortest routing path, helps in minimizing communications overhead, solving the vehicle mobility problem within the cluster only, and so is one distinguished category of routing protocols in VANET [5, 12, 20].

Clustering for Open IVC Networks (COIN) is one significant protocol in the cluster-based routing protocols category specially designed for infrastructure-less VANETs. IVC stands for Inter Vehicular Communication. COIN designed for the Highways assumes that all vehicles travel at constant speed on a Highway. It leads to maximum connectivity among vehicles irrespective of whether a vehicle is a cluster head, cluster member, or gateway in a cluster. It also improves cluster stability with the high speed of vehicles. Vehicular dynamics and driver objectives are the main criteria for the selection of a cluster head. A vehicle with the maximum connected vehicles in the cluster is selected as a cluster head. In this protocol, vehicles moving on a Highway in the same lane, same direction, and in the same radio range are neighbor vehicles. In COIN, the position of a vehicle gathered through GPS is utilized for preventing collision among vehicles [11]. All vehicles contributing to VANET have an inbuilt On-Board Unit (OBU) consisting of a GPS unit. However, to initiate a Sybil attack, the spoofing or jamming of GPS signals can invariably be achieved by the malicious vehicle [21].

1.1 Research Gap

Clustering for Open IVC Networks (COIN) protocol is among the most significant protocols within the cluster-based routing protocols category [11]. It is one of the few protocols specially designed for infrastructure-less VANETs. The literature review of relevant papers has suggested that none of the researchers have so far discussed the perturbation caused by the Sybil attack in COIN or other routing protocol in this category framed for infrastructure-less VANETs. Sybil attack is a dangerous attack that may disrupt the routing mechanisms in infrastructure-less VANETs. As already discussed, the cluster-based routing protocols category, with numerous advantages, is one of the distinguished categories for infrastructure-less VANETs. The infrastructure-less deployments of VANETs are deprived of centralized infrastructure for network management, and the protocols in the cluster-based category have every possibility of being affected by the Sybil attack. In the literature survey, we can add that no researcher has ever discussed the effect of the Sybil attack on cluster-based protocols in VANETs. The objective of the paper is to investigate the possibility of a Sybil attack on cluster-based routing protocols. We have considered COIN, a protocol in this category designed particularly for infrastructure-less VANETs. In this paper, through different experiments, it has been concluded that there is every likelihood of a Sybil attack to disrupt the infrastructure-less VANETs using COIN. This paper is a novel one as it is the first one to confirm the perturbation by Sybil attack in COIN-based VANET.

The rest of the paper is arranged subsequently in the following manner. Section 2 represents Sybil attack using COIN for infrastructure-less VANETs. Section 3 talks about the methodology used to initiate and examine the impact of the Sybil attack on COIN. In Sect. 4, we have discussed the experimental setup for launching a Sybil attack on a COIN-based VANET using simulation on NS-2. Section 5 examines the simulation results, and Sect. 6 covers the conclusion and future work.

2 Sybil Attack Using COIN for Infrastructure-Less VANETs

We have considered an infrastructure-less VANET topology with thirteen vehicles to demonstrate the effect of the Sybil attack using COIN protocol. Let these vehicles are set in two clusters with five vehicles in the Cluster-C1 and six vehicles in the Cluster- C2, as displayed in Fig. 1.

In COIN, a cluster head is a vehicle in each cluster having maximum weight. Vehicle *A* is a vehicle with the highest weight in Cluster-C2, and vehicle *S* is a vehicle with the highest weight in Cluster-C1. At a given instance of time Vehicle *A* in Cluster-C2, and vehicle *S* in Cluster-C1, selected as the cluster heads. In this topology, an entered malicious vehicle *M* has initiated a Sybil attack by setting a Sybil vehicle *S*. It is achieved by *M* through GPS spoofing and by transmitting beacons periodically at different locations. The specific purpose of the malicious vehicle is to generate the Sybil vehicle(s) at such a position so that there are maximum chances of the Sybil vehicle to become the cluster head and to be able to control the entire cluster. The exact location to launch *S* by *M* has to be one where *S* gets associated with a weight that is the maximum number of vehicles compared to all others existing

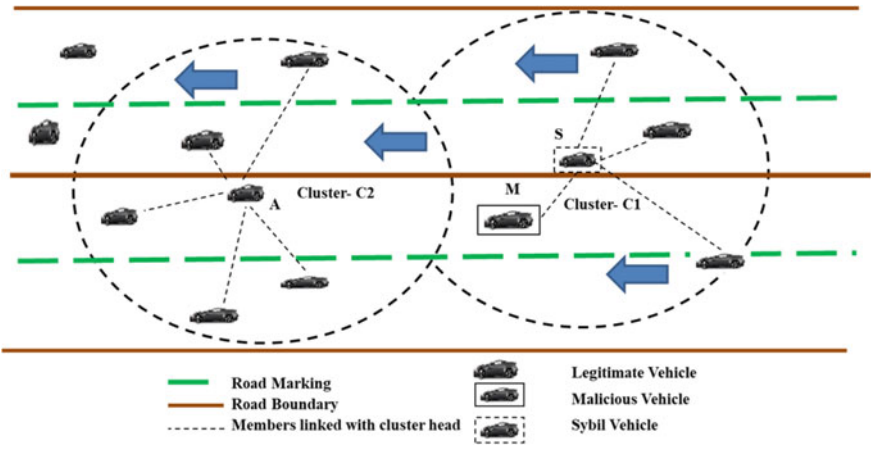


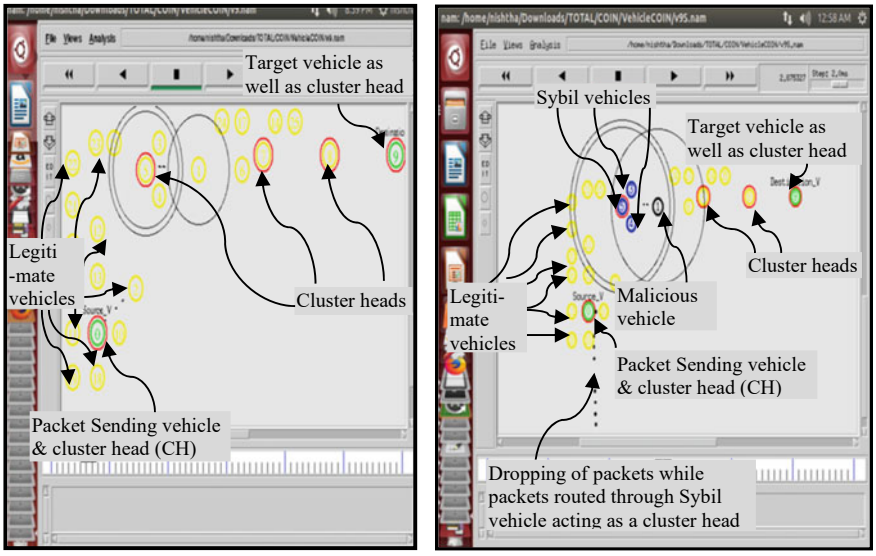
Fig. 1 VANET with Clustering for Open IVC Networks (COIN) protocol under Sybil attack [11]

in the Cluster-C1. In other words, vehicle *S* compared to all other vehicles has the highest one-hop neighbors in Cluster-C1 and is chosen as a cluster head. With the introduction of the Sybil attack through the launch of a Sybil vehicle *S,* managing it to act as a cluster head, it becomes possible for the malicious vehicle *M* to completely manage and control inter-vehicular as well as intra-vehicular communications in the Cluster-C1. It shows that the Sybil attack has entirely disrupted the COIN protocol.

3 Methodology Used

The impact of the Sybil attack in infrastructure-less VANET employing COIN is examined with two different simulation scenarios having similar topology and parameters like equal speed, direction, and similar simulation area for performing simulation experiments.

The point of difference between the two scenarios: in the 1st every vehicle behaves legitimately inclusive of the malicious vehicle in the VANET, as displayed in Fig. 2a, but, in the 2nd scenario, the malicious vehicle has launched the Sybil attack by creating and positioning *3* Sybil vehicles in VANET, as displayed in blue in Fig. 2b. In 1st scenario, the packets routing from the sending vehicle to the target vehicle



(a) **(b)**

Fig. 2 **a** COIN with no Sybil attack. **b** COIN under Sybil attack Total mobile vehicles engaged in this VANET topology = 26, Area = 2000 m *2000 m, Packet Size = 1024B, Speed of each vehicle = 1 m/s, Simulation Time = 6 s

occurs according to the COIN protocol, and the packets reach the target vehicle without any obstruction, as displayed in Fig. 2a.

The cluster head in their respective cluster in both the simulation scenarios is displayed as a vehicle encircled in red color. The malicious vehicle has launched these Sybil vehicles at a position where any one of these three Sybil vehicles is a vehicle with the highest weight in the cluster. The Sybil vehicle behaving as a cluster head enables the malicious vehicle to manage all inter-vehicular and intra-vehicular communication in that cluster. In all simulations under the Sybil attack, we have considered the worst case of disruption by the Sybil attack. In this scenario, packet forwarding from the sending towards the target vehicle routed through an intermediate cluster in which a Sybil vehicle is acting as a cluster head, the malicious vehicle, in that case, does not further forward the packets to the next vehicle according to the standards followed by COIN. The malicious vehicle drops the complete packets and prevents them to reach the target, prohibiting absolute communication in the VANET.

Two simulation scenarios have been designed and implemented to assess the behavior of Sybil attack in COIN-based VANET while changing: (1) size of the packets, (2) vehicle speed, and (3) the number of vehicles involved in the VANET. These scenarios are: (a) 1st, where all vehicles participating in the VANET topology are legitimate vehicles as shown in Fig. 2a, and 2nd, where a Sybil attack is implemented with the launch of three Sybil vehicles, as displayed in Fig. 2b.

4 Experimental Setup

Network Simulator (NS-2.35) was set up on the Ubuntu 14.04.6 LTS Operating System which was augmented by a virtual machine running on VMware Workstation 10 for performing simulation experiments. This network simulator is a specialized distinct event simulator for executing wireless simulations. Vehicles varied from 10 to 28 for carrying a different set of experiments. The packet size is changed from 512 to 9216B under different simulation scenarios. The transmission range is set as 250 m. These vehicles travel with a predefined starting position to the final target in an area of 2000 m X 2000 m for all simulation scenarios. These vehicles travel with a speed of 1 m/s to 10 m/s for various sets of simulations. All other NS2 parameters are kept at their preset values. The outcome of the executed simulation was further analyzed by a developed awk file that uses the respective trace file(s).

5 Simulation Results

The evaluation of the Sybil attack using COIN in VANET is accomplished by using varied simulation programs created for the *2* scenarios, one with no Sybil attack and another under the execution of the Sybil attack. Both these scenarios are executed

with the same set of parameters keeping out three cases: packet size, the speed of each vehicle, and the number of vehicles in VANET. All the simulations are performed for both the scenario (i) 'with no Sybil attack' and 'under Sybil attack' for three varied cases: (i) Throughput & PDR Versus Packet Size, (ii) Throughput & PDR Versus Speed of Vehicles, and (iii) Throughput & PDR Versus Number of Vehicles.

Throughput is defined as the number of packets sent to the target in one unit of time whereas, PDR is defined as the ratio of packets sent by the sender vehicle to the number of packets received by the target vehicle. Hence, Throughput and PDR are the performance parameters in VANET to examine both scenarios 'with no Sybil attack' and 'under Sybil attack' for variation in *3* exclusive cases (i) the packet size, (ii) the speed of each vehicle, and (iii) third is the number of vehicles.

Case 1: To examine the impact of the Sybil attack in VANET using COIN protocol, the values of the packet size have been changed in the 10 simulation programs implemented for each of the two scenarios 'with no Sybil attack' and 'under Sybil attack'. These values for each of the 10 programs were 512B, 1024B, 2048B, 3072B, 4096B, 5120B, 6144B, 7168B, 8192B and 9216B. The 1st scenario is executed with 10 genuine vehicles, and in the 2nd scenario, the Sybil attack is implemented with the same 10 vehicles consisting of one malicious vehicle and three Sybil vehicles. The experimental results attained for the Throughput and PDR have been displayed in Fig. 3a, b.

Case 2: To study the effect of this attack with the change in speed of vehicles using COIN protocol, 20 different simulations for the two different scenarios are performed by escalating the speed of all the vehicles by one unit in the range of 1 to 10 m/s. There

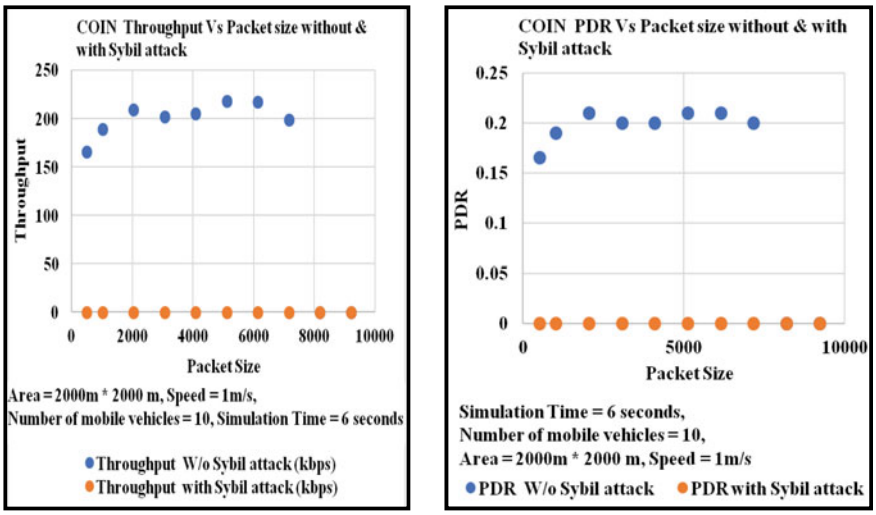


Fig. 3 **a** Throughput versus the packet size (with no & under Sybil attack). **b** PDR versus the packet size (with no & under Sybil attack)

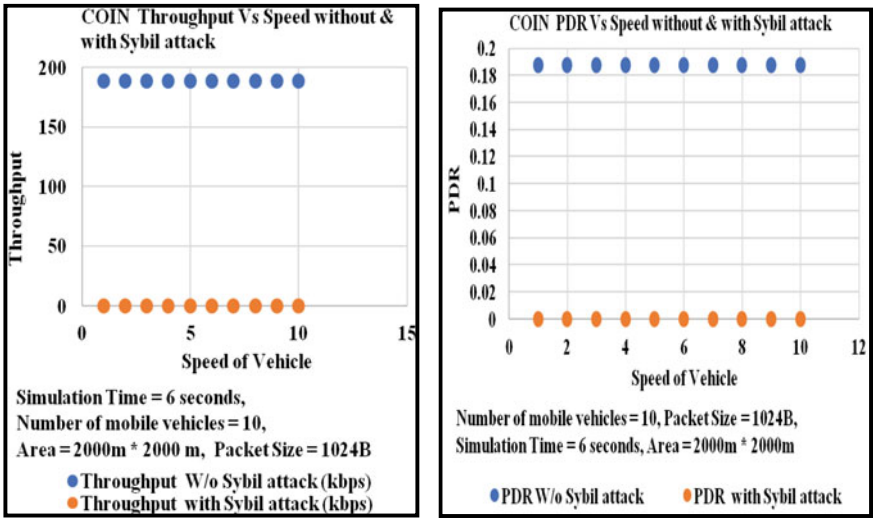
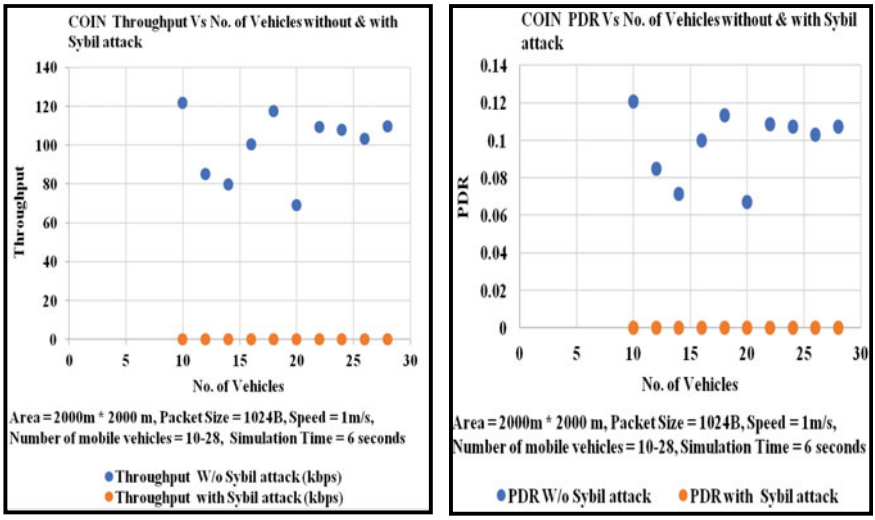


Fig. 4 **a** Throughput versus speed of each vehicle (with no & under Sybil attack). **b** PDR versus speed of each vehicle (with no & under Sybil attack)

are 10 legitimate vehicles in the 1st simulation scenario and 10 vehicles comprising one malicious vehicle, and three Sybil vehicles in the 2nd scenario. Performance analysis for both the scenarios has been worked out with calculation of Throughput and PDR, shown in Fig. 4.

Case 3: The impact of the Sybil attack with the change in the number of vehicles in the VANET using COIN is investigated with the execution of 20 simulation programs. The number of the participating vehicle in the VANET topology for both scenarios is altered from 10 in the 1st program, adding two in each successive program, reaching 28 in the last program.

All graphs represent that changing the size of packets, changing the speed, or increasing the number of participating vehicles can never prevent a Sybil attack in a VANET using COIN protocol. However, with a malicious vehicle intending to launch a Sybil attack, it is ever possible for it to launch a Sybil attack and disrupt the communication in the VANET. In all simulation scenarios using COIN 'Under Sybil attack' while forwarding the packets from the sending to the target, the malicious vehicle has launched one out of the three Sybil vehicles as a cluster head of an intermediate cluster. In this scenario, at any time when the packets are forwarded through this cluster head, the malicious vehicle drops the entire packets, obstructing communication in VANET. On the top, it can be observed that throughput and PDR have a relationship of direct proportionality. Also, a closer look at the figures shows that the variation in the speed of the participating vehicles does not affect the Sybil attack in a COIN-based infrastructure-less VANET (Fig. 5).



(a) **(b)**

Fig. 5 **a** Throughput versus number of vehicles (with no & under Sybil attack). **b** PDR versus number of vehicles (with no & under Sybil attack)

6 Conclusion and Future Work

In the existing paper, employing COIN, a protocol in the cluster-based routing protocols category for infrastructure-less VANETs, we have performed simulations on NS-2 for the two scenarios, one with no Sybil attack and another with the Sybil attack implementations. In all simulations, with the Sybil attack implementation, the malicious vehicle has launched a Sybil attack with the initiation of three Sybil vehicles. The malicious vehicle, in reality, is the controlling physical vehicle for its three framed Sybil vehicles. In all simulations, whenever packets from source to the target are advanced through an in-between cluster managed by the malicious vehicle to manifest Sybil vehicle a cluster head, it completely dropped the packets received through the Sybil vehicle. These packets from the specified cluster were not routed to the next forwarding vehicle as per COIN specifications and did not arrive at the target, ceasing complete communication in VANET. It has confirmed that the Sybil attack is a peculiar, most deadly attack capable of completely disrupting the network communication in an infrastructure-less VANET. Such disruption by a Sybil attack withdraws the purpose behind VANETs deployment. Thus, the infrastructure-less VANETs require a mechanism to protect them against Sybil attacks. In the future, we are working on mechanisms to safeguard infrastructure-less VANETs using cluster-based routing protocols from the adverse effects of the Sybil attack. This work will help the researchers to develop detection and prevention mechanisms for Sybil attack for protocols in the cluster-based routing protocols category of VANETs, and at the same time to design new protocols in this category that are secure from the Sybil attack.

References

1. J. Liu, J. Wan, Q. Wang, P. Deng, K. Zhou, Y. Qiao, A survey on position based routing for vehicular ad hoc networks. Telecommun. Syst. **62**, 15–30 (2016)
2. M.S. Al-Kahtani,Survey on security attacks in vehicular ad hoc networks (VANETS), in 6th International Conference on Signal Processing and Communication Systems, pp. 1–9. IEEE, Gold Coast (2012)
3. Nishtha, M. Sood, Exploring the effects of sybil attack on pure ad hoc deployment of vanet, inP.K. Singh, A.K. Kar, Y. Singh, M.H. Kolekar, S. Tanwar (eds) 2nd ICRIC 2019, LNEE, vol. 597, pp. 349–363. Springer, Cham (2019)
4. F. Li, Y. Wang, Routing in vehicular ad hoc networks: a survey. vehicular technology magazine. IEEE **2**(2), 12–22 (2007)
5. M.M. Nasr, A.M. Abdelgader, Z.G. Wang, L.F. Shen, Vanet clustering based routing protocol suitable for deserts. Sensors (Basel, Switzerland) **16**(4), 478 (2016)
6. F. Papadimitratos, A.D.L. Fortelle, K. Evenssen, R. Brignolo, S. Cosenza, Vehicular communication systems: enabling technologies, applications, and future outlook on intelligent transportation. Commun. Mag. IEEE **47**(11) 84–95 (2009)
7. S.A.A. Shah, E. Ahmed, M. Imran, S. Zeadally, 5g for vehicular communications. Commun. Mag. IEEE **56**(1), 111–117 (2018)
8. E.C. Eze, S. Zhang, E. Liu, J.C. Eze, Advances in vehicular ad-hoc networks (vanets): challenges and road-map for future development. Int. J. Autom. Comput. **13**(1), 1–18 (2016)
9. S. Hosmani, B. Mathapati, Efficient vehicular ad hoc network routing protocol using weighted clustering technique. Inf. J. Inform. Technol. **13**, 469–473 (2021)
10. Y. Luo, W. Zhang, Y. Hu, A new cluster based routing protocol for vanet, in 2nd International Conference on Networks Security, Wireless Communications and Trusted Computing 2010, pp. 176–180, IEEE, Wuhan, China (2010)
11. J. Blum, A. Eskandarian, L. Hoffman, Mobility management in ivc networks. Intell. Veh. Symposium, IEEE, Columbus, USA (2003)
12. M.A. Mujahid, K.A. Bakar, T.S.J. Darwish, F.T. Zuhra, Cluster-based location service schemes in vanets: current state, challenges and future directions. Telecommun. Syst. **76**, 471–489 (2021)
13. Nishtha, M. Sood, Exploring the possibility of sybil attack in position based routing protocols in vanets: a case study of greedy perimeter coordinate routing (gpcr), in U. Batra, N. Roy, B. Panda (eds.) REDSET-2019, CCIS, vol. 1230, pp. 79–89, Springer, Singapore (2020)
14. J.R. Douceur, The sybil attack, in P. Druschel, M.F. Kaashoek, A.I.T. Rowstron (eds.) 1st International Workshop on Peer-to-Peer Systems, pp. 251–260, Springer, London, UK (2002)
15. C. Harsch, A. Festag, P. Papadimitratos, Secure position-based routing for vanet, in 66th Vehicular Technology Conference, pp. 26–30, IEEE, USA (2007)
16. K. Malathi, R. Manavalan, Detection and localization of sybil attack in vanet: a review. Int. J. Res. Appl. Sci. Eng. Technol. **2**(ix), 282–293 (2014)
17. J. Newsome, E. , Shi, D. Song, A. Perrig, The sybil attack in sensor networks: analysis & defenses, in 3rd International Symposium on Information Processing in Sensor Networks, pp. 259–268, IEEE, Berkeley (2004)
18. Y.S. Devi, M. Roopa, Performance analysis of routing protocols in vehicular adhoc networks. Mater. Today: Proc. (2021)
19. L. Zhou, Z.J. Haas, Securing ad hoc networks. IEEE Netw. **13**(6), 24–30 (1999)
20. S. Ebadinezhad, Design and performance evaluation of improved dfaco protocol based on dynamic clustering in vanets. SN Appl. Sci. **3** (2021)
21. R.H. Mitch, R.C. Dougherty, M.L. Psiaki, S.P. Powell, B.W. O'Hanlon, J.A. Bhatti, T.E. Humphreys, Signal characteristics of civil gps jammers, in 24th International technical meeting of the satellite division of the institute of navigation. pp. 1907–1919, Portland, OR (2011)

Image Processing and Computer Vision

Neuromorphic Computing: Review of Architecture, Issues, Applications and Research Opportunities

Hitesh Vora, Preeti Kathiria, Smita Agrawal, and Usha Patel

Abstract In this new enhancing world of computer, it is required to have a high-speed computer at cheap cost and small size. As the AI, ML and Big Data is enhancing the need of faster computer it becomes a necessity. One of the solutions for faster computers is Neuromorphic computers. This is the special kind of computer which are inspired by the biological concept of the human brain. It works like the human brain works. The human brain has the capacity of doing 1 exaFlop calculation per second, which is far better than the capacity of supercomputers. Taking inspiration from the human brain now Intel and IBM is trying to make a computer that can function like the human brain. This device known as the neuromorphic computer will start the new era of computers and AI. This paper discusses the comparison of Neuromorphic computing with Von Neumann computing, algorithms used for neuromorphic computing, concepts of neuromorphic computing, building blocks, application, and research opportunities. Here some research opportunities and applications is discussed in which this new system can work perfectly. This paper clarifies the view on Neuromorphic computing.

Keywords Synapses · Neuron · Axon · Biology · Neural network · Von Neumann · Spiking neural network

H. Vora · P. Kathiria (✉) · S. Agrawal · U. Patel
Institute of Technology, Nirma Universtiy, Ahmedabad, India
e-mail: preeti.kathiria@nirmauni.ac.in

H. Vora
e-mail: 19MCA171@nirmauni.ac.in

S. Agrawal
e-mail: smita.agrawal@nirmauni.ac.in

U. Patel
e-mail: ushapatel@nirmauni.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_28

1 Introduction

Starting from the first computer to the supercomputer like Summit AI one of the world's fastest supercomputers, the technology enhances in every stage. As the technology is growing the device is becoming more and more powerful, small in size, and less power consumption. Humans always try to make a computer device that can work like a human itself. But this kind of computer/robot needs devices that are quick in the calculation and small in size. The technology is developing in the field of AI (Artificial Intelligent), ML (Machine Learning), DL (Deep Learning), Big Data, and Space Science all this field needs a fast device with small size and less power consumption.

At this point in time summit is one of the most powerful computers created by the human. It has the capacity of 200 quadrillion calculations per second. To do it, it has 9216 IBM power9 CPU, 25 Gigabytes per second between nodes speed, 27,648 NVIDIA Tesla GPUs. But to handle all these things it needs a big space and 15,000 litter water per minute to maintain the cooling and 13 Mega Watt power [1]. It is not possible to use in space/satellite or at any other place like robotics. In the field of Big data Analytics it is required to have a quick device which can work on multiple un-structured data and respond quickly. As the solution to this kind of problem neuromorphic computing becomes blessing. For the solution of this, it needs to look at nature. The main motive of this paper is to define some applications and issues of neuromorphic computing that need to have some research for development of realistic chip.

2 Why Need to Look at Biology

Nature designs each and everything very mysteriously. When it's come to biology it becomes too mysterious and too unique. One of the most complex and powerful organs of humans is the "Brain". Whatever thing is created by a human is just because of the brain. As biology understands the brain and its power, it has Billions of Neurons (in some study it shows that 100 billion). 86 Billion neurons mean 1 Quadrillion synapses and 4,00,000 km Neuer fiber which is enough to reach the moon. With specification human brain can do 1 exaFlop calculation per second. And to do it human brain required only 20 watts of power which is less than the power required by a single light bulb [2]. This is the reason to look at biology for the new and fast devices like neuromorphic computers (Fig. 1).

The Von Neumann architecture is the most basic architecture on which all the computers are designed. This architecture has some limitations and this limitation is low computation power. In this architecture, the computer has 4 units, a control unit, ALU (Arithmetic and Logical Unit), RAM (Random Access Memory), and an input–output unit. Its main weakness is that it has memory separation with the control unit and ALU [6]. Due to this problem, more time is invested in controlling and data

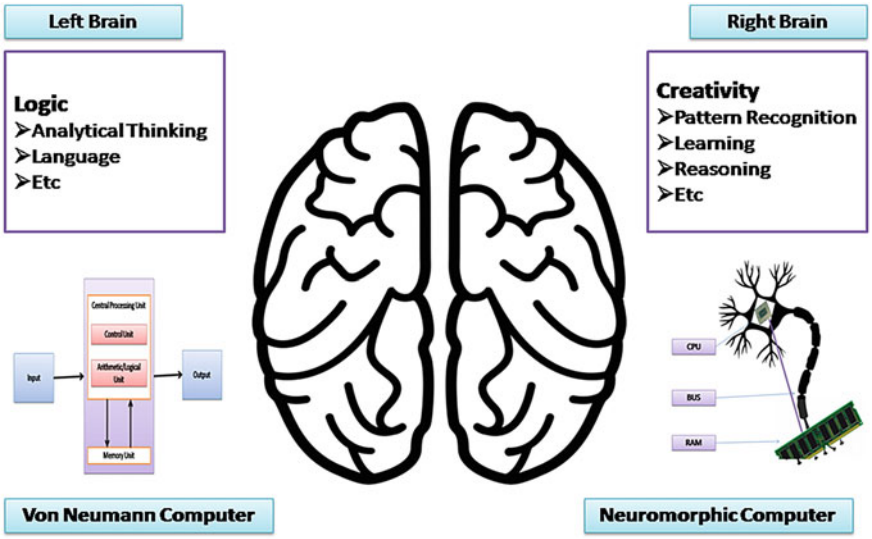


Fig. 1 Brain and computer architecture [3–5]

transferring rather than computing. Moreover in this kind of architecture transistors have only 3 connections. This kind of computer is suitable for calculation purposes but it is not effective where pattern recognition, reasoning, and learning are needed. For that purpose neuromorphic computers are required.

3 Neuromorphic Computer Architecture

This kind of computer adopts the new architecture which is similar to the brain architecture. The human brain has neurons, which are mainly divided into the 3 parts, CELL BODY, AXON, and SYNAPSES; here CELL BODY is the part where all the calculation is done, AXON is used for transferring the data between CELL BODY and SYNAPSES, SYNAPSES are nothing but the storage unit of neurons where all the information store [7]. Moreover, SYNAPSES are connected with other 1000 neurons which makes it faster and parallel.

As shown in Fig. 2b, neuromorphic Neuron design. But it is the design of the single neuron. For the functioning like the brain, it needs to have 10 Billion neurons [7]. It becomes too complex to handle work of all the neurons in the system.

As shown in Fig. 3 each neuron has a router in it which is connected with all other neurons in the computer. As all the neurons are connected with other neurons some algorithm or methods is required to handle the working and connect all the neurons. As the solution of this, Neural Network works.

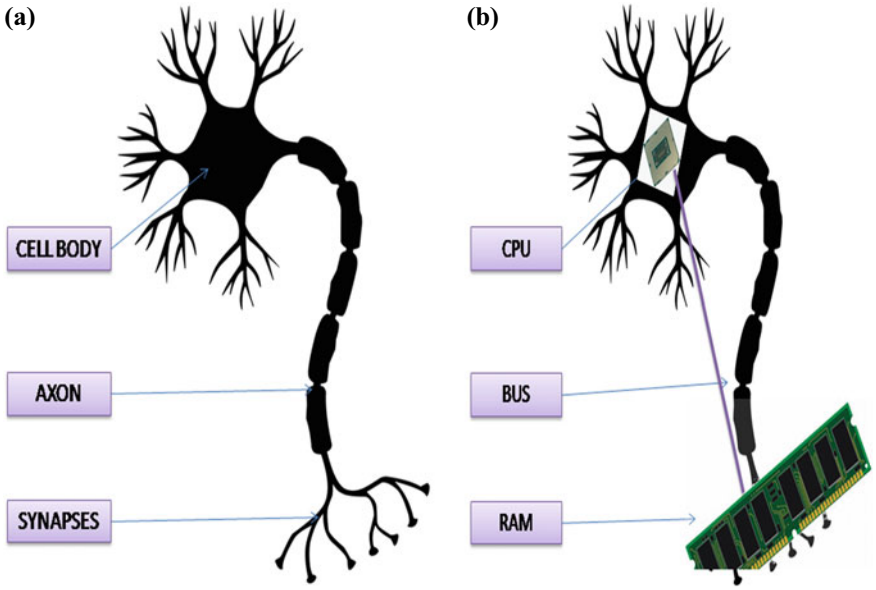


Fig. 2 **a** Brain Neuron [8], **b** Architecture single Neuron [5]

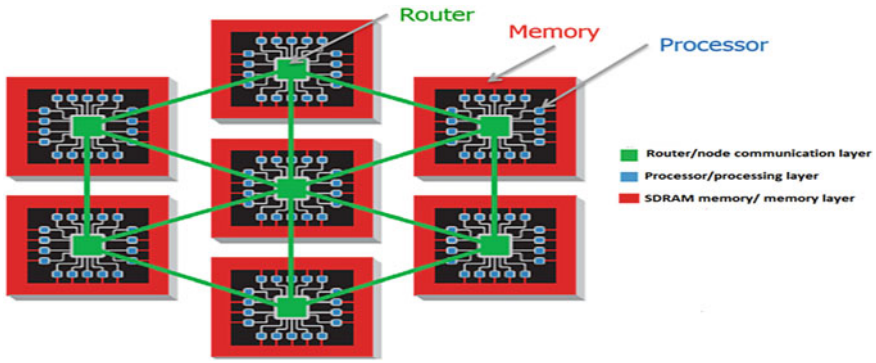


Fig. 3 Architecture of neuromorphic computer

3.1 Neural Network and Neuromorphic Computer

Neural Network is also inspired by the concept of biology. The neural network is nothing but a set of neurons that are connected with each other. It means all the neurons are connected with all neurons of the next layer. It may have several layers and each layer may have a different number of neurons in it.

As shown in Fig. 4, it is an example of a neural network. In any fully connected neural network, all the neurons are connected with all the neural of the next layer.

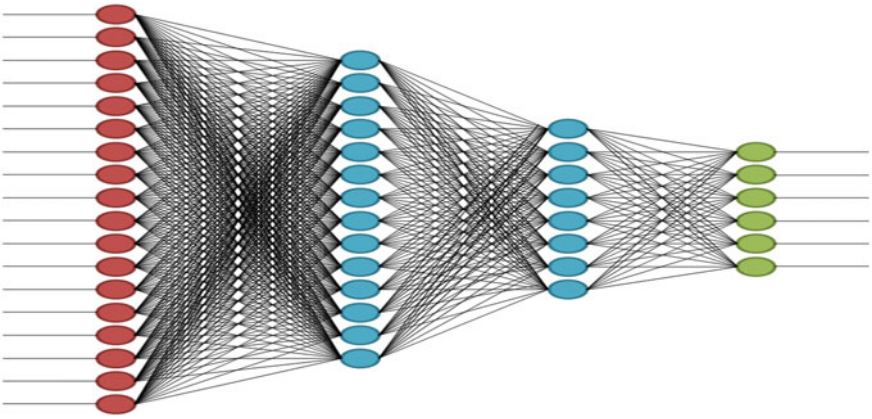


Fig. 4 Neural network

Here in Fig. 4, a red color circle represents the input layer, a blue color circle for the hidden layer, and a green color circle for the output layer. As far as machine learning and deep learning are concerned, a more hidden layer means more accurate output and a better model. It is also possible to add more hidden layers and more neurons in that hidden layer. Output layer neurons have the final output of any learning.

In the neuromorphic computers, all the neuron is connected with all the neurons of the next layer. This creates a more complex structure of the network but at the time it provides a system in which one CPU with its RAM is connected with other thousand CPUs which are also connected with others and so on [7].

As shown in Fig. 5, each neuron of the layer is connected with all neurons of the next layer. Due to this kind of structure, the computer becomes faster. As all neurons are connected with all other neurons, failure of some neurons in the network/computer never affects the accuracy and speed of the computer. In the first phase, all the neurons learn some specific work and store its output in their own memory, when the next time any similar work is given to the computer it will do the work very quickly. As the network grows with time all the neurons have their own working experience as our brain has. And then the computer becomes able to work like the human brain.

3.2 Neuromorphic Concepts

The below concepts play a vital role in neuromorphic computers. These concepts describe the working of neuromorphic computing. Each point describes the behavior and working methodology.

Spiking: In neuromorphic computers, the signals are communicated between the neurons through the voltage or the current spikes [7]. This communication is quite different from the current digital, analog, or binary signals, which rely on

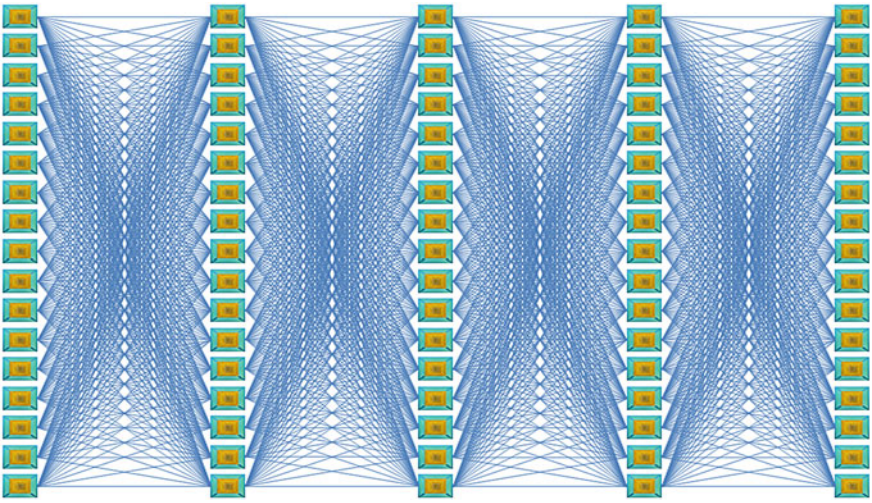


Fig. 5 Neurons connnection

the manipulation of continuous signals. The Spiking signaling system is fully time encoded and transmitted via "action potentials".

Plasticity: This device is designed to learn from past experiences. The term plasticity is used for the response from the device to some input through the past experience [9]. This device will change its properties of elements depending on the past history.

Fan-in/fan-out: It showed the connection between devices. It is not possible to connect each neuron with all other neurons of the circuit due to physical limitations. So to determine how many neurons are interconnected with each other in a device, further study is needed. In the human brain, this number of connections is on too large a scale, i.e., 10,000.

Hebbian learning/dynamic resistance change: This is the long-term change in the synapses resistance after repeated spiking by the neurons. It is referred to as spike time-dependent plasticity (STDP). This can be also called the "device that is fire together, wire together".

Adaptability: The biological brain starts with the multiple connections from which, through the selection and learning process, some are selected and others not. This process takes some time to learn and improve fault tolerance. Moreover as time passes it finds the best path of communication. In the same way, the neuromorphic devices will also be able to learn this adaptability of the brain functionality. In a neuromorphic computer, the architecture will not be fixed from the beginning like in today's computer system.

Criticality: The human brain works on the critical point where the brain has plasticity enough that it can be switched from one state to other state and neither too stable nor very volatile at the same time. In the same way, the computer must be able to explore many close-lying states.

Accelerators: The computer-like brain needs some intermediate states to work with some special applications. For that, it will be able to combine with other sensors.

3.3 Building Blocks

The neuromorphic computer has the following building blocks for smooth functioning.

Synapses/Memristor: This is the most advanced element in this new kind of computer. This has two main properties, i.e., switching and plasticity. This is implemented in the two-terminal devices such as a memristor [10].

Soma/Neuristor: This device provides two important functionalities, i.e., integration and threshold spiking. A possible implementation of this device consists of a capacitor in parallel with a memristor.

Axon/Long wire: The role of the axon is as simple as a wire. Its main functionality is to provide connections between neurons. Even though it's simple functionality it needs more research because the device is required high-speed data transmission, as the human brain can transfer data between two neurons at such a high speed that cannot be achieved by any man-made devices. Moreover, some other functionality is going to be added like signal-conditioning [7]. For that more research is required.

Dendrite/ Short wire: The role of this wire is to provide a signal from multiple neurons into one single neuron. This emphasizes the 3-D nature of the connectivity of the human brain [7]. Some new research shows that it can also help to detect the pattern. One of the research shows that it can detect over a hundred patterns.

Fan-in/fan-out: Neurons have a connection to many thousands of other neurons. One axon may connect to the ten thousand or more than ten thousand dendrites. But due to the current limitation of the electrics, it can connect with a few ten of terminals. To have the connection like the brain more research is required to do on this fan-in/fan-out. The human brain has billions of connections.

3.4 Algorithms for Neuromorphic Computing

The neuromorphic computer still required a deep research to select most preferable algorithm. Many organizations and researcher believe that Spiking Neural Network(SNN) is the most preferable algorithm. In earlier phase of Neuromorphic computing many researcher believed to use ANN (Artificial Neural Network) but with development it also becomes limited due to its own limitation [11]. Currently SNN is working on many Neuromorphic chips. Many developers believed to use some new algorithm which should be able to connect multiple boards of chips as well as able to handle more and more connections. Here a small discussion is about SNN.

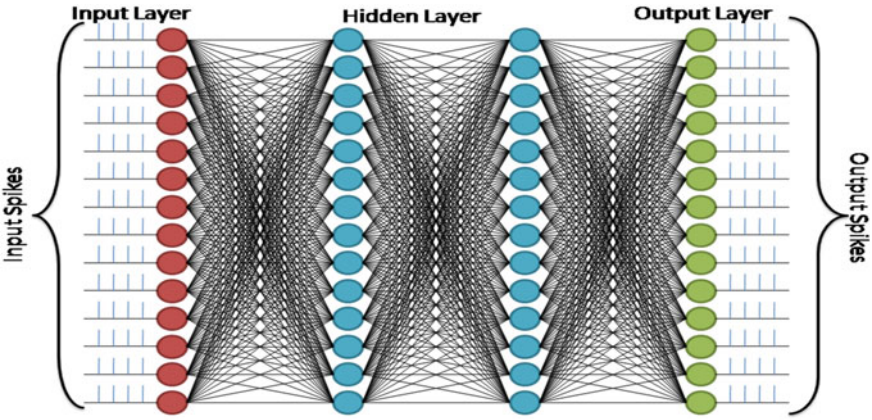


Fig. 6 Full spiking neural network

Spiking Neural Networks.

The neuromorphic computer works on the Spiking concept. In this approach, neurons do not transmit information to each propagation but it transmits to the next layer of neurons if and only if required. It can be identified by the value known as the threshold. When the working neurons reach the threshold then the neuron fire the signals to the other neuron which in turn, increases or decreases their potentials in response to the signal [12, 13].

Here in Fig. 6, it shows architecture of SNN. This network is most useful in neuromorphic computing as it passes the signal by learning from the past experience. The following figure shows working of SNN.

As shown in Fig. 7, SNN activates some or all neurons of the next level. Here all activate lines are shown in red color. This always depends on the experience. So that each time neurons compute something it store output in it. As next time same kind of problem comes, the same neurons will do the work so that computation speed can be increased. This way it becomes possible to use the feature same as humans have.

4 Loihi-Intel

Loihi is the chip design by Intel. It is a neuromorphic research test chip designed by the Intel labs that uses the Spiking Neural Network. The chip has 128-neuromorphic cores. This chip was presented in 2018 at the Neuro Inspired Computational Element workshop in Oregon (Fig. 8).

This is one of the milestones in the development of neuromorphic devices. From a device like loihi, Intel is going to design different chips that are more powerful and faster than this which will change the era of computer science.

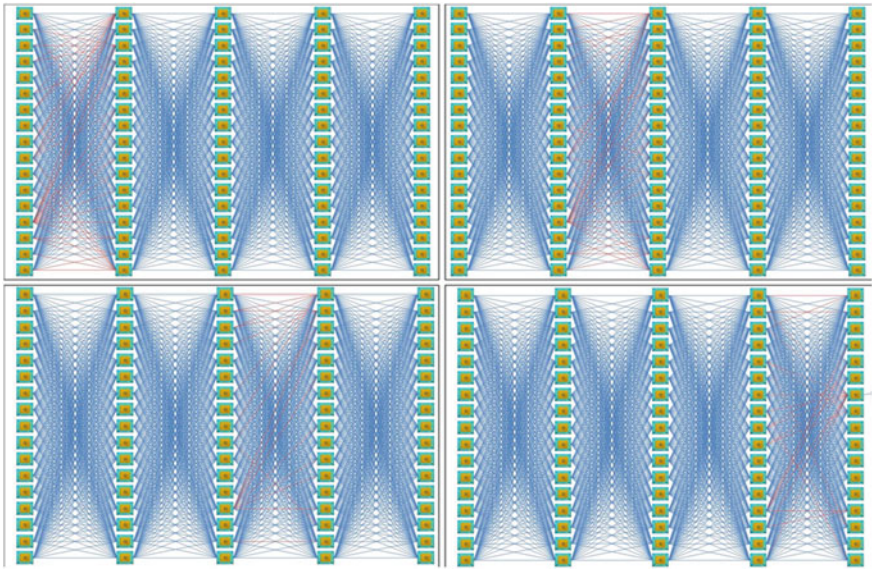


Fig. 7 Working of spiking neural network

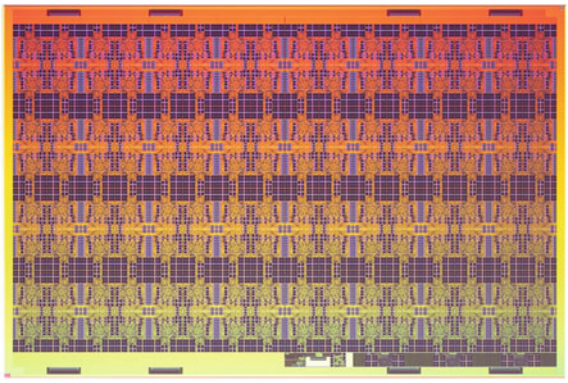


Fig. 8 Loihi chip by Intel [14]

Intel is planning to develop a chip that will be as powerful as a human brain. But it is not a task of a single phase. So it is divided into different parts. The next chip that is going to develop is Kapoho Bay which has 1- 2 chips consisting of 130,000,000- 260,000,000 Synapses and 131,072 - 262,144 Neurons. In the next phase, Intel will develop a device which will be known as Wolf Mountain. It will have 4 chips consisting of 520,000,000 Synapses and 524,288 Neurons. The next task will be having device called as Nahuku which will have 8–32 Chips consisting of 1,040,000,000- 4,160,000,000 Synapses and 1,048,576- 4,194,304 Neurons. In the next phase, it will develop a device known as Pohoiki Beach which will have 64 chips consisting of 8,320,000,000 Synapses and 8,388,608 Neurons. And at the end,

the final required chip will be developed which will be known as Pohoiki Spring which will have 786 chips consisting of 99,840,000,000 Synapses and 100,633,296 Neurons. It will be able to work like a human brain. The final goal the neuromorphic computing is to create a device that has 10 Billion Neurons and 100 Trillion Synapses.

5 Applications

The neuromorphic computer will use for the new era's technology including AI, ML, DL, and Big data. It is also a fact that the real use of a neuromorphic computer is not clear. But as the architecture and working methods show it will be used in the field of image processing, text mining, data classification, natural language processing, anomaly detection, speech-based systems [15], etc. Implementing the neural network on the hardware device helps in fast computation, consumes low power energy as well as more accurate. This device is also helpful in the field of neuroscience where it becomes easier to understand the working of neurons more clearly and effectively.

Vision-based system: In a real-world application, this kind of system is useful in the vision-based system where biology visual is needed. It can be used in sense based systems, touch-based systems, somatosensory, etc. This kind of device is used in the system where deep monitoring is required like medical treatment or monitoring. As this computer works like a brain, it uses in brain monitoring, retina implant, wearable fall detector for elderly people, etc.

Robotics: This device is specially designed to be used in the field of robotics. Its size is small and power consumption is low as well as it is like the human brain that is why it is used in the robots for the smart and more accurate humanoid robots.

Control system: Neuromorphic computers can be used in a control system. As it can work very quickly, it can use in the real-time application where all decisions and all the actions need to take on time. It can be used in the gaming application as well where real-time response is needed.

Image processing: The most common application of this computer is image processing. As the system is also based on neural networks and image processing also used a neural network, this system can respond more quickly and more accurately. This computer is used in image classification, region-based clustering, edge detection, image compression, image filtering, image segmentation, and feature extraction. Image classification, object detection, and recognition is the most common and used application.

Number identification: Another common application is number identifying. In this method also neuromorphic computer can give quick and more accurate results. It can be also used in the house number reading, identifying number plat of vehicle, traffic light reading, etc.

Sound recognition: Another application is the sound-based recognition task. Speech recognition is becoming more popular as Goggle and Amazon start the device that works on speech recognition. It can also be used in music recognition. As the

neuromorphic devices use the spiking neural network, they can do a real-time task so that they can identify the speed, remove the noise, and distinguish speech as well.

Video identification: Another most common application is video-based identification. Neuromorphic can be used for identifying the object from the video frame. It is also possible to detect the motion in the video.

Natural language processing: It is possible to analyze any language using natural language processing. It is a task in which the system identifies the meaning of any language by analyzing the statements given to it.

The fully working system will be more useful in real-time applications. There will be more applications of the neuromorphic computer than described above.

6 Research Opportunities

As the neuromorphic computer is in the developing phase there are lot of areas in which research is needed. The area of research in neuromorphic computing covers the whole area from physical research to artificial neural network selection.

Neural Network selection: The selection of the most appropriate neural network still needs research. There are many kinds of networks that can use in the neural network, but the best network is still a question (Fig. 9).

Synapses can connect with only limited other synapses. This is the limitation of the hardware devices available to the date. Moreover, the memory which synapses have is too limited. That's why more research is required for this as well.

Managing multiple boards: The connection between multiple boards/chips is also a requirement of the neuromorphic device. But managing more than one chip/board concurrently is an issue. It will be required a core system that can handle all the board and devices. It is also an area of research.

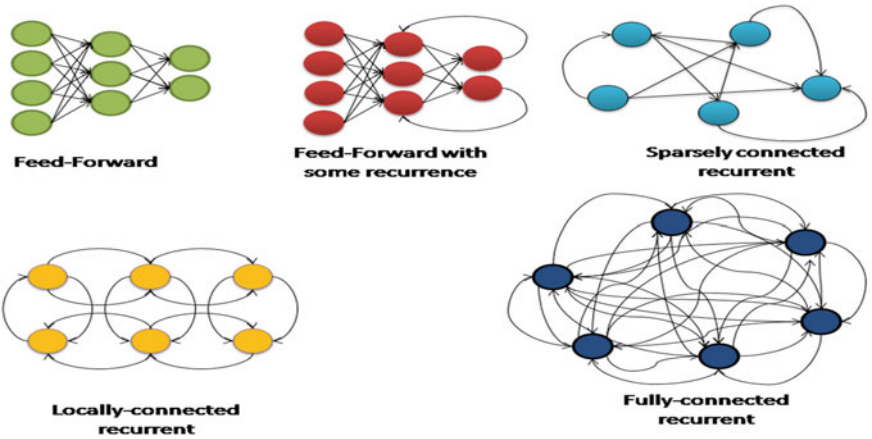


Fig. 9 Neurons topologies

Axon and short wire: Axon and short wire have to be too quick so that they can transfer the data quickly. Delay must be minimum for working as the brain. The brain transfers the data between the neurons at such a speed so that we are not able to recognize the difference between the real-time and delay of analysis done by the brain. We have to develop such a device that can transfer the data at such a speed as the brain has.

Physical devices: The physical devices need by the neuromorphic computer are still an area of research. It needs more research to have a device that can work like the human brain. The handling of all the neurons and synapses and well as a network is a big issue that needs more research to be done on it.

The physical material for developing the neuromorphic computer is also important because at present there is no natural element that can do all the required work at high speed. So it is needed to develop such an artificial element that can do the task.

7 Conclusion

This new kind of device is required to integrate approaches of computer science, engineering, scientist, and material science. The new device needs a new kind of architecture and methodology does create it. It is also necessary to create novel nonmaterial, which opens up the quantum material structure of physical objects. The development is not possible in one step, it has to be done in phases. In near future, it will be possible to have a device that can able to recognize the pattern and able to learn from past experience. This new device is going to change the future of computer science. It will start the new era of computer science and information technology.

References

1. B. Acun, D. J. Hardy, L. V. Kale, K. Li, J. C. Phillips, J. E. Stone. Scalable Molecular Dynamics with NAMD on the Summit System
2. British Neuroscience Association European Dana Alliance for the Brain: Science of the brain an introduction for young students
3. Von Neumann Architecture and Modern Computers, Felix Ukpai Ogban, Iwara I. Arikpo, Idongesit Efaeminde Eteng.
4. Photonic Neuromorphic Computing: The Future of AI, web content available at https://www.youtube.com/watch?v=hBFLeQlG2og
5. IBM's Incredible TrueNorth Chip || Neuromorphic Computing, web content available at https://www.youtube.com/watch?v=X2TYAcr36r0
6. F.U. Ogban, I.I. Aripo, I.E. Eteng, Von Neumann architecture and modern computers. Glob. J. Math. Sci. **6**(2) (2007)
7. Neuromorphic Computing: From Materials to Systems Architecture- Report of a Roundtable Convened to Consider Neuromorphic Computing Basic Research Needs
8. Basic Neural Units of the Brain: Neurons, Synapses and Action Potential Jiawei Zhang
9. "Neuromorphic Adaptive Plastic Scalable Electronics", IEEE PULSE, JANUARY/FEBRUARY 2012

10. K. Saeki, R. Iidaka, Y. Sekine, and K. Aihara, "Hardware neuron models with cmos for auditory neural networks," in Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on, vol. 3. IEEE, 2002, pp. 1325–1329.
11. A historical survey of algorithms and hardware architectures for neural inspired and neuromorphic computing applications, Conrad D. James, James B. Aimone , Nadine E. Miner , Craig M. Vineyard, Frederick H. Rothganger , Kristofor D. Carlson , Samuel A. Mulder , Timothy J. Draelos , Aleksandra Faust , Matthew J. Marinell , John H. Naegle , Steven J. Plimpton
12. Towards spike-based machine intelligence with neuromorphic computing, Kaushik Roy, Akhilesh Jaiswal & Priyadarshini Panda
13. Photonics for artificial intelligence and neuromorphic computing, Bhavin J. Shastri, Alexander N. Tait, T. Ferreira de Lima, Wolfram H. P. Pernice, Harish Bhaskaran, C. D. Wright & Paul R. Prucnal
14. Web content available at https://en.wikichip.org/wiki/intel/loihi
15. A Survey of Neuromorphic Computing and Neural Networks in Hardware Catherine D. Schuman, Thomas E. Potok, Robert M. Patton, J. Douglas Birdwell, Mark E. Dean, Garrett S. Rose, and James S. Plank IEEE- 19 May 2017
16. A. Disney, J. Reynolds, C.D. Schuman, A. Klibisz, A. Young, J.S. Plank, Danna: A neuromorphic software ecosystem. Biologic Inspired Cognitive Architec **17**, 49–56 (2016)
17. S. Mitra, S. Fusi, G. Indiveri, Real-time classification of complex patterns using spike-based learning in neuromorphic vlsi. IEEE Trans. Biomed. Circuits Syst. **3**(1), 32–42 (2009)
18. R.S.N. Ntoune, M. Bahoura, C.-W. Park, "Fpga-implementation of pipelined neural network for power amplifier modeling", in New Circuits and Systems Conference (NEWCAS), IEEE 10th International. IEEE **2012**, 109–112 (2012)
19. R. Woodburn, A. Astaras, R. Dalzell, A. F. Murray, and D. K. McNeill, Computing with uncertainty in probabilistic neural networks on silicon, in Proceedings of Symposium on neural computation, 2000
20. Neural Networks and Neuroscience-Inspired Computer Vision David Daniel Cox, Thomas Dean
21. Convolutional networks for fast, energy-efficient neuromorphic computing**,** Steven K. Esser, Paul A. Merolla, John V. Arthur, Andrew S. Cassidy, Rathinakumar Appuswamy, Alexander Andreopoulos, David J. Berg, Jeffrey L. McKinstry, Timothy Melano, Davis R. Barch, Carmelo di Nolfo, Pallab Datta, Arnon Amir, Brian Taba, Myron D. Flickner, and Dharmendra S. Modha

Computational Intelligence Approaches for Heart Disease Detection

Roseline Oluwaseun Ogundokun, Sanjay Misra, Peter Ogirima Sadiku, Himanshu Gupta, Robertas Damasevicius, and Rytis Maskeliunas

Abstract A disease is an occurrence that affects one or more areas of a person's body. Various diseases are on the rise as a result of changing lifestyles and patrimonial values. Heart disease (HD) is the most serious of all disorders, and its consequences are much more dangerous than those of any other disease. Therefore, early detection of HD will reduce the death rate of people. Computational Intelligence (CI) approaches can be employed for the early diagnosis and detection of HD. Hence, employed the use of two computational intelligence approaches. The study compared a variety of computational intelligence strategies for heart disease detection. A comparison analysis was drawn using Two computational intelligence techniques: Decision Tree (DT) and K-Nearest Neighbor (KNN). A feature extraction algorithm which is Autoencoder was employed to reduce the number of attributes required to describe the heart disease dataset. The performance of each approach was measured using heart disease databases obtained from the National Health Service (NHS) database and uncertainty matrix success assessment metrics.

R. O. Ogundokun (✉)
Department of Computer Science, Landmark University, Omu Aran, Nigeria
e-mail: ogundokun.roseline@lmu.edu.ng

S. Misra
Center of ICT/ICE Research, Covenant University, Ota, Ogun State, Nigeria
e-mail: sanjay.misra@covenantuniversity.edu.ng

P. O. Sadiku
Department of Computer Science, University of Ilorin, Ilorin, Nigeria
e-mail: sadiku.po@unilorin.edu.ng

H. Gupta
Birla Institute of Technology Pilani, Pilani, India
e-mail: f20150339h@alumni.bits-pilani.ac.in

R. Damasevicius · R. Maskeliunas
Kaunas University of Technology, Kaumas, Lithuania
e-mail: robertas.damasevicius@ktu.lt

R. Maskeliunas
e-mail: Rytis.Maskeliunas@ktu.lt

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_29

Keywords Steganography · Least Significant Bit · PSNR · MSE · Comparative analysis

1 Introduction

Heart Disease (HD) is a major cause of demise in today's world. Millions of people die from HD yearly, according to the World Health Organization (WHO) [1]. HD is a condition of the heart that inhibits the normal rhythm of the heart. Since the coronary veins produce blood to the heart, tightening or obstruction are the deuce utmost usual causes of HD. According to projections, beyond 30 million individuals would decease from heart failure by 2040. HD ailment has many characteristics that may impair the edifice or operation of the heart. Physicians and surgeons experience numerous difficulties in reliably and efficiently identifying heart failure. As a result, creating an intelligent HD detection approach to forecast HD in its early stages at a low cost is important. 25% of persons suffering from HD decease unexpectedly without any prior signs [2]. HD is a single of the utmost common forms of ailments that could damage the heart and lead to a heart attack. HD can be reduced with timely care and knowledge of disease symptoms. Several causes to increase the risk of heart disease have been discovered by several experts. Smoking, alcohol, elevated gore saturated fat, diabetes mellitus, obesity, increased blood pressure, and other risk factors may all raise the likelihood of HD and blood vessel disease [3]. Other conditions, such as stress and alcohol, are regarded as contributory hazard influences and might raise the risk of heart failure. Their meaning and incidence, on the other hand, have yet to be determined. Any of the risk factors may be modified, regulated, or handled. As a result, the individual with more risk factors has a greater probability of being attacked by the HD [3].

Many studies show that computational intelligence methods have sparked a lot of interest in the investigative community, and they can provide a lot of consistency in classification problems as compared to other data classification models [4]. In the field of statistics, LR is a type of numerical intelligence methodology. It's a statistical analysis of a binary dependent variable. The LR rep's results are presented as succeeding/abort or victory/misplace [5]. It's likewise used in the medicinal industry to address a variety of issues. In today's science, SVMs are one of the utmost common and widely employed CI techniques. SVMs are primarily used for supervised learning and are employed in categorization, regression, and the identification of outliers. Image recognition, Bioinformatics, textual information categorization, hand-written categorization, and other fields have seen a lot of progress with SVM [6]. Many studies have shown that SVM outperforms other classification methods including computational algorithms, decision trees, and instance-centered learning approaches [7]. Several CI and ML methods [8–25] have recently been employed in numerous medical diagnoses. The study contrasted some computational intelligence strategies such as DT, RF, and KNN that can envisage HD more accurately and quickly in this

article. Many methods, such as RapidMiner, WEKA, and others, may provide classification outcomes. However, there are several drawbacks to using such methods. It used too much memory in RapidMiner, causing the machine to slow down [26]. It also encourages the use of less formula which is more costly. WEKA, on the other hand, does not use the modern technique. WEKA's key concerns are scaling, data processing, and simulation [27].

Hence, employed the use of two computational intelligence approaches. The study compared a variety of computational intelligence strategies for heart disease detection. A comparison analysis was drawn using Two computational intelligence techniques: DT, and KNN. A feature extraction algorithm which is Autoencoder was employed to reduce the number of attributes required to describe the heart disease dataset. The performance of each approach was measured using heart disease databases obtained from the National Health Service (NHS) database and uncertainty matrix success assessment metrics.

The article is structured as thus: the subsequent section discussed the associated previous researches that have adopted the use of CI for diagnosis and detection of diseases. Section 3 discusses the materials and methods used for the implementation and testing of the system. Section 4 discusses the results and finding discovered from the implementation. The last section concludes the article and future works are suggested for the research.

2 Literature Review

The diagnosis of HD in an individual is problematic and demands a diversity of details, lab experiments, and instruments [27]. This investigation pursued to encourage this technique using novel knowledge, for instance, Machine Learning (ML), rather than substituting the traditional procedure for detecting and prognosing the likely-hoods of HD. The ML technique is not modern, and it has been employed for a diversity of purposes. [28] projected a cloud-centered verdict support system to help cardiologists all through the evaluation process. For prognostication of heart disease, these gadgets employed ML procedures. The system was intended to deliver support economically, with the potential to link with existing structures. In [29], the author employed a technique of utilizing numerous bunching approaches on an HD dataset to discover the superlative resolution for exploiting the estimation accuracy. In their analysis of employing Naive Bayes (NB), Decision Trees (DT), Support Vector Models (SVM), and further approaches, ML techniques were discovered to be effective in prognosis HD using historical data [30]. Their outcomes demonstrated that the SVM delivered the superlative results when compared with further techniques.

Bashir et al. (2019) employed a feature selection method to upsurge the accuracy of the heart disease forecast. Different techniques were used in the experiment, including Nave Bayes, Random Forest, and others. Due to the feature selection method, the performance showed a high level of precision. According to a report, most heart disorders are not always measurable clearly by administering an ECG (a

kind of examination for diagnosing a heart's functioning capability) [31]. To help the medical practitioners, an electronic healthcare gadget [32–35] with an extensive variety of purposes was established precisely for the medical division to manage patients' time, resources, and, most notably, to assist surgeons in saving the patient's lifetime. Furthermore, a research analysis published by the researcher [36] elucidated the significance of classification approaches and the essentials of the approaches presently employed in the medical industry. Several researchers have successfully utilized data mining approaches in health care conditions. Similarly, additional investigation likened the outcomes of innumerable classifiers employed to deuce diverse gadgets: MATLAB and WEKA. Generally, the accuracy of DT, Linear SVMs, and further approaches extended from 52% to 67.7%, even though the accuracy was fair. Conferring to the researches discussed herein, innumerable kinds of approaches have variance in the prognosing assessments. As a consequence, dimensionality reduction and function manufacturing could support to advance of the data gathering procedure, which as a result could assist to enhance the accuracy of the system [37].

3 Material and Method

This section discussed the proposed system, HD datasets used for the system implementation and testing.

3.1 Proposed System

This study used two computational intelligence techniques which are DT and KNN. Figure 1 shows the computational flow of the proposed HD detection system. The heart disease datasets are inputted into the system and are passed into the feature extraction technique which in this case it is autoencoders. The datasets then go through the preprocessing stage which is the training and testing process after which the preprocessed datasets are passed into the two computational intelligence classifiers used in this study and then the system decides whether an individual is positive or negative with HD diagnosis. The system performance is then finally evaluated using the confusion matrix.

3.2 Description of Heart Disease Dataset

Figure 2 shows the excel sample heart disease dataset used for this study implementation and testing.

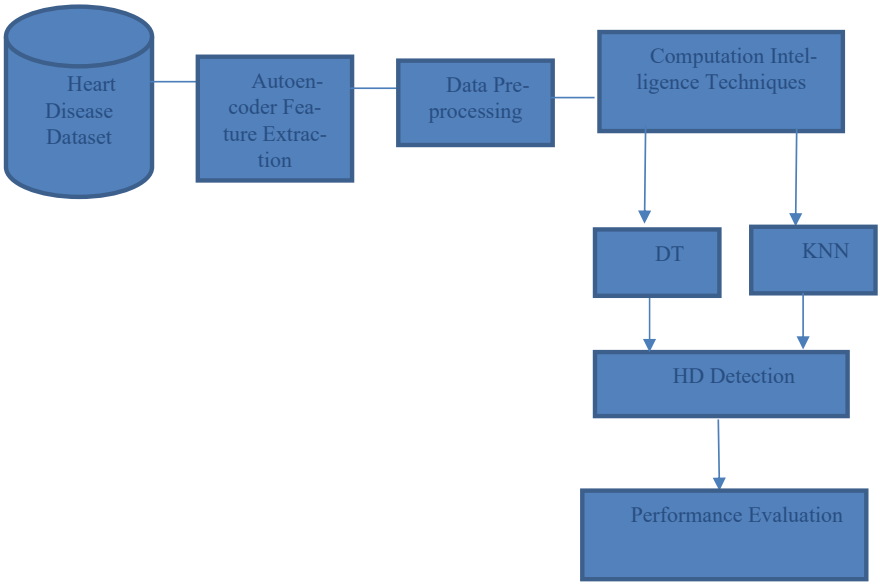


Fig. 1 Proposed heart disease detection system flow

4 Results and Discussion

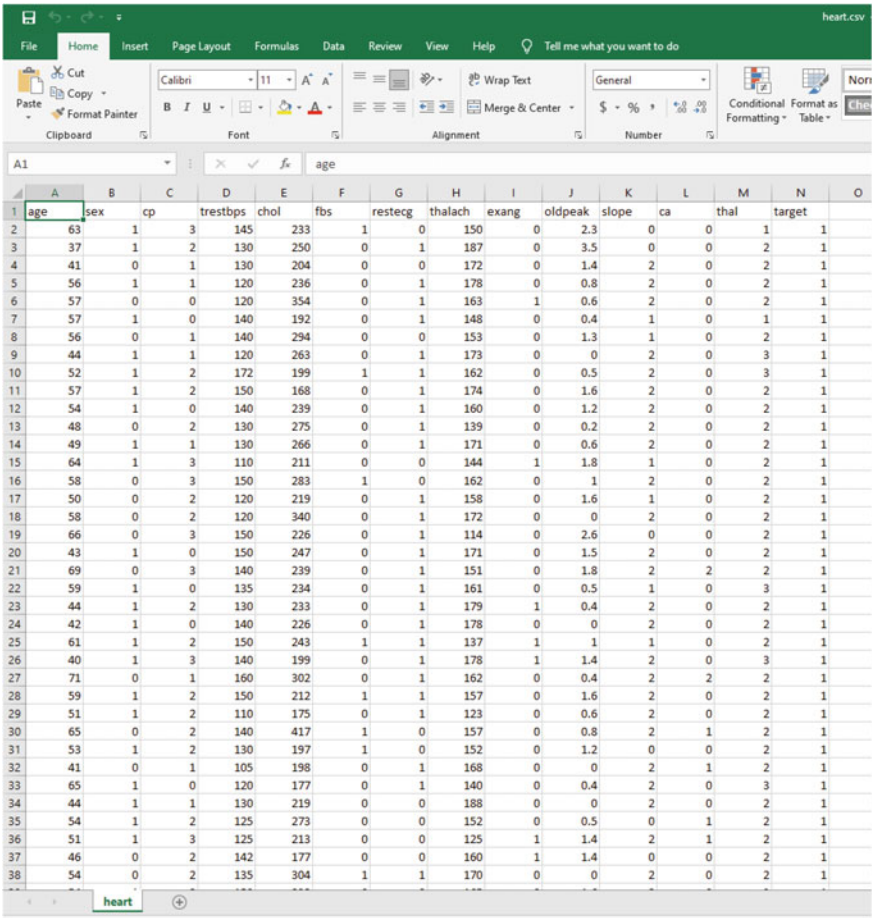
This section discussed the implementation finding interfaces and also the finding discovered were discussed similarly in this section.

In Fig. 1, the TN is 82, FP is 63, FN is 68, and TP is 86 and Fig. 2 displayed that TN is 110, FP is 38, FN is 100, and TP is 52.

Table 1 shows the classification report for the execution of autoencoder with KNN classifier and Table 2 shows the classification report for the execution of autoencoder with DT classifier (Fig. 3).

5 Discussion

Autoencoder feature extraction technique was used in this study with two computational intelligence algorithms which are Decision Tree (DT) and K-Nearest Neighbor (KNN). It is shown in Table 3 and Fig. 4 that the autoencoder with decision tree accuracy of 56.17% outperformed that of the autoencoder with K-nearest neighbor with an accuracy of 54%. Autoencoder with DT also outperformed in measures such as sensitivity with 55.84% over that of autoencoder with KNN with 34.21% and F1 score of Autoencoder plus DT having 56.77% and autoencoder plus KNN having 42.98%. Autoencoder plus KNN outperformed that of autoencoder plus DT with a specificity of 74.32 to 56.55% and precision of 57.78 to 57.72% (Fig. 5).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
2	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
3	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
4	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
5	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
6	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
7	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
8	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
9	44	1	1	120	263	0	1	173	0	0	2	0	3	1
10	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
11	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
12	54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
13	48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
14	49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
15	64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
16	58	0	3	150	283	1	0	162	0	1	2	0	2	1
17	50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
18	58	0	2	120	340	0	1	172	0	0	2	0	2	1
19	66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
20	43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
21	69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
22	59	1	0	135	234	0	1	161	0	0.5	1	0	3	1
23	44	1	2	130	233	0	1	179	1	0.4	2	0	2	1
24	42	1	0	140	226	0	1	178	0	0	2	0	2	1
25	61	1	2	150	243	1	1	137	1	1	1	0	2	1
26	40	1	3	140	199	0	1	178	1	1.4	2	0	3	1
27	71	0	1	160	302	0	1	162	0	0.4	2	2	2	1
28	59	1	2	150	212	1	1	157	0	1.6	2	0	2	1
29	51	1	2	110	175	0	1	123	0	0.6	2	0	2	1
30	65	0	2	140	417	1	0	157	0	0.8	2	1	2	1
31	53	1	2	130	197	1	0	152	0	1.2	0	0	2	1
32	41	0	1	105	198	0	1	168	0	0	2	1	2	1
33	65	1	0	120	177	0	1	140	0	0.4	2	0	3	1
34	44	1	1	130	219	0	0	188	0	0	2	0	2	1
35	54	1	2	125	273	0	0	152	0	0.5	0	1	2	1
36	51	1	3	125	213	0	0	125	1	1.4	2	1	2	1
37	46	0	2	142	177	0	0	160	1	1.4	0	0	2	1
38	54	0	2	135	304	1	1	170	0	0	2	0	2	1

Fig. 2 Heart disease dataset sample

Table 1 Autoencoder + KNN classification report

Code	Precision	Recall	F1-score	Support
0	0.51	0.74	0.60	144
1	0.58	0.34	0.42	155
accuracy			0.53	299
Macro avg	0.54	0.54	0.51	299
Weighted avg	0.54	0.53	0.51	299

Table 2 Autoencoder + DT classification report

Code	Precision	Recall	F1-score	Support
0	0.55	0.57	0.56	145
1	0.58	0.56	0.57	154
accuracy			0.56	299
Macro avg	0.56	0.56	0.56	299
Weighted avg	0.56	0.56	0.56	299

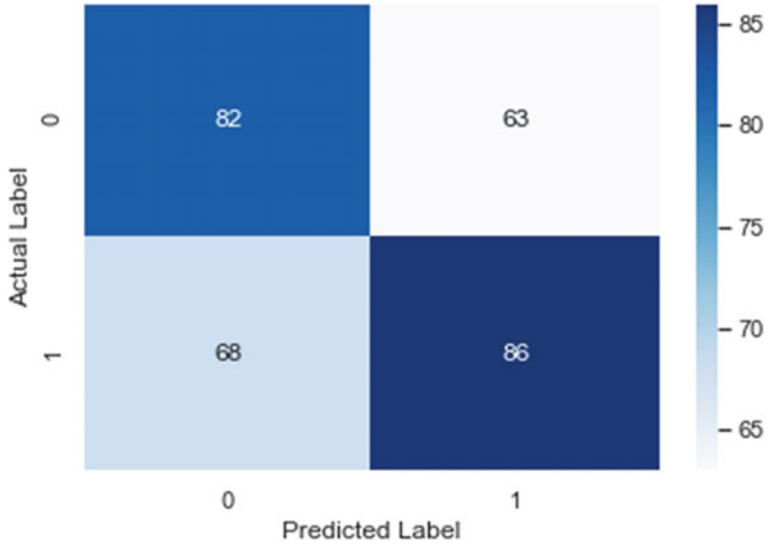


Fig. 3 Confusion matrix for Autoencoder + DT

Table 3 Performance evaluation of the system

Measures	Autoencoder + DT Values (%)	Autoencoder + KNN Values
Accuracy	56.19	54
Sensitivity	55.84	34.21
Specificity	56.55	74.32
Precision	57.72	57.78
F1 Score	56.77	42.98

6 Conclusion

In recent years, substantial progress has been made in the development of medicinal datasets, as well as improvements in implementation metrics and efficiency, which are heavily driven by exploratory plans, activities, and data processing types. Heart

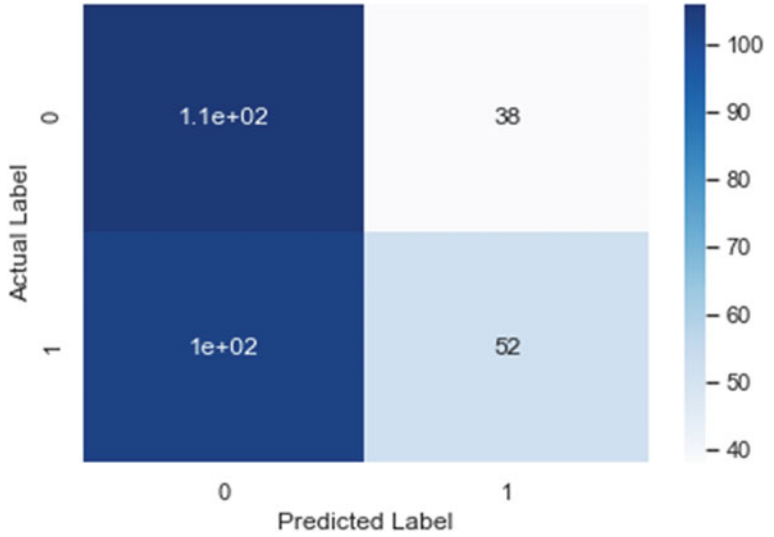


Fig. 4 Confusion matrix for Autoencoder + KNN

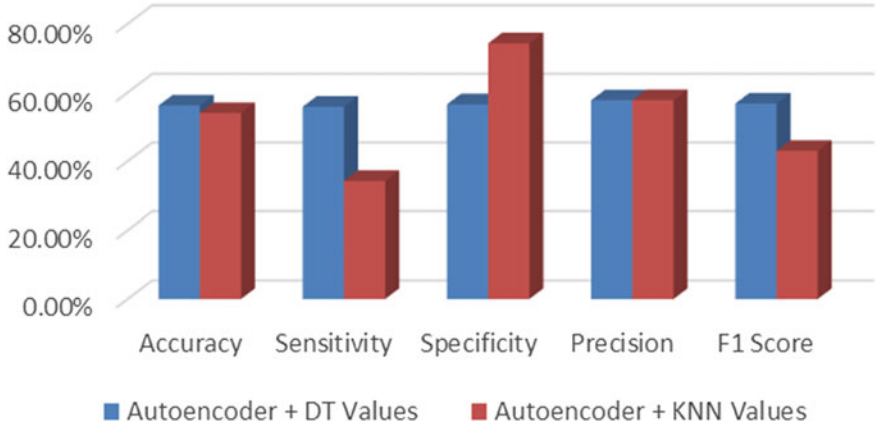


Fig. 5 Comparative analysis between the two computational bits of intelligence

disease (HD) is the most serious of all disorders, and its consequences are much more dangerous than those of any other disease. Therefore, early detection of HD will reduce the death rate of people. Computational Intelligence (CI) approaches can be employed for the early diagnosis and detection of HD. Hence, employed the use of two computational intelligence approaches. The study compared a variety of computational intelligence strategies for heart disease detection. It is therefore concluded that CI which is Autoencoder with DT can be adopted for medical diagnosis and disease detection as it gives more efficient and effective accuracy and

sensitivity. In the future, researchers can combine the autoencoder feature extraction with other classifiers or even engage the use of another FE with CI classifiers. Larger HD datasets can likewise be employed for system execution and testing.

Acknowledgments The authors appreciate the sponsorship from Covenant University through its Centre for Research, Innovation and Discovery, Covenant University, Ota Nigeria.

References

1. Coronary Artery Disease. https://my.clevelandclinic.org/ health/diseases/16898-coronary-artery-disease.
2. R.O. Bonow, D.L. Mann, D.P. Zipes, P. Libby, *Braunwald's heart disease: A textbook of Cardiovascular Medicine"*, vol. 9 (Saunders, New York, 2012)
3. Risk Factors for Coronary Artery Disease. https://www.healthline.com/health/coronary-artery-disease/riskfactors.
4. K. Bache, M. Lichman, *UCI machine learning repository* (University of California, School of Information and Computer Science, Irvine, CA, 2013)
5. C.R. Boyd, M.A. Talson, W.S. Copes, Evaluating trauma care: The TRISS method. Trauma score and the injury severity score. J. Trauma **27**, 370–378 (1987)
6. K. Durgesh, S.L. Bhambhu, Data classification using support vector machine. J. Theor. Appl. Inf. Technol. **12**, 1–7 (2009)
7. T. V. Gestel et al., Benchmarking Least Squares Support Vector Machine Classifiers, Vol. 54. Kluwer Academic Publishers (Machine Learning), 2004, pp. 5–32.
8. D.K. Bangotra, Y. Singh, A. Selwal, N. Kumar, P.K. Singh, W.C. Hong, An intelligent opportunistic routing algorithm for wireless sensor networks and its application towards e-healthcare. Sensors **20**(14), 3887 (2020)
9. T.O. Oladele, R.O. Ogundokun, A.A. Kayode, A.A. Adegun, M.O. Adebiyi, Application of Data Mining Algorithms for Feature Selection and Prediction of Diabetic Retinopathy. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11623 LNCS, pp. 716–730 (2019)
10. A.A. Adeyinka, M.O. Adebiyi, N.O. Akande, R.O. Ogundokun, A.A. Kayode, T.O. Oladele, A Deep Convolutional Encoder-Decoder Architecture for Retinal Blood Vessels Segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11623 LNCS, pp. 180–189 (2019)
11. M. Elhoseny, M.A. Mohammed, S.A. Mostafa, K.H. Abdulkareem, M.S. Maashi, B. Garcia-Zapirain, M.S. Maashi, A new multi-agent feature wrapper machine learning approach for heart disease diagnosis. Comput. Mater. Contin **67**, 51-71 (2021)
12. X.Y. Gao, A. Amin Ali, H. Shaban Hassan, E.M. Anwar, Improving the accuracy for analyzing heart diseases prediction based on the ensemble method. Complexity (2021)
13. A.J. Swift, H. Lu, J. Uthoff, P. Garg, M. Cogliano, J. Taylor, D.G. Kiely, A machine learning cardiac magnetic resonance approach to extract disease features and automate pulmonary arterial hypertension diagnosis. Eur. Heart J. Cardiovascul. Imaging **22**(2), 236-245 (2021)
14. S.I. Ayon, M.M. Islam, Diabetes prediction: A deep learning approach. Int. J. Inf. Eng. Electron. Bus. **11**(2), 21–27 (Mar. 2019)
15. M. Gunasekaran, R. Varatharaian, M.K. Priyan, Hybrid recommendation system for heart disease diagnosis based on multiple kernels learning with adaptive neuro-fuzzy inference system. Multimed. Tools. Appl. **77**(4), 4379–4399 (2018)
16. M.K. Hasan, M.M. Islam, M.M.A. Hashem, Mathematical model development to detect breast cancer using multigene genetic programming, in 5th International Conference on Informatics, Electronics, and vision (ICIEV), Dhaka, pp. 574–9 (2016)

17. M.R. Haque, M.M. Islam, H. Iqbal, M.S. Reza, M.K. Hasan, Performance evaluation of random forests and artificial neural networks for the classification of the Liver disorder, in International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, pp. 1–5 (2018)
18. F.E. Ayo, R.O. Ogundokun, J.B. Awotunde, M.O. Adebiyi, A.E. Adeniyi, Severe Acne Skin Disease: A Fuzzy-Based Method for Diagnosis. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12254 LNCS, pp. 320–334 (2020)
19. J. Rasheed, A.A. Hameed, C. Djeddi, A. Jamil, F. Al-Turjman, A machine learning-based framework for the diagnosis of COVID-19 from chest X-ray images. Interdiscipl Sci Comput Life Sci **13**(1), 103–117 (2021)
20. Y. Kumar, G. Yadav, P.K. Singh, P. Arora, A PHR-based system for monitoring diabetes in mobile environment. In: S. Paiva (eds) Mobile solutions and their usefulness in everyday life. EAI/Springer Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-319-93491-4_7
21. N. Kumar, R. Iqbal, S. Misra, J.J. Rodrigues, Bayesian coalition game for contention-aware reliable data forwarding in vehicular mobile cloud. Futur. Gener. Comput. Syst. **48**, 60–72 (2015)
22. R.O. Ogundokun, P.O. Sadiku, S. Misra, O.E. Ogundokun, J.B. Awotunde, V. Jaglan, Diagnosis of long sightedness using neural network and decision tree algorithms. J. Phys. Conf. Ser. **1767**(1), 012021). IOP Publishing (2021)
23. R.K. Behera, M. Jena, S.K. Rath, S. Misra, Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data. Inform. Process. Manage. **58**(1), 102435 (2021)
24. K. Rangra, K.L. Bansal, Comparative study of data mining tools. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **4**(6), 216–223 (Jun. 2014)
25. S. Dwivedi, P. Kasliwal, and S. Soni, Comprehensive study of data analytics tools (RapidMiner, Weka, R tool, Knime), in Symposium on Colossal data analysis and Networking (CDAN), Indore, pp. 1–8 (2016)
26. P.K. Singh, Y. Singh, M.H. Kolekar, A.K. Kar, J.K. Chhabra, A. Sen, Recent Innovations in Computing, Proceedings of ICRIC 2020, Lecture Notes in Electrical Engineering book series (LNEE), volume 701, Springer (2020)
27. K. Chen, A. Mudvari, F.G.G. Barrera, L. Cheng, T. Ning, Heart murmurs clustering using machine learning, in 2018 14th IEEE International Conference on Signal Processing (ICSP), pp. 94–98 (2018)
28. E. Maini, B. Venkateswarlu, A. Gupta, Applying machine learning algorithms to develop a universal cardiovascular disease prediction system, in International Conference on Intelligent Data Communication Technologies and Internet of Things, pp. 627–632 (2018)
29. S. Kodati, R. Vivekanandam, G. Ravi, *Comparative Analysis of Clustering Algorithms with Heart Disease Datasets Using Data Mining Weka Tool, in Soft Computing and Signal Processing* (Springer, Singapore, 2019), pp. 111–117
30. K. Deepika, S. Seema, Predictive analytics to prevent and control chronic diseases, in 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT) (2016)
31. T.R. Reed, N.E. Reed, P. Fritzson, Heart sound analysis for symptom detection and computer-aided diagnosis. Simul. Model. Pract. Theory **12**(2), 129–146 (2004)
32. P. Amnarayan et al., Measuring the impact of diagnostic decision support on the quality of clinical decision making: development of a reliable and valid composite score. J. Am. Med. Informatics Assoc. **10**(6), 563–572 (2003)
33. C.-S. Lee, M.-H. Wang, A fuzzy expert system for diabetes decision support application. IEEE Trans. Syst. MAN Cybern. B Cybern. **41**(1), 139–153 (2011)
34. C.B. Rjeily, G. Badr, E. Hassani, E. Andres, Medical data mining for heart diseases and the future of sequential mining in medical field, in Machine Learning Paradigms, pp. 71–99 (2019)
35. K. Shameer, K.W. Johnson, B.S. Glicksberg, J.T. Dudley, P.P. Sengupta, Machine learning in cardiovascular medicine: are we there yet? Heart **104**(14), 1156–1164 (2018)

36. D. Tomar, S. Agarwal, A survey on data mining approaches for healthcare. Int. J. Bio-Sci. Bio-Technol. **5**(5), 241–266 (2013)
37. V.V. Ramalingam, A. Dandapath, M. Karthik Raja, Heart disease prediction using machine learning techniques: a survey, Int. J. Eng. Technol. **7**(2.8), 684–687 (2018)

An Analysis of Different Machine Learning Algorithms for Image Classification

Ankur Chaturvedi, Vikram Rajpoot, Meghansh Bansal, and Hanuman Das Agrawal

Abstract The aim of this paper is to bring together two areas, which are deep learning and machine learning for image classification. This paper checks or compares the accuracy level of any image classification dataset by using machine learning and deep learning algorithms. The target of this paper is the ability of a marked classifier for image classification. Machine learning is a branch of artificial intelligence that helps the system to learn automatically and improve with experience. Deep learning is a subfield of machine learning that involves learning from a large amount of data with complex computation. In this paper, we have used the plant village dataset. Plant village dataset is a collection of more than 50,000 expertly administered images on healthy and infected leaves of different crops. The images are 250 × 250 px. In this paper, machine learning algorithms such as K-nearest neighbors, logistic regression, support vector machines, and deep learning algorithms like ANN, CNN, and RBFNN are applied to the dataset and their performance in classifying the images has been compared.

Keywords Convolutional neural network · Radial basis neural network · K-nearest neighbors · Logistic regression · Support vector machine · k-means · Decision tree · Random forest

A. Chaturvedi · V. Rajpoot · M. Bansal (✉) · H. Das Agrawal
GLA University, Mathura, Utter Pradesh, India
e-mail: meghansh.bansal_cs.aiml19@gla.ac.in

A. Chaturvedi
e-mail: ankur.chaturvedi@gla.ac.in

V. Rajpoot
e-mail: vikram.rajpoot@gla.ac.in

H. Das Agrawal
e-mail: hanuman.agrawal_cs.aiml19@gla.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_30

1 Introduction

In today's time, the internet contains trillions of images and videos, which motivates the research of algorithms that can find and analyze the image or video to provide the user with the most accurate search with speed. There have been great milestones covered in image labeling, object classification, and image classification and recognition by many different researchers and scientists all across the globe. Since artificial intelligence and machine learning have shown unimaginable results in the area of object detection and image classification, many algorithms have been designed and tried out in these fields. Many different models using different techniques, parameters, and working theories have shown many different but promising results that are applicable in the real world. In the field of artificial intelligence, the neural network has great leads in image classification and object detection.

The focal point of this paper is on supervised learning algorithms such as logistic regression, support vector machines, random forest, K-nearest neighbors, decision tree, and deep learning algorithms like artificial, convolutional, and radial basis function neural networks. Logistic regression, SVM, KNN, decision tree, and random forest are used to calculate the dependent variable by finding the predictor variable of a certain dataset.

Source of Data: The data is taken from arXiv:1511.08060[cs.CY]. Plant village dataset is an open access repository of images on plant health.

2 Literature Survey

2.1 Image Data Pre-processing

The data acquired from different sources are generally very messy and required refining before giving it as input in different machine learning [1] models. It is the task of refining the image data by also creating new and improved data from the existing one. It allows us to work with different algorithms on the data [2]. In this paper, the authors have discussed many different techniques like the remote sensing classification process to improve classification and important factors that take part in this process like data pre-processing, size of the dataset, selection of features, and post-classification processing [3]. They have also performed image segmentation, edge detection, contour extraction, and grayscale conversion of images and then feed into the deep neural network to perform classification on the iris dataset.

The image is just an array of pixels ranging from 0 (zero) to 255. Image pre-processing includes converting colored images to the grayscale image to reduce computation but in exchange for some information. It also includes changing the size and shape of the image and standardizing [4]. The authors have discussed in this paper multiple image processing algorithms to add security while saving them. In addition, they have used python libraries pillow and OpenCV.

2.2 Methodology

The aim of this paper is to classify the images with deep learning algorithms and machine learning algorithms and their related tools. After data pre-processing of the image, we will select the best classifier for the image datasets which can classify the image based on their result and compare to evaluate which algorithms have the best performances for image classification problems [5, 6]. The authors have discussed the application and working of ANN in medical diagnosis [7]; have discussed different CNN model architectures, presented, and also have shown comparative analysis [8]; have applied RBFNN on top of different CNN architecture for improving the results of image classification [9]. In this paper, the authors have analyzed the random forest in the task of classification and regression and have given the conclusion that random forest performs well than other machine learning algorithms like logistic regression or linear regression [10]. They have used machine learning techniques reviews for diabetic prediction [11]. They have used a deep dense face detector, a method that does need segmentation, bounding box, and SVM classifier, and have suggested that the results can be further improved with data augmentation techniques [12]. The authors have created a deep layer convolutional network, used pre-trained model MCDNN, and applied it to MNIST, NIST SD 19, and many other datasets [13]. They have discussed ANN and its variations like KDD, PSO, GA, Fuzzy logic, ACO, ICS, ABC, with their limitations [14]. In this paper, the authors have discussed deep convolutional networks with a combination of other techniques on MNIST and CIFAR-10 datasets.

As deep learning has been performed in classifying the images [15], the machine learning algorithms have also given beautiful results in image classification [16]. The authors have applied three different machine learning algorithms that are decision tree, SVM, and KNN to classify the satellite images. They used Pleiades satellite images data in Taiwan. The result showed that SVM has the best accuracy [17]. The authors have used the iris flower database and applied many classification algorithms such as neural networks, AdaBoost, support vector machines, logistic regression, random forest, and KNN [18]. In this paper the authors have compared SVM and CNN for the task of image classification and have concluded that traditional methods perform well on small datasets while CNN is better for larger datasets.

After applying the classification models to the data, it is also necessary to check how your model has performed, so accuracy matrices are also an integral part of any project.

2.3 Classification Method

We have to use supervised learning algorithms for classification and a few unsupervised algorithms like k-means and also deep learning algorithms like convolutional neural network, artificial neural network, etc. For this, we have used python libraries like scikit-learn and Keras with TensorFlow as backend.

Artificial Neural Network

Artificial neural networks (ANNs), simply called neural networks, are the computation system designed basically inspired by the working of natural neurons in a human brain.

In 1943, Warren McCulloch and Walter Pitts gave the idea of a calculative model for neural networks. D. O. Hebb designed a learning hypothesis in the 1940s, which came to know as Hebbian learning. Rosenblatt created a perceptron. Today, artificial neural networks are used in regression, time series prediction, fitness approximation, classification tasks, pattern recognition, decision-making, data pre-processing, data filtering, etc.

In ANN the artificial neurons are called nodes which are fully connected. Every node works as a processing unit. Node one receives the input signal, processes it, and then passes it to all the connected nodes. The input is a real number and the output is produced by applying some nonlinear functions. The two nodes are connected with some weights; these weights get adjusted as the model gets learned or trained. The model can also have a threshold so that random and wild cases get ignored. Generally, these nodes are designed in the form of layers. Many different layers perform different functions on their input and pass it on to the next layer.

The first layer which is used to take in the data to train is known as the input layer; after that comes the layer which causes main computation and analysis and is known as hidden layers; and then comes the last layer which is used to produce output and known as output layer. Signals may transfer from the input layer to the output layer through hidden layers multiple times.

Components of Artificial Neural Network

- *Connections and weights*: Each connection from one layer to another layer provides the output from one neuron to the next layer. To each connection, weights are assigned that represent its relative importance. When value passes through the connection it gets multiplied by the weight and through learning the weights are changed to best fit the data.
- *Propagation function*: The propagation function is the function used to calculate the output upon input in an artificial neuron, and a bias term can be added to this.
- *Rectified linear unit layer*: The rectified linear unit layer exchanges any negative value in the pooling layer with zero.
- *Hyperparameter*: A hyperparameter is a constant term that is set before learning. Hyperparameters are like the number of hidden layers, rate of learning, input size, etc.

Model

The model is a sequential model:

1. *Flatten layer*: It flattens the passed input shape without changing the batch size (Table 1).

2. *Dense*: Dense layer gives the output = activation function (dot product (input, kernel) + bias)
   Output shape is (batch, number of units)
   Units: 3
   Activation: Rectified linear unit
   Kernel initializer: Glorot Uniform or Xavier normal initializer (extracts samples from a truncated normal distribution centered on zero with standard deviation = sqrt (2/(number input unit in weight + number of output unit in weight))
   Bias initializer: Zero matrix of the same dimension as input shape.
3. *Dropout layer*: The dropout layer randomly chooses and sets the input to zero. It helps in preventing the problem of overfitting. The inputs which are not set to zero are scaled up.
   Rate: 0.8
4. *Dense*:
   Units: 1
   Activation: Linear
   Model: "sequential"

Convolutional Neural Network

A convolutional neural network, also known as CNN, is a subdomain of deep neural networks, generally applied in visual imaginary. The applications are image recognition, recommender system, image classification, etc.

Convolutional neural networks are inspired by biological processes, from the working of the visual cortex of animals. In animals, the visual cortex is responsible for detecting light in their receptive field. The receptive field of distinct neurons in the visual cortex partially cover each other to create the whole visual.

A convolution network is made up of an input layer, hidden layer, and an output layer. The hidden layer consists of a series of convolutional layers that combine with multiplication or dot product. They work the same as cortex neurons in a living animal as they process the input and pass it to the next convolution layer as input.

In CNN, the neurons are not fully connected, rather it uses a 3D structure for the image to get analyzed in the specific region by neurons. The activation function is

Table 1 The layout of an artificial neural network model

Type of layer	Output shape	No. of parameters
flatten (Flatten)	(Batch size, Batch size)	0
dense (Dense)	(Batch size, 3)	589,827
dropout (Dropout)	(Batch size, 3)	0
dense_1 (Dense)	(Batch size, 1)	4

Total params: 589,831
Trainable params: 589,831
Non-trainable params: 0

common "ReLU", and then additionally connected to pooling layers, normalization layers, and fully connected dense layers.

The convolutional networks are used in many tasks which have great importance in many real-life applications. Recognition of handwritten digits was the first application where CNN was used. There has been a lot of improvement in the structure of CNN from that time.

Components of Convolutional Neural Network:

- *Input layer*: The first layer is the input layer in the convolutional neural network structure. This layer takes the data as input, resizes it, and passes it to the further layers for feature extraction. Input shape = (batch size, dimension one, dimension two, number of channels)
- *Convolution layer*: The next layer is the convolutional layer which is used for filtering images and to find out features from images and also use for calculating the match feature points during testing. The kernel passes over to a small area of the image and multiplies the filter value to the actual value of the pixel in the kernel selected part. The multiplications are summed up creating one single value or summary of that part of the image. The two-color dimensional image has only a two-dimensional matrix over which the two-dimensional filter move on. But with a three-dimensional color image, the filter need is to be three-dimensional to independently work on each of the color channels separately and then add the value to give the convoluted output.
- *Pooling layer*: The drawn-out features are then inputted into the pooling layer. This takes large images and diminishes them down while maintaining the most important knowledge in them. To diminish the images, it applies some defined functions like taking max value or average value, etc. to diminish the receptive field to a single value.
- *Batch normalization*: It is also known as the Batch norm, which is a method used to make ANN more stable and faster through normalization of the input by re-centering and rescaling. It normalizes by applying subtraction of the mean of the batch and division by the standard deviation of batch on the output of the previous layer. However, after this normalization, the weights are not optimized for the next layer. So, the batch normalization multiplies each layer by standard deviation term which is known as gamma, and mean deviation term which is known as beta.
- *Rectified linear unit layer*: The ReLU layer applies the rectified linear function on the input that converts all negative values into zero maintaining stability.

Model

Model is a sequential model (Table 2):

1. *Convolutional layer*: This layer creates a kernel that convolves with input to produce output with the same shape as input and channel as the number of units.
   Number of filters: 32
   Input shape: (64, 64, 3)

Table 2 The layout of the convolutional neural network model

Type of layer	Output shape	No. of parameters
Convolution	(Batch size, 62, 62, 32)	896
Batch normalization	(Batch size, 62, 62, 32)	128
Max pooling	(Batch size, 31, 31, 32)	0
Dropout	(Batch size, 31, 31, 32)	0
Convolution	(Batch size, 29, 29, 32)	9248
Batch normalization	(Batch size, 29, 29, 32)	128
Max pooling	(Batch size, 14, 14, 32)	0
Dropout	(Batch size, 14, 14, 32)	0
Flatten	(Batch size, 6272)	0
Dense	(Batch size, 32)	200,736
Batch normalization	(Batch size, 32)	128
Dropout	(Batch size, 32)	0
Dense	(Batch size, 32)	1056
Batch normalization	(Batch size, 32)	128
Dense	(Batch size, 1)	33

Total Parameters: 212,481
Trainable parameters: 212,225
Non-trainable parameters: 256

Activation function: Rectified linear unit
Padding: valid
Kernel initializer: Glorot Uniform
Bias initializer: Zero matrix

2. *Batch normalization*:
   Axis $= -1$
3. *Pooling*: Downsamples the input by choosing the maximum value in the window of set pool size along the feature axis.
   Pool size: (2, 2)
4. *Dropout layer*: The dropout layer randomly chooses and sets the input to zero. It helps in preventing the problem of overfitting. The inputs which are not set to zero are scaled up.
   Rate: 0.2
5. *Convolutional layer*:
   Number of filters: 32

6. *Batch normalization*:
   Axis = −1
7. *Pooling*:
   Pool size: (2, 2)
8. *Dropout layer*:
   Rate: 0.2
9. *Flatten layer*: It flattens the passed input shape without changing the batch size.
10. *Dense*: Dense implements the output = activation function (dot (input, kernel) + bias). Output shape is (batch, number of units)
    Units: 32
    Activation: Rectified linear unit
    Kernel initializer: Glorot Uniform or Xavier normal initializer (extracts samples from a truncated normal distribution centered on zero with standard deviation = sqrt (2/(number input unit in weight + number of output unit in weight))
    Bias Initializer: Zero matrix of the same dimension as input shape.
11. *Batch normalization*:
    Axis = −1
12. *Dropout layer*:
    Rate: 0.2
13. *Dense*:
    Units: 32
14. *Batch normalization*:
    Axis = −1
15. *Dense*:
    Units: 1
    Activation: Linear
    Model: "sequential"

Radial Basis Function Neural Network

Radial basis function neural networks are a type of neural network that uses radial basis functions as their hidden layer activation function. Broomhead and Lowe, researchers at Royal Signals and Radar Establishment, were the first ones to formulate it in their paper in 1988.

In a radial basis function neural network, layers can be divided into three parts; an input layer, hidden layers consisting of non-linear radial basis function, and a linear layer for output.

A radial basis function is that function which is real-valued and whose value only depends on the distance between the input and any fixed point in the cartesian so that f(x)=f (||x||), or some other point c, called the center so that f(x)=f(||c-x||). The function that satisfies this property is called a radial basis function. The radial basis

neural networks are trained in a two-step process. In the first step, centers and weights are determined using any clustering algorithm and in the second step, the weights between the hidden layers and output layer are calibrated in such a way that the error matrices are minimized over the whole dataset.

Component of Radial Basis Function Neural Network

In a radial basis function neural network, the starting is for the inputs of the network, the middle has hidden layers having radial nonlinear functions, and the final corresponds to the output. There are many activation functions but Gaussian functions are generally implemented.

Gaussian function:

$$f(x) = exp\left(\frac{-(x-p)^2}{2q^2}\right)$$

for some arbitrary real constants, p and q. Its name is given after the mathematician Carl Friedrich Gauss.

Now the output of the node k in the output section can be calculated by

$$y_k = \sum_{j=1}^{n} w_{jk}\phi_j(x)$$

Model

The model is a sequential model (Table 3):

1. *Flatten layer*: It flattens the passed input shape without changing the batch size.
2. *Radial basis function layer*:
   Number of units: 1
   Function: Gaussian
3. *Dropout layer*: The dropout layer randomly chooses and sets the input to zero. It helps in preventing the problem of overfitting. The inputs which are not set to zero are scaled up.
   Rate: 0.8

Table 3 The layout of radial basis function neural network model

Layer (type)	Output shape	Param #
Flatten	(Batch size, 196,608)	0
RBF layer	(Batch size, 1)	196,609
Dropout	(Batch size, 1)	0
Dense	(Batch size, 1)	2

Total Parameters: 196,611
Trainable parameters: 196,611
Non-trainable parameters: 0

4. *Dense*: Dense calculates the output = activation function (dot (input, kernel) + bias). Output shape is (batch, number of units)
   Units: 1
   Activation: Linear
   Model: "sequential"

Logistic Regression

Logistic regression is a classification algorithm based on a statistical model that uses a logistic function or sigmoid function to predict the probability of a categorical dependent variable. In this algorithm, we can find the decision boundary that separates the class. A threshold value can be set to predict, to which class the data belongs. Based on this threshold they obtained the estimated probability is classified into class for example in an email classification model, if predict value >= 0, 5, then the algorithm classifies the email as spam or not.

Support Vector Machine

It comes under the category of supervised learning technique which can be used for both classification and regression. But in this paper, we will use it as a classification. In this algorithm, each data item is a point in n-dimensional space with the value of each attribute we perform the classification task by SVM to find the optimal hyperplane that can classify the data into different classes.

Random Forest

It is the most flexible and easy algorithm used for the classification task. A forest is composed of a tree. It consists of a decision tree that operates as a whole. Each tree performs the task of splitting the class prediction. The class which has more votes is taken as a prediction of our model.

K-Nearest Neighbour

KNN algorithms have both implementations in classification and regression and it is easy to implement. They depend on the value of k. The best value of k can be found by the elbow method (where there is a considerable stop in the decreasing value of Euclidean distance). There is no need to build a training model.

Decision Tree

Decision tree is the most popular and powerful tool for classification problems. Its structure is just like a tree. Decision tree constructs nodes, branch, and leaf nodes. During the classification path from the root to the node, it follows to test the feature and make decisions along the path. It is expensive to construct and fastest for classifying unknown records.

3 Performance Analysis

For neural networks

Accuracy: Accuracy is one of the metrics for the evaluation classification model. Informally, accuracy is the fraction of predictions our model got right.

Loss: The loss function is one of the important components of the neural network. Loss is nothing but a prediction error of our neural network. There are many loss functions like mean square error, binary cross-entropy, categorical cross-entropy, sparse categorical cross-entropy, etc (Figs. 1, 2 and 3, Tables 4, 5 and 6).

Results

For logistic regression, SVM, random forest, KNN, supervised machine learning, decision tree algorithms:

Accuracy = TP + TN/TP + TN + FP + FN

F1score = 2 * recall * precision/recall * precision

Precision = TP/TP + FP

Recall = TP/TP + FP

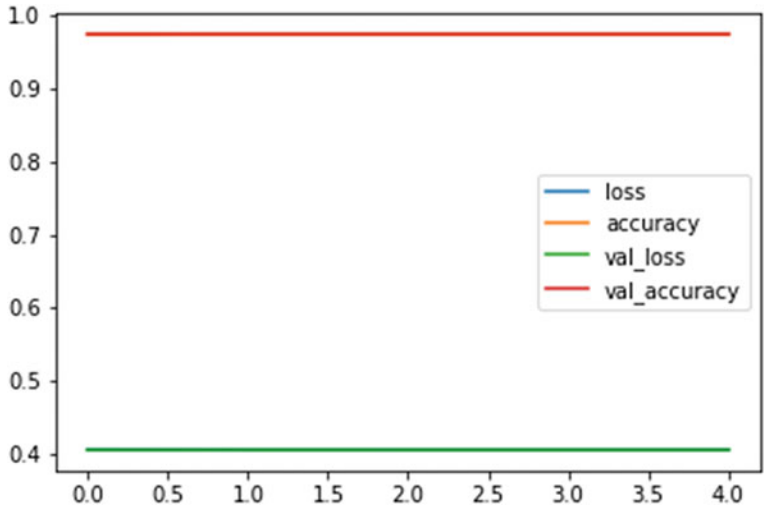


Fig. 1 Training phase of artificial neural network

Table 4 Results of the first model (artificial neural network)

Model	Training time	Accuracy of training	Loss in training	Accuracy over validation data	Loss over validation data
Artificial neural network	2 h 9 min 13 s	0.9737	0.4059	0.9737	0.4059
		Training data: 38,029 images		Validation data: 16,276 images	

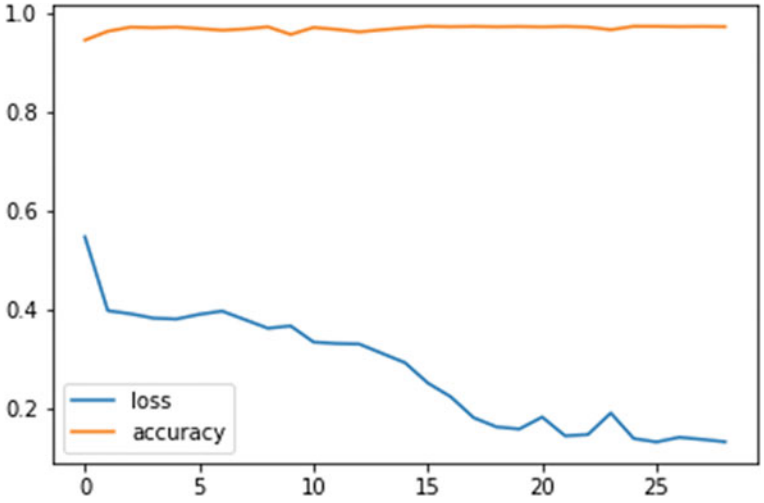


Fig. 2 Training phase of convolutional neural network

Table 5 Results of the second model (convolutional neural network)

Model	Training time	Accuracy of training	Loss in training	Accuracy over validation data	Loss over validation data
Convolutional neural network	4 h 17 min 51 s	0.9727	0.1327	0.9736	0.1272
		Training images: 43,442 images		Validation data: 10,863 images	

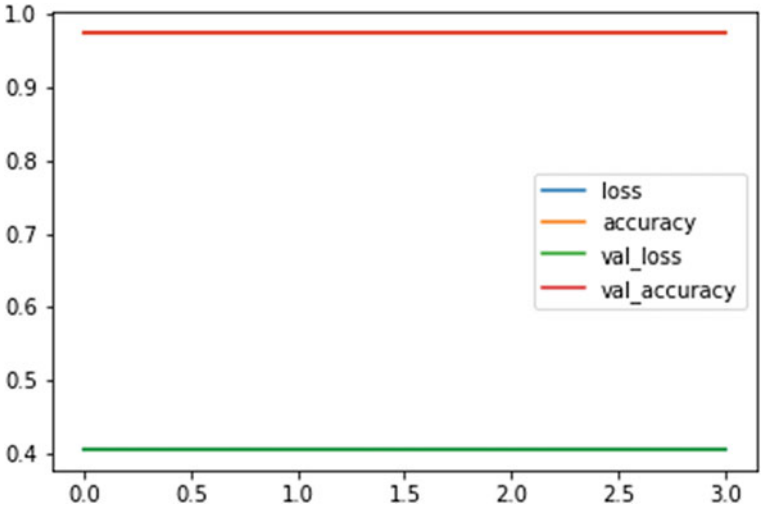


Fig. 3 Training phase of radial basis function neural network

Table 6 Results of the third model (radial basis function neural network)

Model	Training time	Accuracy of training	Loss in training	Accuracy over validation data	Loss over validation data
Radial basis function neural network	7 h 23 min 15 s	0.9737	0.4059	0.9737	0.4059
		Training data: 38,029 images		Validation data: 16,276 images	

Table 7 Results of all supervised machine learning algorithms

Models	Accuracy	F1 score	Precision	Recall
Logistic regression	0.8598	0.8588	0.8588	0.8588
Support vector machine	0.5923	0.5931	0.5933	0.5934
Decision tree	0.5521	0.5543	0.5520	0.5520
Random forest	0.55	0.55	0.55	0.55
KNN	0.4398	0.4391	0.4391	0.4391

Obtained result: No one model gives 100% accuracy. All models show an accuracy range between 50 and 90% (Table 7).

4 Conclusion

We analyzed the accuracy of all the models.

Now the result shows that logistic regression in machine learning algorithms and CNN in deep learning show the best performance in classifying the image, and decision tree, random forest, SVM, k-means algorithms performance of image classification are less. ANN and RBFN showed the same good accuracy but the loss was high compared to CNN which means they have shown more errors, while KNN performance is very bad and has very low accuracy. Convolutional neural network has given a very good accuracy. With all the simulation and discussion over working of different algorithms, it can be stated that convolutional neural networks have proved best in classifying images. Also, it is possible that by increasing the amount of data, accuracy can be increased and the chances of overfitting can be decreased.

References

1. D. Choudhary, S. Malasri Machine learning techniques for estimating amount of coolant required in shipping of temperature sensitive products. Int. J. Emerg. Technol. Adv. Eng. **10**(10), 67–70 (2020)

2. D. Lu, Q. Weng A survey of image classification methods and techniques for improving classification performance. Int. J. Remote Sens. **28**(5), 823–887 (2007)
3. W. Zhou, X. Ma, Y. Zhang, Research on image preprocessing algorithm and deep learning of iris recognition. J. Phys: Conf. Ser. **1621**, 012008 (2020). https://doi.org/10.1088/1742-6596/1621/1/012008
4. A. Goel, R.K. Bhujade, A functional review, analysis & comparison of position permutation based image encryption techniques. Int. J. Emerg. Technol. Adv. Eng. **10**(7), 97–99 (2020)
5. A. Egba, Okonkwo, R. Obikwelu, Artificial neural networks for medical diagnosis: a review of recent trends. Int. J. Comput. Sci. Eng. Surv. **11**, 1–11 (2020). https://doi.org/10.5121/ijcses.2020.11301
6. G. Chandra Mohan, S.S. Pattnaik, An ANN ensemble based ECG signal classification approach for accurate arrhythmia detection. Int. J. Emerg. Technol. Adv. Eng. **10**(8), 57–60 (2020)
7. S. Patel, A comprehensive analysis of convolutional neural network models. Int. J. Adv. Sci. Technol. **29**, 771–777 (2020)
8. M. Amirian, F. Schwenker, Radial basis function networks for convolutional neural networks to learn similarity distance metric and improve interpretability. IEEE Access, 1–1 (2020). https://doi.org/10.1109/ACCESS.2020.3007337
9. M. Schonlau, R.Y. Zou, The random forest algorithm for statistical learning. Stand. Genomic Sci. **20**(1), 3–29 (2020). https://doi.org/10.1177/1536867X20909688
10. A. Nahar, S. Sharma, Machine learning techniques for diabetes prediction: a review. Int. J. Emerg. Technol. Adv. Eng. **10**(3), 28–34 (2020)
11. S.S. Farfade, M.J. Saberian, L.J. Li, Multi-view face detection using deep convolutional neural networks. ICMR (2015)
12. D. Ciregan, U. Meier, J. Schmidhuber, Multi-column deep neural networksfor image classification (2012). arXiv:1202-2745
13. R. Bala, D. Kumar, Classification using ANN: a review. Int. J. Comput. Intell. Res. **13**(7), 973–1873 (2017)
14. Z. Sun, F. Li, H. Huang, Large scale image classification based on CNN and parallel SVM 2017 *International Conference on Neural Information Processing* (Springer, Cham, Manipal), pp. 545–555
15. Y. Peng, Z. Zheng, Spectral clustering and transductive SVM based hyperspectral image classification. Int. J. Emerg. Technol. Adv. Eng. **10**(4), 72–77 (2020)
16. I. Nurwauziyah, S. Umroh Dian, I.G.B. Putra, M.I. Firdaus, Satellite image classification using Decision Tree, SVM and k-Nearest Neighbor. Department of Geomatics, National Cheng Kung University, Tainan, Taiwan, July 2018
17. S.R. Alty, S.C. Millasseau, P.J. Chowienczyk, A. Jakobsson, Cardiovascular disease prediction using support vector machines. 376–379 (2006)
18. P. Wang, E. Fan, P. Wang, Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. Pattern Recogn. Lett. **141**, 61–67 (2021). ISSN 0167–8655, https://doi.org/10.1016/j.patrec.2020.07.042.

Biotic Disease Recognition of Cassava Leaves Using Transfer Learning

Rahul Sharma and Amar Singh

Abstract Healthy plants are essential for supporting life on Earth. Biotic plant diseases adversely impact food crops resulting in economical losses to farmers. Early and accurate diagnosis of plant diseases can aid farmers to take timely preventive measures. Accurately identifying disease symptoms and adopting corrective steps is a challenging task for the farmers. Small-scale farmers' livelihood depends upon agricultural outputs. Computer vision techniques can be employed to accurately detect plant diseases. In this paper, the transfer learning approach employing the pre-trained neural networks MobileNetV2 and ResNet50 is presented. The ResNet50 model outperformed MobileNetV2. ResNet50 model was able to classify cassava plant diseases with a training accuracy of 97.12% and validation accuracy of 94.67%. In addition, a hybrid ResNet-SVM model is also explored and a comparative analysis is done. Deep learning models using the transfer learning approach can be easily applied to other crops to significantly contribute to precision agriculture.

Keywords Computer vision · Transfer learning · Image classification · Feature extraction

1 Introduction

Cassava (Manihot esculentra) is the third-largest source of carbohydrates worldwide [1]. Cassava is a starchy tuberous root crop cultivated worldwide Common names of cassava are yuca, manioc [2], mogo, agbeli, tapioca, balinghoy, etc. Cassava is consumed as a staple food crop in different developing countries like Africa. The tuberous root is consumed either in boiled form or is used for the extraction of tapioca (starch). Cassava can grow in adverse weather conditions, degraded soil with less

R. Sharma (✉) · A. Singh
Lovely Professional University, Phagwara, Jalandhar, India

A. Singh
e-mail: amar.23318@lpu.co.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_31

agriculture value consuming very few resources, and still produce reasonable yields. Fields with good quality soil and higher levels of water supply can produce higher yields.

Cassava crop is mostly grown by small-scale, low-income, and smallholder farmers. Less resource-consuming cassava crop provides an opportunity for low-income farmers to earn more and at the same time contribute to meeting the rising demand of dried cassava and starch [3]. All these factors have contributed to the increased importance of cassava crop in the Agriculture sector. The yield of cassava crop is adversely affected by different biotic diseases like cassava mosaic diseases (CMD), cassava brown streak disease (CBSD), cassava bacterial blight (CBB), cassava green mite (CGM), etc. Outbreaks of CMD and CBSD can cause major or complete yield loss [4]. A lack of proper knowledge and inability to diagnose crop diseases results in economic loss to the farmers. Guidance from the field experts is not always possible due to various reasons. Moreover, different experts may not have consciences.

Empowering farmers by providing accessible, low-cost, accurate tools will aid them in timely detection of the disease and the adoption of appropriate preventive measures for controlling the spread of the disease. In today's digital age, a lot of digital content is generated every day. Moreover, the computational power of digital devices has improved. Computer vision techniques like deep neural networks can be used to extract and interpret complex features from a given dataset. Instead of relying on human expertise to diagnose plant diseases, convolutional neural networks can be trained to diagnose plant diseases.

The objective of the study is to demonstrate an image classification approach based on transfer learning models for the accurate prediction of cassava plant diseases. This paper presents the transfer learning approach where the previously learned knowledge acquired while solving the problem of one domain is utilized for solving another problem. Instead of developing a deep neural network model from scratch pre-trained models are applied. The knowledge acquired by a pre-trained model is utilized for providing an efficient and accurate prediction of cassava plant diseases. In this paper, we have used pre-trained models MobileNetV2 and ResNet50. MobileNetV2 is a lightweight model developed by Google and is designed to target mobile devices with less computation power.

This paper is organized into 6 sections. Section 1 gives the introduction of cassava plant diseases; Sect. 2 presents a literature review of the work done; Sect. 3 presents the overview of the cassava dataset used in the study. Section 4 presents the materials and methodology discussing the transfer learning approaches studied; Sect. 5 analyzes the results and Sect. 6 concludes the paper.

2 Literature Review

Ramcharan et al. proposed a transfer learning based approach for cassava plant disease detection [1]. The proposed model was applied to the dataset of cassava images collected from the fields of Tanzania and the disease images were classified with high accuracy. Ferentinos (2018) proposed deep learning based models trained using an 88 k image dataset for the detection of plant diseases and was able to achieve 99.53% accuracy [6].

Sangbamrung et al. (2020) proposed a CNN-based model [7] capable of identifying cbsc cassava disease with a precision of above 90%. Classification of cassava leaves into healthy or unhealthy (cbsd infected) was done with a 14 layer CNN model.

Ayu et al. (2021) developed a python GUI-based application utilizing the MobileNetV2 deep learning model [8] to classify plant diseases. The developed approach achieved 65.6% accuracy with test data. Sharma (2020) proposed two models using Tensorflow and python. Ist CNN model was trained using full images whereas 2nd CNN model was trained using the segmented image dataset. The CNN model trained using a segmented dataset performed with an accuracy of 98.6% [9].

Cristin (2020) presented an image classification approach. Input images were segmented using fuzzy C-means clustering [10]. HOG features were extracted from the segmented images which were used for image classification using a deep belief network optimized by rider optimization algorithm and Cuckoo Search. Chen et al. (2020) applied transfer learning using the VGGNet pre-trained CNN model and classified rice plants with an accuracy of 92% [11]. PlantDoc [12] dataset was compiled and presented by Singh et al. (2020). The dataset can be utilized for testing or developing intelligent systems for plant disease detection.

Kalvakolanu et al. (2020) proposed transfer learning using resnet 34 and resnet 50 to demonstrate a deep learning approach with 99.44% training accuracy [13].

Training reliable CNN-based models require a huge dataset which may not be technically or economically feasible for everyone. Argüeso (2020) FSL algorithm [14] capable of learning the new plant diseases images with a small dataset with high accuracy.

Rice diseases cause major crop loss. Wang (2021) presented an approach based on transfer learning and the Bayesian optimization method [15]. Koonce [18] presented a work discussing the importance of the ResNet50 network in solving different computer vision problems. The ResNet50 model is already trained on the Imagenet dataset. Instead of training the neural network from the scratch, the pre-trained weights can be downloaded and the final few layers are modified to swiftly develop an accurate classifier. Kumar et al. (2021) extracted the dataset features using K-means clustering and Otsu's threshold [21]. Features extracted to train Adaboost classifier and achieved 85% accuracy.

Various techniques/algorithms used by different researchers for image classification, feature extraction, and image enhancement are shown in Table 1.

Table 1 Techniques/algorithms used for image classification

Paper	Algorithms/pre-trained model used	Remarks
Xiao et al. [22]	EfficientNet-B3 Transfer Learning	Classification of jamming signals with 97.5% accuracy
Ghosh and Bandyopadhyay [23]	Proposed CNN Model based on VGG16 with fivefold cross-validation	Chest X-ray image classification with 99.39% accuracy
Zhang et al. [24]	Studied VGG16, InceptionV3, ResNet, DenseNet and proposed efficient CNN model	Parotid Tumor image classification with an accuracy of 97.78% and InceptionV3 with an accuracy of 96.11%
Pham et al. [25]	Compared ANN performance with well known pre-trained models like AlexNet, VGG16, and ResNet50 Feature Selection with Grey Wolf optimization method	Mango plant leaves disease detection with proposed CNN classified with 89.41% accuracy
Shrivastava et al. [26]	Studied 10 different pre-trained models AlexNet, VGG16, ResNet152V2, InceptionV3, etc	Best image classification by VGG16 with 93.11% accuracy
Alencastre-Miranda et al. [27]	ResNet101, AlexNet, VGG-16, and GoogLeNet	Efficient defect detection using AlexNet
Aversano et al. [28]	VGG-19, Xception, and ResNet50	VGG-19 performed with the top accuracy of 97%
Reddy et al. [29]	A comparative study was done using GaussianNB, SVM, SGD, LinearDiscriminant, RandomForest, LogisticRegression, DecisionTree and Proposed DNN model	For classification of the type of attacks in the IoT the proposed DNN model performed with an accuracy of 98.28%
Kaur et al. [30]	Soft computing approaches (Genetic Optimization, Fuzzy, Neural network, other Optimization algorithms like ABC, ACO, PCO, etc., were identified for image enhancement	Compared to traditional image enhancements methods better performance can be achieved using the soft computing based approached
Singh et al. [32]	Proposed a new soft computing-based approach three-parent genetic algorithm which is an extension of genetic algorithm	The proposed algorithm is used for optimal feature selection. The proposed approach outperforms other approaches for WMNs

(continued)

Table 1 (continued)

Paper	Algorithms/pre-trained model used	Remarks
Singh et al. [33]	Improved Parallel 3-Parent Genetic algorithm was proposed compared with 16 search and optimization algorithms	The benchmark performance of the algorithm was best in 14 out of 30 functions
Singh et al. [34]	A soft computing-based Parallel Big Bang Big Crunch optimization approach was proposed	A new efficient routing approach was proposed and was applied to the wireless mesh network

3 Dataset

In this paper, cassava plants dataset with 9436 labeled and 12,595 unlabeled images collected by Artificial Intelligence Lab, Makerere University [5]. The experts annotated the cassava images into five classes. Images with complex, overlapping backgrounds, different light conditions, low-resolution images, multiple diseases images, similar symptoms, etc., make learning and predicting the image classes a challenging task. Dataset of cassava leaves belong to five categories. Four of them are common cassava diseases and one healthy category. Cassava plants with diseases like cassava mosaic diseases (CMD), cassava brown streak disease (CBSD), cassava bacterial blight (CBB), cassava green mite (CGM) are collected and annotated by experts. In this paper, labeled images shown in Table 2 are used.

The dataset is split into 80–20 train-test split using 80% images are used for training the model and 20% images are used for validation and testing. All the dataset images are of size greater than 500 × 500 pixels, with a horizontal and vertical resolution of 96 dpi and bit depth of 24. During preprocessing, the images are resized to 224 × 224 × 3 pixels. Figure 1 shows the sample images of the dataset.

Table 2 Cassava dataset images per class

Class	Number of images
Cassava Mosaic Diseases (CMD)	2658
Cassava Brown Streak Disease (CBSD)	1443
Cassava Bacterial Blight (CBB)	466
Cassava Green Mite (CGM)	773
Healthy	316

(a) cassava bacterial blight (CBB)

(b) cassava brown streak disease (CBSD)

(c) cassava green mite (CGM)

(d) cassava mosaic diseases (CMD)

(e) Healthy

Fig. 1 Sample images of cassava dataset

4 Materials and Methodology

4.1 Transfer Learning

The CNN model learns with experience. The convolution layer (Conv2D) of the CNN model has a set of learnable filters. The kernel filter matrix is applied to the whole images of the training dataset. The model learns and updates the weights during

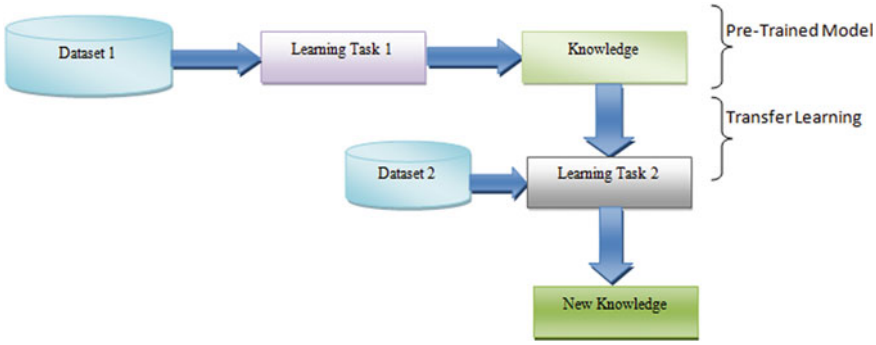


Fig. 2 Transfer learning approach

each training cycle. Huge datasets, computational power, fine-tuning of hyperparameters, domain knowledge, etc., are required for developing an accurate model from a scratch.

Transfer learning is really useful for solving problems in different but related areas [16]. In transfer learning, a pre-trained model is used as a starting point for the prediction [17] of a new task and is very useful for solving problems with small or inconsistent datasets. Figure 2 presents an overview of transfer learning.

4.2 Pre-trained Models (ResNet50 and MobileNetV2)

In this paper, we have applied transfer learning using ResNet50 neural network using Python-based computer vision and deep learning libraries. Residual Networks ease the training of very deep neural networks. ResNet152 model had 152 layers which is eight times the number of layers in the VGG net. ResNet50 neural network is trained on the Imagenet dataset. ResNet model makes use of skip connections which aid in reducing the problem of vanishing gradient. Skip connections of the ResNet model allow the output of earlier layers to be inputted to later layers skipping some intermediate layers. ResNet50 model proposed identity mapping and residual mapping [19]. ResNet 50 makes use of residual blocks with multiple convolution operations with activation maps (16, 128, 256, 512, 1024) [20].

MobileNetV2 is a lightweight memory-efficient model based on an inverted residual structure developed by Google and is designed to target mobile devices with less computation power [31]. MobileNet is smaller and faster in size than many of the other well-known pre-trained models. Due to the small disk size and less number of parameters this model is considered great for mobile devices. Table 3 shows the model size on disk and the number of parameters involved.

Table 3 Performance parameters of ResNet 50 and MobileNetV2 models

Model	Size	Top-5 accuracy on imagenet (%)	Parameters
ResNet50	98 MB	92.1	25,636,712
MobileNetV2	14 MB	90.1	3,538,984

4.3 Transfer Learning Hybrid Model (CNN-SVM)

CNN models are heavily used in computer vision problems. Fully connected layers of the CNN model perform the task of image classification. The logistic layer or Softmax layer is the last layer of the CNN model. For binary classification logistic layer is used whereas for a multiclass classification softmax layer is used.

In this paper, we have analyzed the hybrid CNN-SVM model. The CNN model architecture is ResNet and is used for feature extraction. The extracted features are inputted to the Support Vector Machine (SVM) model for image classification. The pre-trained model is already trained on large datasets. Features extraction using a pre-trained CNN model can be done.

1. Without further training the model using the new dataset
2. First training the CNN model on the new dataset and then extracting the image features

In this paper, the hybrid model is not trained on the cassava dataset. The already learned feature extraction capability of the pre-trained model is exploited. The MATLAB R2021a software is used for experimentation.

5 Results and Discussion

5.1 Transfer Learning Using ResNet 50 and MobileNetV2 Models

Pre-trained models are already trained on huge datasets (Imagenet). During training, the pre-trained model's weights are optimized and these weights are saved. In this paper, while using pre-trained models we have downloaded the weights of the ResNet50 and MobileNetV2 networks, modified the final layers of the models to classify leaf images of the cassava leaf dataset. Both the models are compiled with "binary_crossentropy" loss, "Adam" optimizer with a learning rate of 1e-5 m for 10 epochs. The model is retrained on the cassava plant dataset. Figures 3 and 4 present the training/validation accuracy and loss of ResNet50 and MobileNetV2 models, respectively.

The ResNet 50 model performed with the best training accuracy of 97.12 and 94.67 validation accuracy. After running for 10 epochs the MobileNetV2 model performed with the best validation accuracy of 92.42% and training accuracy of 94.02%.

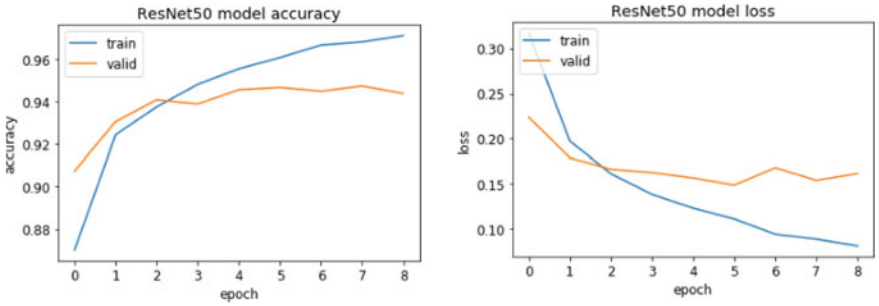


Fig. 3 ResNet50 Training/validation accuracy and loss using cassava plant dataset

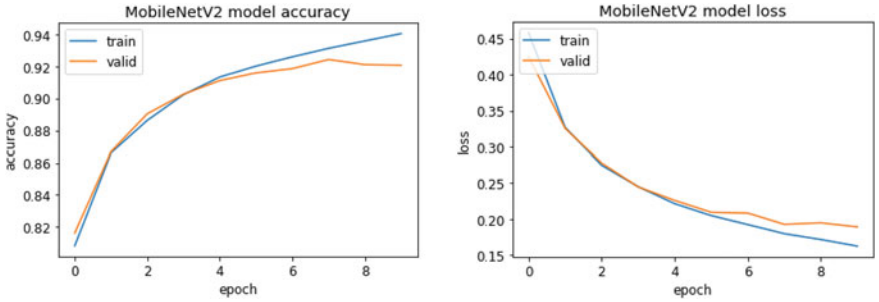


Fig. 4 MobileNetV2 Training/validation accuracy and loss using cassava plant dataset

Randomly selected unseen test images of infected and healthy cassava plant leaves were inputted into the trained models for the classification of the diseases. The confusion matrix of both models is shown in Fig. 5. The confusion matrix is obtained for the five classes of the cassava image dataset.

ResNet50 model performs better than the MobileNetV2 model. Different performance parameters for comparative analysis of these models like accuracy, precision, recall, and f1 score are shown in Table 4.

Figure 6 shows the comparative validation accuracy and loss of ResNet50 and MobileNet V2 models.

5.2 Transfer Learning Hybrid Model CNN-SVM

The ResNet-SVM integrated approach was employed for the classification of cassava leaves shown in Fig. 7. In the first step, the ResNet model was used for feature extraction. The fully connected layers of ResNet were removed and the important features extracted by the "pool5" layer of the ResNet model were obtained. In the second step, the extracted features were inputted to the SVM for final classification.

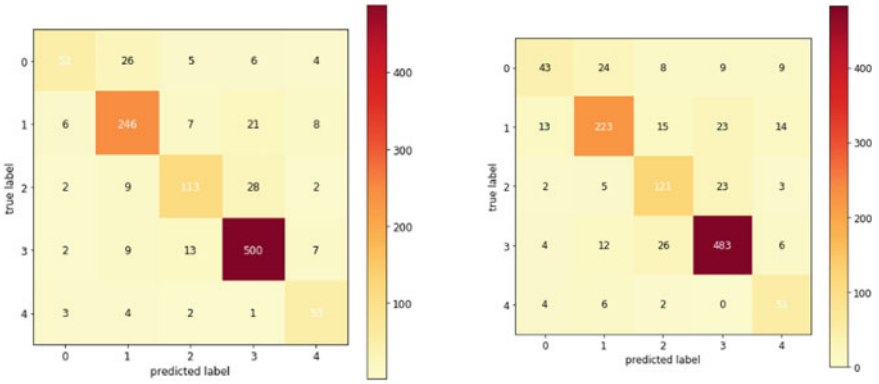


Fig. 5 Confusion matrix of pre-trained models

Table 4 Performance parameters of ResNet 50 and MobileNetV2 models

Model	Validation accuracy	Test accuracy (%)	Precision	Recall	F1 score
ResNet50	94.67	85.385	81.187	78.599	79.319
MobileNetV2	92.42	81.576	73.863	74.83	73.695

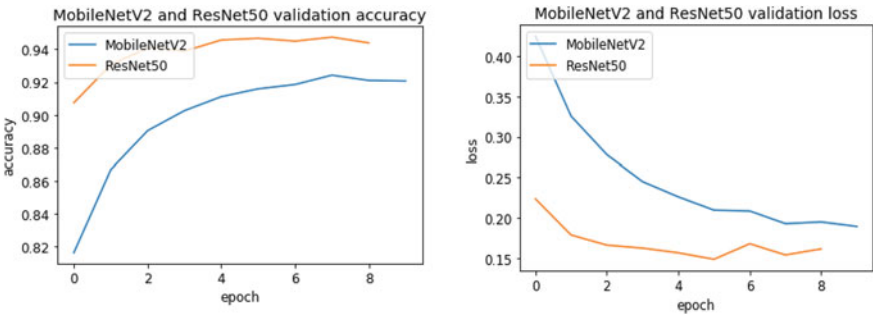


Fig. 6 Comparative training graphs of ResNet50 and MobileNetV2 models

The CNN-SVM model classified test images of cassava leaves with an accuracy of 72.58%. The ResNet model architecture is shown in Fig. 8.

The ResNet-SVM model accuracy can be improved by adopting different optimization approaches. Features extracted by the ResNet model can further be improved by first training the ResNet neural network using the cassava dataset and then using the trained model for feature extraction.

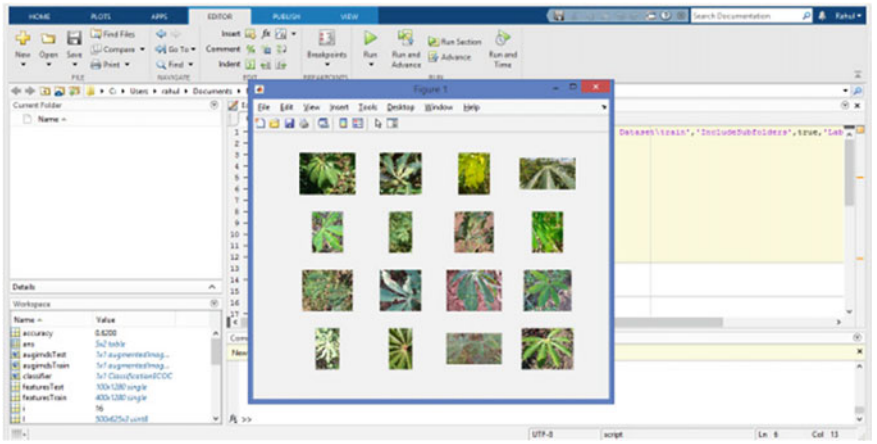


Fig. 7 Cassava dataset sample images

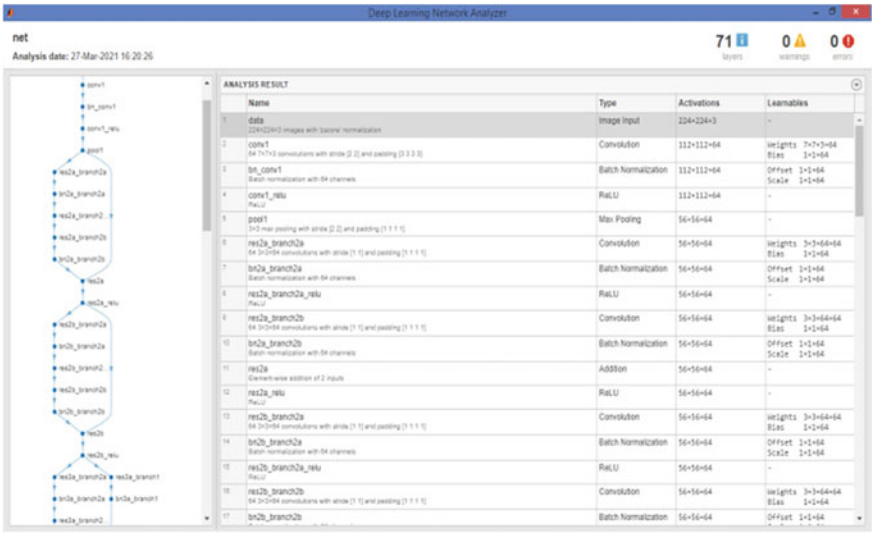


Fig. 8 ResNet model analyzed using MATLAB

6 Conclusions

In this paper, transfer learning using the pre-trained network ResNet50 and MobileNetV2 models is proposed for the classification of cassava plant diseases. The experiment results demonstrate that the transfer learning approach is very well suited for the detection of cassava plant diseases on the unbalanced dataset. The ResNet-SVM hybrid model was also analyzed. The performance of the model can

further be improved by optimizing the feature extraction. Comparative performance of the models reveals that the ResNet50 model with Adam optimizer performs better when compared with the MobileNetV2 model. In the future, work on identifying the severity of the cassava plant disease can be worked. In addition, various soft computing based approaches can be used for feature extraction and fine-tuning of the hyperparameters of the pre-trained models. Also if the plant leaves are infected by more than one type of disease, the current model will not be able to predict the diseases properly. Image segmentation along with transfer learning needs to be explored in the future.

References

1. A. Ramcharan, K. Baranowski, P. McCloskey, B. Ahmed, J. Legg, D.P. Hughes, Deep learning for image-based cassava disease detection. Front. Plant Sci. **8**, 1852 (2017)
2. V. Kuete, Physical, hematological, and histopathological signs of toxicity induced by African medicinal plants, in *Toxicological Survey of African medicinal plants*. (Elsevier, 2014), pp. 635–657
3. R. Howeler, N. Lutaladio, G. Thomas, *Save and grow: cassava. A guide to sustainable production intensification*. (Fao, 2013)
4. J.M. Thresh, G.W. Otim-Nape, J.P. Legg, D. Fargette, African cassava mosaic virus disease: the magnitude of the problem. Afr. J. Root Tuber Crops **2**(1/2), 13–19 (1997)
5. E. Mwebaze, T. Gebru, A. Frome, S. Nsumba, J. Tusubira, iCassava 2019 fine-grained visual categorization challenge (2019). arXiv:1908.02900
6. K.P. Ferentinos, Deep learning models for plant disease detection and diagnosis. Comput. Electron. Agric. **145**, 311–318 (2018)
7. I. Sangbamrung, P. Praneetpholkrang, S. Kanjanawattana, A novel automatic method for cassava disease classification using deep learning. J. Adv. Inf. Technol. **11**(4) (2020)
8. H.R. Ayu, A. Surtono, D.K. Apriyanto, Deep learning for detection cassava leaf disease. *In Journal of Physics: Conference Series* (Vol. 1751, No. 1, p. 012072). (IOP Publishing 2021)
9. P. Sharma, Y.P.S. Berwal, W. Ghai, Performance analysis of deep learning CNN models for disease detection in plants using image segmentation. Inf. Process. Agric. **7**(4), 566–574 (2020)
10. R. Cristin, B.S. Kumar, C. Priya, K. Karthick, Deep neural network based Rider-Cuckoo Search Algorithm for plant disease detection. Artif. Intell. Rev. 1–26 (2020)
11. J. Chen, J. Chen, D. Zhang, Y. Sun, Y.A. Nanehkaran, Using deep transfer learning for image-based plant disease identification. Comput. Electron. Agric. **173**, 105393 (2020)
12. D. Singh, N. Jain, P. Jain, P. Kayal, S. Kumawat, N. Batra, PlantDoc: a dataset for visual plant disease detection. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD* (2020), pp. 249–253
13. A.T.S Kalvakolanu, Plant disease detection from images. arXiv:2003.05379 (2020)
14. D. Argüeso, A. Picon, U. Irusta, A. Medela, M.G. San-Emeterio, A. Bereciartua, A. Alvarez-Gila, Few-shot learning approach for plant disease classification using images taken in the field. Comput. Electrn. Agric. **175**, 105542 (2020)
15. Y. Wang, H. Wang, Z. Peng, Rice diseases detection and classification using attention based neural network and bayesian optimization. Expert Syst. Appl. 114770 (2021)
16. L. Torrey, J. Shavlik, Transfer learning. In *Handbook of Research On Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. (IGI global, 2010), pp. 242–264
17. H.C. Shin, H.R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, R.M. Summers, Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging **35**(5), 1285–1298 (2016)

18. B. Koonce, ResNet 50. In *Convolutional Neural Networks with Swift for Tensorflow*. (Apress, Berkeley, CA, 2021), pp. 63–72
19. B. Li, D. Lima, Facial expression recognition via ResNet-50. Int. J. Cognit. Comput. Eng. (2021)
20. M. Loey, G. Manogaran, M.H.N. Taha, N.E.M. Khalifa, Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection. Sustain Cities Soc **65**, 102600 (2021)
21. C.S. Kumar, V.K. Sharma, A.K. Yadav, A. Singh, Perception of plant diseases in color images through adaboost. In *Innovations in Computational Intelligence and Computer Vision* (Springer, Singapore, 2021), pp. 506–511
22. Y. Xiao, J. Zhou, Y. Yu, L. Guo, Active jamming recognition based on bilinear EfficientNet and attention mechanism. IET Radar Sonar Navig (2021)
23. S. Ghosh, M. Bandyopadhyay, Detection of coronavirus (COVID-19) using deep convolutional neural networks with transfer learning using chest X-ray images. Mach. Learn. Approach. Urban Comput. **3**, 63 (2021)
24. H. Zhang, H. Lai, Y. Wang, X. Lv, Y. Hong, J. Peng, C. Chen, Research on the classification of benign and malignant parotid tumors based on transfer learning and a convolutional neural network. IEEE Access **9**, 40360–40371 (2021)
25. T.N. Pham, L. Van Tran, S.V.T. Dao, Early disease classification of mango leaves using feed-forward neural network and hybrid metaheuristic feature selection. IEEE Access **8**, 189960–189973 (2020)
26. V.K. Shrivastava, M.K. Pradhan, M.P. Thakur, Application of pre-trained deep convolutional neural networks for rice plant disease classification. in *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)* (IEEE, 2021), pp. 1023–1030
27. M. Alencastre-Miranda, R.M. Johnson, H.I. Krebs, Convolutional neural networks and transfer learning for quality inspection of different sugarcane varieties. IEEE Trans. Industr. Inf. **17**(2), 787–794 (2020)
28. L. Aversano, M.L. Bernardi, M. Cimitile, M. Iammarino, S. Rondinella, Tomato diseases classification based on VGG and transfer learning. in *2020 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor)* (IEEE, 2020), pp. 129–133
29. D.K. Reddy, H.S. Behera, J. Nayak, P. Vijayakumar, B. Naik, P.K. Singh, Deep neural network based anomaly detection in Internet of Things network traffic tracking for the applications of future smart cities. Trans. Emerg. Telecommun. Technol. e4121 (2020)
30. G. Kaur, N. Bhardwaj, P.K. Singh, An analytic review on image enhancement techniques based on soft computing approach. in *Sensors and Image Processing* (Springer, Singapore, 2018), pp. 255–265
31. 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*, pp. 4510–4520 (2018)
32. A. Singh, S. Kumar, A. Singh, S.S. Walia, Three-parent GA: a global optimization algorithm. J. Mult Valued Logic Soft Comput **32** (2019)
33. A. Singh, S. Kumar, A. Singh, S.S. Walia, Parallel 3-parent genetic algorithm with application to routing in wireless mesh networks. in *Implementations and Applications of Machine Learning* (Springer, Cham, 2020), pp. 1–28
34. S. Kumar, A. Singh, S. Walia, Parallel big bang-big crunch global optimization algorithm: performance and its applications to routing in WMNs. Wireless Pers. Commun. **100**(4), 1601–1618 (2018)

A Sentiment Detection Tool for Multiple Domains

Priya Shrivastava and Dilip Sharma

Abstract The sentiment is opinion or thought about a particular product or a specific situation. Sentiment analysis is applied to extract the subject information from text. Sentiment is computed using feature level, entity level, sentence level, document level. This paper proposed a sentiment detection tool that detects the sentiment (positive or negative) of any input comment from the web in multidomain-based on sentence-level and document level (text analysis) by analyzing the Polarity score of each and individual word. This work will allow users to choose the type of content they want to read or focus on. It will help in analyzing what kind of impact that text is going to leave on the reader. This will help in detecting the Polarity of text. We designed a sentiment detection tool to obtain the Polarity of sentence and document. This work aims to divide or distribute the content according to their sentiments.

Keywords Artificial intelligence · Polarity detection · Sentiment analysis · Web comments

1 Introduction

Sentiment analysis is a technique that is applied to text to determine whether it is positive, negative, or neutral. This procedure helps in collecting customer's experience about a particular product. Sentiment analysis has a wide range of applications because views are central to almost all human activities; whenever we need to decide, we look for others' opinions. This project is aimed to check the sentiments and Polarity in a sentence. In this paper, we consider text-domain only. There is an increase in the use of digital media. Peoples use social media platforms to share their reviews about a particular entity. The customer reads these reviews before buying any product. So, it is essential to determine the sentiments of the text.

P. Shrivastava (✉) · D. Sharma
GLA University Mathura, Mathura, India

D. Sharma
e-mail: dilip.sharma@gla.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_32

We will analyze the user-generated text and predict the sentiments of the text, whether it is positive or not. This paper will help peoples to filter the content they want to read or buy. It will help people to check the emotion behind any content. It will determine the attitude of the speaker or a writer on the topic.

For Ex: This movie is amazing.

This sentence represents a favorable opinion of the movie. Like this example, first, we will determine the sentiment of the sentence. Then it will calculate its Polarity, which will tell that the document is how much percent positive sentiment the document or sentence contains.

Polarity = Positive sentiment.

Total Sentiments = 1.

1.1 Applications

Our sentiment analysis can be used for market research and to determine the demand for a particular product. It can be used to extract detailed information to monitor the selling pattern of a specific brand. Sentiment analysis is helpful in various domains like

- **Social media monitoring**: As social is growing day by day, its horizons are becoming wider. So, social media, blogging platforms are the platform to share opinions. So, it is essential to recognize what type of opinion or emotions that person was trying to convey or that content is even healthy or not. Social blogs we read have an impact on our mind that can be positive or negative.
- **Customer support (Analyzing reviews)**

  We check other reviews and opinions before taking any decision. So, it's difficult to check what type of responses they are? If we are purchasing something online, there are types of reactions that are difficult to judge. So this can help us by telling the Polarity of reviews.
- **Individual Consumers**

  Opinion mining is helpful in the recommendation system. Individual consumer buys the product based on the review of other consumers on the same product. Using this analysis, a consumer can do comparative analysis [1]. Authors in the paper [12] proposed a framework for consumer review analysis. This study discussed the behavior and concern of the consumer.
- **Question Answering** [6, 7]

  The idea behind building a question answering system is to automatically answer the questions asked by human/sentiment analysis is useful to depict the sentiment of questions and analyze how the sentiments affect the overall answer. It is

essential to include the semantic features to improve the quality of the question answering system.

- **Online Advertising** [1]

  Opinion mining behaves like a marketing tool that can automatically extract the user's opinion based on the user's review. It plays a vital role in the online advertising of a particular product. Advertisements may be based on customer's interests and ratings.

- **Detection of Flames** [6, 6]

  Sentiment analysis helps determine flames, use of hatred speech, and negative words. The sensitivity of the topic can be computed by opinion mining.

Figure 1 shows the different multimodal sentiment analysis application [24]. Nowadays, users mainly depend on the internet world where they share their feelings, suggestions, and thoughts on the social platform, which will help them where they choose their favorite item or material from them. But the sentiment analysis has both cons and pros where they help out to detect the fake news from social media and other platforms. On the other side, many people generate rumors or fake news to mislead users to make their profit.

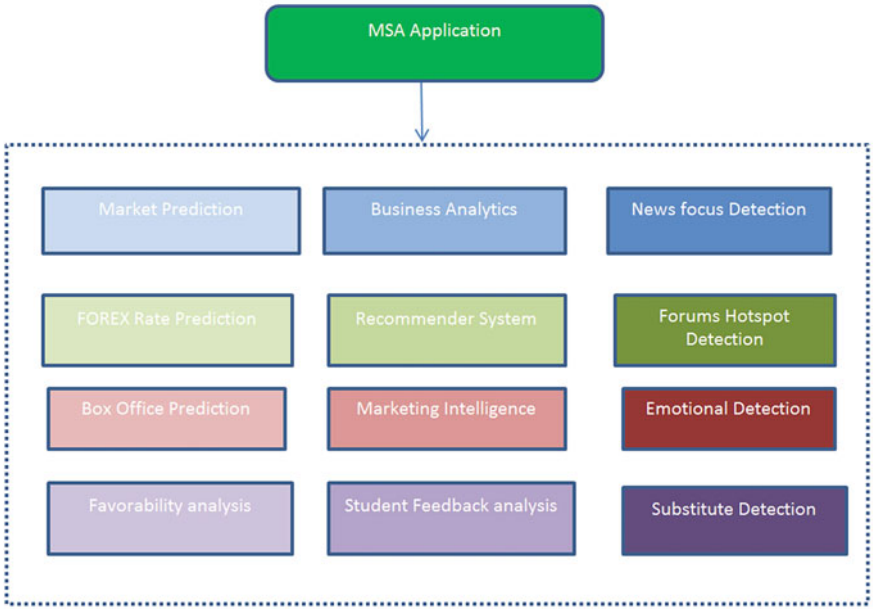


Fig. 1 Multimodal Sentiment analysis application

2 Literature Review

Text can exist in two forms. It may be subjective or objective. Sentiment analysis is applied to a subjective entity. A lot of work has been done by many researchers in this field. Authors in paper [1] explained the application areas where sentiment analysis can be applied. They adopted a supervised machine learning approach. They used a movie review dataset to classify whether it is positive or negative. The Tf-idf approach is applied for word embedding, but the limitation of this approach is that it cannot consider the distributions of terms across various documents. Authors in paper [2] expanded this work by using rating scales for sentiment computation. They used the bag of words (unigram) approach for feature extraction. They applied machine learning models like naïve Bayes and SVM and achieved better results. Authors in paper [3], used seed adjectives for sentiment computation. They have generated more positive and negative adjectives by recursively finding synonyms and antonyms of these seed adjectives from WordNet. This approach performed better in comparison to existing manually picked word lists. In paper [4], the authors used tweets extracted from some selected investors Twitter feeds. They assigned Polarity to these tweets based on opinion words of tweets. This method worked well for stock analysis. Authors in the paper [15] suggested a novel approach for creating a stock market lexicon. This is used to predict Twitter investors' sentiment. Authors in the paper [5] used approx 5.1million product reviews dataset from Amazon.com. A max entropy POS tagger is applied to identify the tags of words and apply various machine learning models like Naïve Bayesian, RandomForest, and Support Vector Machine to predict the review's sentiment.

In this paper [8], the researchers applied the mutual information technique to determine the document's Polarity. It is based on the assumption that the same opinion word has the same Polarity in the same domain. Still, in the paper [9], the authors assumed that the same opinion word has different sentiments in the same domain. They used Conditional Mutual Information (CMI) method to compute the Polarity of documents. Authors in a study [10] worked on context-dependent opinion words. Context-dependent words are the words which have different meanings in a different context. They used an online dictionary to generate context-dependent words, and an interaction information method was used to find the sentiments. Researchers in paper [11], used a word sense disambiguation algorithm for sentiment detection. The authors applied an extended gloss overlap algorithm, and the accuracy increased to 60%. Authors in paper [13] proposed a language parser. They designed the parser for phrase sentiment detection. Researchers in paper [14] use an unannotated corpus. They used a sentiment dictionary. They consider domain-specific sentiment identification [16]. This paper is concerned with reducing users' time and efforts for searching the products using recommendation and proposed a Naïve Bayes classifier that computes the reviews accurately and combines them to give a final rating for the product. Paper [17] discusses different techniques and algorithms for text classification and summarises the cons and pros of the different methods [18].

This paper proposed a model for Polarity estimation of user reviews using aspect-based sentiment analysis; this model has three phases data preprocessing, aspect co-occurrence calculation and Polarity estimation using multimodal sentiment dataset and Twitter dataset [19]. This paper proposed framework commitments which identify the member of any specific community for a particular event by using automated snowball sampling scores to identify community members and publish content 20]. This paper discusses different models and their features values and datasets and presents a comparative analysis for the different datasets [22]. This paper proposed a grey-box adversarial attack and defence framework for sentiment classification and discussed above the differentiability and reconstruction for adversarial attack and defence [23]. This paper investigates to improve the interpretability of deep learning neural networks for natural language processing tasks using machine translation and sentiment analysis. Papers [25–34] also discussed various issues, applications, types, frameworks, and algorithms related to sentiment analysis and option mining.

3 Proposed Algorithm

The Proposed algorithm includes the following steps.

(1) Created two lists. First, one is for positive sentiment words, and the other one is for negative sentiment words.
(2) Then, we extract users' comments or reviews from the webpage for Polarity identification.
(3) Extract the opinion words from these reviews and perform matching.
(4) Scoring the Polarity for each sentence.
(5) At the end, Polarity score is displayed to the user as a result.

3.1 Architecture for Proposed Work

The proposed architecture is shown in Fig. 2.

In this work (Fig. 2.), first, we extract the comments from the webpage and separate the sentences using firebase, then extract opinion words to prepare the verb or term list from comments. These opinion words are a match with an existing manually created lexicon, and separate into the three lists: positive, negative, and neural then compute Polarity for each sentence separately and then resemble the total Polarity score for the input value and at last classify the comments into positive and negative sentiment word for suggesting the user.

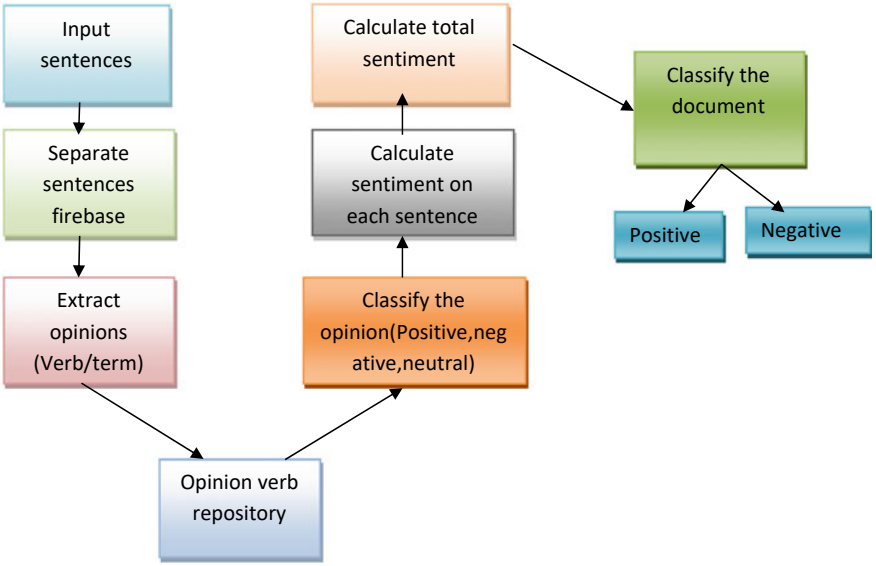


Fig. 2 Representation of proposed architecture

3.2 Dataset

We use IMDB dataset [21] from Kaggle, having a 50 k movie review for natural language processing. Table 1 represents the parameters of dataset.

This dataset used primarily for binary sentiment classification. The parameter values iare shown in Table 1. This dataset provides positive and negative movie reviews with 25,000 highly polar movie reviews for training and 25,000 for testing. Dataset and their levels are shown in Fig. 3.

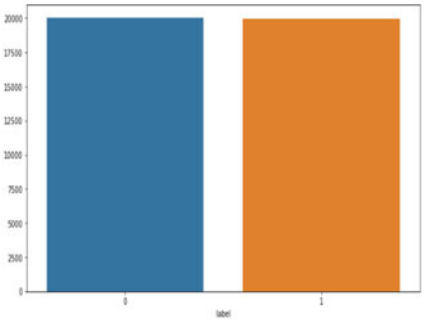
We can clearly see some differences between positive and negative movie reviews separated by training and testing dataset between these two word-clouds as shown in Fig. 4.

Table 1 Detail of dataset

Parameter	Values
No. of negative reviews	25,000
No. of positive reviews	25,000
Rating criteria	1–5
Size of dataset	25,000 × 25,000

	text	label
0	I grew up (b. 1965) watching and loving the Th...	0
1	When I put this movie in my DVD player, and sa...	0
2	Why do people who do not know what a particula...	0
3	Even though I have great interest in Biblical ...	0
4	Im a die hard Dads Army fan and nothing will e...	1
5	A terrible movie as everyone has said. What ma...	0
6	Finally watched this shocking movie last night...	1
7	I caught this film on AZN on cable. It sounded...	0
8	It may be the remake of 1987 Autumn's Tale aft...	1
9	My Super Ex Girlfriend turned out to be a plea...	1

Dataset Screenshot



Dataset label

Fig. 3 Screenshot and label of dataset

Positive movie reviews word-cloud

Negative movie reviews word-cloud

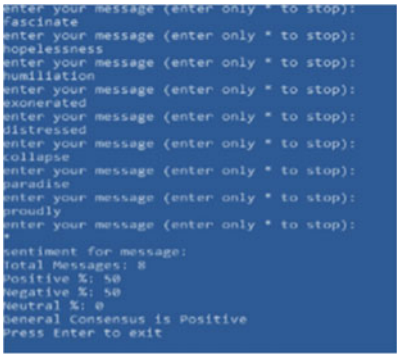
Fig. 4 Word-cloud of positive and negative movie reviews

4 Tools Demo Results

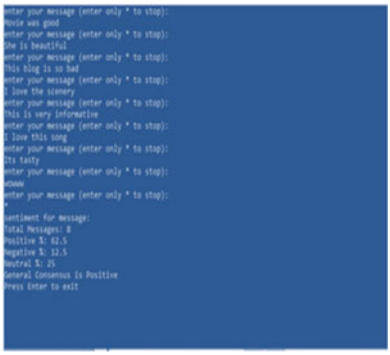
To design the sentiment prediction tools, we consider simple words, sentences, and phrases for sentiment detection. Our tool is also able to predict movie reviews and app reviews. This tool performed well on social media comments. These are some results that our tool obtained.

5 Result Analysis

For sentiment prediction, we work over different sceneries that were explained in Fig. 5a–c to check the sentiment among word, sentence, and paragraph, respectively, where we identify the word, sentence, and paragraph with predefined lexicon values and compute the Polarity for each phase and give a better result per the user demand.

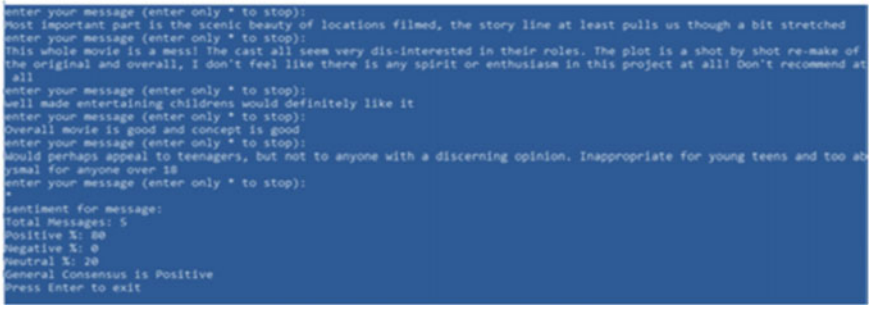


(a)



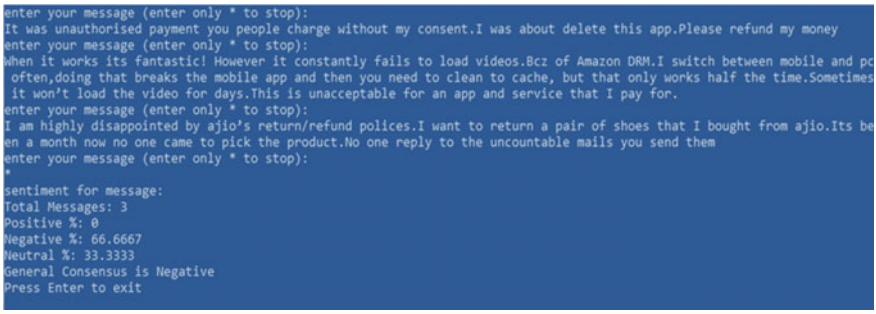
(b)

(c)

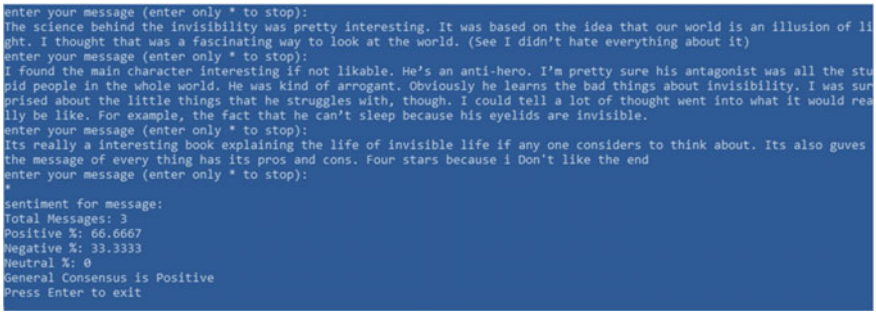


(d)

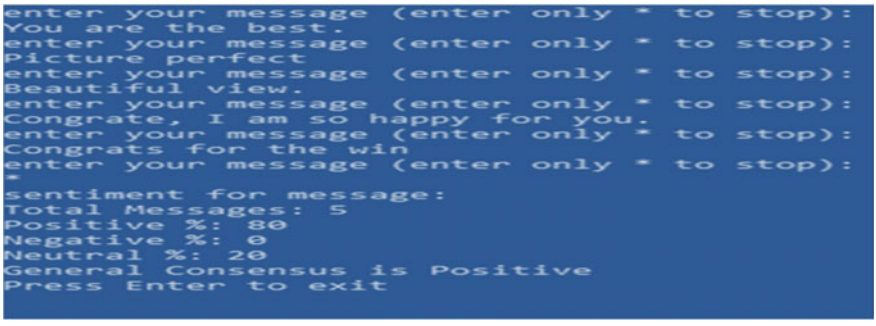
Fig. 5 **a** Word sentiment check **b** Sentence sentiment check **c** Paragraph sentiment check **d** Movie review sentiment. **e** App review sentiment check **f** Book review sentiment detection **g** Social media comment sentiment detection **h** Article review sentiment check **i** Online shopping review sentiment check



(e)



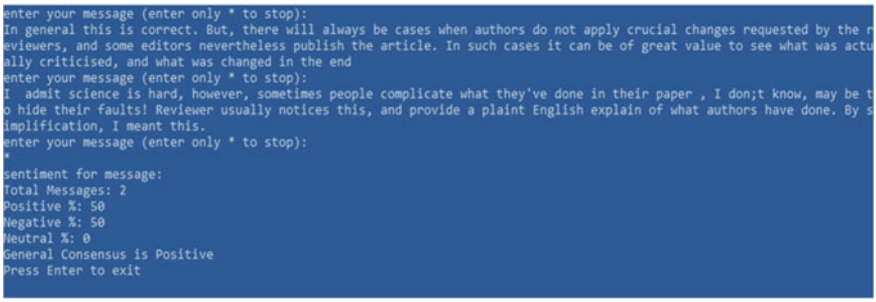
(f)



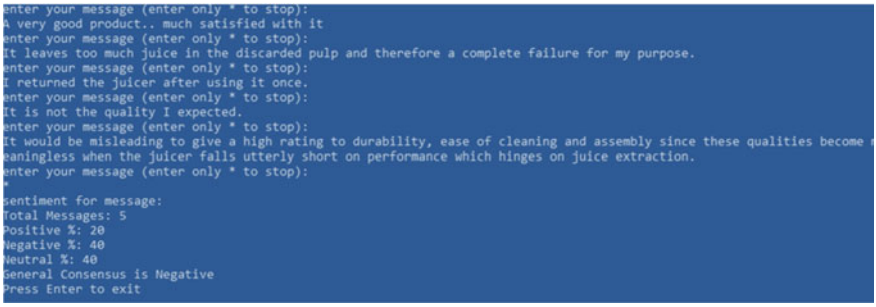
(g)

Fig. 5 (continued)

Over this, we work on the movie reviews and explore the reviews in negative and positive using Polarity score whose results are shown in Fig. 2d as same as we use our tool for Book reviews, article reviews, app reviews, online social reviews, and also on social media comment and all those are shown in Fig. 5g–i, respectively.



(h)



(i)

Fig. 5 (continued)

6 IMDB Movie Reviews Dataset Results

On the other side, we use the IMDB movie reviews dataset for sentiment detection, So Fig. 6a shows 80% accuracy by computing tf-idf vectorization using tf-idf Gaussian Naïve Bayes, Fig. 6b scores 73% accuracy using N-gram and GNB machine learning classifier and Fig. 6c shows the results having 80% accuracy using the AdaBoost classifier.

6.1 Conclusion and Future Work

In this paper, we designed a tool to determine the sentiment analysis over text document in multidomain and in future, we can use the multi-models among neural networks and transformers.

```
Valid Accuracy: 0.7978
Test Accuracy: 0.8056
[[1990  505]
 [ 467 2038]]
Classification Report:                precision    recall  f1-score   sup
port

           0       0.81      0.80      0.80      2495
           1       0.80      0.81      0.81      2505

    accuracy                           0.81      5000
   macro avg       0.81      0.81      0.81      5000
weighted avg       0.81      0.81      0.81      5000
```

(a)

```
Valid Accuracy: 0.741
Test Accuracy: 0.7346
[[1990  505]
 [ 467 2038]]
Classification Report:                precision    recall  f1-score   sup
port

           0       0.74      0.72      0.73      2495
           1       0.73      0.75      0.74      2505

    accuracy                           0.73      5000
   macro avg       0.73      0.73      0.73      5000
weighted avg       0.73      0.73      0.73      5000
```

(b)

```
Valid data Accuracy is : 0.8064
Test Accuracy: 0.8012
[[1955  540]
 [ 454 2051]]
Classification Report:                precision    recall  f1-score   sup
port

           0       0.81      0.78      0.80      2495
           1       0.79      0.82      0.80      2505

    accuracy                           0.80      5000
   macro avg       0.80      0.80      0.80      5000
weighted avg       0.80      0.80      0.80      5000
```

(c)

Fig. 6 **a** Tf-idf Gaussian Naïve Bayes **b** N-gram and GNB **c** AdaBoost classifier

References

1. S.V. Wawre, S.N. Deshmukh, Sentiment classification using machine learning techniques. Int. J. Sci. Res. (IJSR) **5**(4), 819–821 (2016)
2. B. Pang, L. Lee, Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. arXiv:cs/0506075 (2005)
3. N. Godbole, M. Srinivasaiah, S. Skiena, Large-scale sentiment analysis for news and blogs. Icwsm **7**(21), 219–222 (2007)
4. A. Chatterjee, W. Perrizo, Investor classification and sentiment analysis. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE (2016), pp. 1177–1180
5. X. Fang, J. Zhan, Sentiment analysis using product review data. J. Big Data **2**(1), 1–14 (2015)

6. B. Liu, L. Zhang, A survey of opinion mining and sentiment analysis. In *Mining text data.* Springer, Boston, MA (2012), pp. 415–463
7. G.A. Miller, WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
8. A. Rashid, N. Anwer, M. Iqbal, M. Sher, A survey paper: areas, techniques and challenges of opinion mining. Int. J. Comput. Sci. Issues (IJCSI) **10**(6), 18 (2013)
9. B. Pang, L. Lee, Opinion mining and sentiment analysis. Found Trends Inf Retriev **2**(1/2), 1–135 (2008)
10. R. Feldman, Techniques and applications for sentiment analysis. Commun. ACM **56**(4), 82–89 (2013)
11. P.D. Turney, Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. arXiv:cs/0212032 (2002)
12. W.J. Ye, A.J. Lee, Mining sentiment tendencies and summaries from consumer reviews. IseB **19**(1), 107–135 (2021)
13. S.W. Chan, M.W. Chong, Sentiment analysis in financial texts. Decis. Support Syst. **94**, 53–64 (2017)
14. S. Deng, A.P. Sinha, H. Zhao, Adapting sentiment lexicons to domain-specific social media texts. Decis. Support Syst. **94**, 65–76 (2017)
15. N. Oliveira, P. Cortez, N. Areal, Stock market sentiment lexicon acquisition using microblogging data and statistical measures. Decis. Support Syst. **85**, 62–73 (2016)
16. R.V.B. Vangara, K. Thirupathur, S.P. Vangara, Opinion mining classification using naive bayes algorithm. Int. J. Innovat. Technol. Explor. Eng. (IJITEE) **9**(5), 495–498 (2020)
17. R.V.B. Vangara, S.P. Vangara, V.K. Thirupathur, A survey on natural language processing in context with machine learning. Int. J. Anal. Exp. Modal Anal, 1390–1395(2020)
18. A. Banjar, Z. Ahmed, A. Daud, R.A. Abbasi, H. Dawood, Aspect-based sentiment analysis for polarity estimation of customer reviews on Twitter. CMC-Comput. Mater. Continua **67**(2), 2203–2225 (2021)
19. M.A. Jarwar, R.A. Abbasi, M. Mushtaq, O. Maqbool, N.R. Aljohani, A. Daud, I. Chong, CommuniMents: a framework for detecting community based sentiments for events. Int. J. Semant. Web Inf. Syst. (IJSWIS), **13**(2), 87–108 (2017)
20. H.U. Khan, A. Daud, U. Ishfaq, T. Amjad, N. Aljohani, R.A. Abbasi, J.S. Alowibdi, Modelling to identify influential bloggers in the blogosphere: a survey. Comput. Hum. Behav. **68**, 64–82 (2017)
21. Kaggle Dataset. https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews (2019/5/11)
22. Y. Xu, X. Zhong, A.J. Yepes, J.H. Lau, Grey-box adversarial attack and defence for sentiment classification. arXiv:2103.11576 (2021)
23. S. Luo, H. Ivison, C. Han, J. Poon, Local interpretations for explainable natural language processing: a survey. arXiv:2103.11072 (2021)
24. R. Kaur, S. Kautish, Multimodal sentiment analysis: a survey and comparison. Int. J. Serv. Sci. Manag. Eng. Technol. (IJSSMET), **10**(2), 38–58 (2019)
25. A. Samuel, D.K. Sharma, x A spatial, temporal and sentiment based framework for indexing and clustering in twitter blogosphere. J. Intell. Fuzzy Syst. **32**(5), 3619–3632 (2019)
26. S. Rathi, S. Shekhar, D.K. Sharma, Opinion mining classification based on extension of opinion mining phrases. in *Proceedings of International Conference on ICT for Sustainable Development* (Springer, Singapore, 2016), pp. 717–724
27. D.K. Sharma, Hindi word sense disambiguation using cosine similarity. in *Proceedings of International Conference on ICT for Sustainable Development*. (Springer, Singapore, 2016), pp. 801–808
28. S. Garg, D.K Sharma, Sentiment classification of context dependent words. in *Proceedings of International Conference on ICT for Sustainable Development* (Springer, Singapore, 2016), pp. 707–715
29. R. Pradhan, D.K. Sharma, A frequency-based approach to extract aspect for aspect-based sentiment analysis. In *Proceedings of Second International Conference on Computing, Communications, and Cyber-Security* (Springer, Singapore, 2021), pp. 499–510

30. R. Pradhan, V. Khandelwal, A. Chaturvedi, D.K. Sharma, Recommendation system using lexicon based sentimental analysis with collaborative filtering. in *2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)* (IEEE, 2020), pp. 129–132
31. Y. Agarwal, R. Katarya, D.K. Sharma, Deep learning for opinion mining: a systematic survey. in *2019 4th International Conference on Information Systems and Computer Networks (ISCON)* (IEEE, 2019), pp. 782–788
32. Y. Agarwal, D.K. Sharma, R. Katarya, Sentiment/opinion review analysis: detecting spams from the good ones!. in *2019 4th International Conference on Information Systems and Computer Networks (ISCON)* (IEEE, 2019), pp. 557–563
33. S. Garg, D.K. Sharma, Feature based clustering considering context dependent words. in *2015 1st International Conference on Next Generation Computing Technologies (NGCT).* (IEEE, 2015), pp. 713–718
34. M. Rajput, R. Agrawal, D.K. Sharma, A study and comparative analysis of web personalization techniques. in *2013 IEEE Conference on Information & Communication Technologies* (IEEE, 2013), pp. 1293–1298

Content-Based Image Retrieval (CBIR): A Review

Deepti Agrawal, Apurva Agarwal, and Dilip Kumar Sharma

Abstract This paper gives an overview of various CBIR (Content-Based Image Retrieval) Techniques. CBIR is a system that utilizes different image features, like texture, color, and shape information, to fetch different images from a huge database. CBIR systems are mostly used in the medical field as they collect images from a huge database depending on the similarities. This aids in diagnosing patients. The paper describes CBIR as the first step followed by addressing the characteristics of an image (i.e., image features). Further, the techniques and applications of CBIR are discussed. The recent papers described in literature review are used to enhance the study of CBIR-based applications. This paper is useful for the researchers willing to pursue their research in content-based image retrieval.

Keywords CBIR · CBIR applications · Image features · Similarity measures

1 Introduction

Images are more useful than text since they contain more information as compared to a text. As a result, the majority of information has indeed been stored in the database in a digital form. Due to the continuous advancements in digital image processing, data storage has reached an optimal level, thereby making image search and retrieval a difficult task [1]. For image retrieval, there are two main approaches. The first one is text-based, which is done manually by a human. This method necessitates the use of text to represent an image and thus it needs a significant amount of effort and time. To get around this limitation, a new approach named as CBIR (Content-Based Image Retrieval) has been developed which describes images based on their features like texture, color, and shape [2].

The size of the database is growing with each day. This is mainly due to the technological advancements, increased computing space, and a variety of storage devices,

D. Agrawal (✉) · A. Agarwal · D. K. Sharma
Department of Computer Engineering and Applications, GLA University, Mathura, UP, India

D. K. Sharma
e-mail: dilip.sharma@gla.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_33

necessitating the invention and implementation of efficient image retrieval systems. First-ever commercial CBIR was IBM's QBIR (Query by Image Content Retrieval). Berkeley's Blobworld, Vhoto, WebSeek, VisualSeek, Photobook, and Virage [3] are a few other frameworks. Image segmentation methods are used after feature extraction in CBIR. The best results are obtained by comparing the images and finding the best matches from selected features. An efficient CBIR system can query an image through many databases and return a collection of the most relevant results.

A comprehensive study of CBIR with various image features is provided in this review paper. The frameworks for CBIR operate only with the local features of images. For CBIR images, their visual content is already saved in a database using features like texture, color, and form. CBIR is also defined as QBIC (Query By Image Content) and CBVIR (Content-Based Visual Information Retrieval). The key factor of CBIR is image retrieval and feature extraction. This paper summarizes all of the CBIR types, and the coming sections discuss a thorough study of CBIR followed by feature description and applications.

2 Literature Review

Images are often viewed more favorably than text. The retrieval process has become more complex as the number of images in the repository has increased. Content-based and text-based methods are being used for scanning and retrieving to make this complex job simpler. The aim is to deal with images to find identical images depending on multiple features. There are several methods for image retrieval. CBIR can help with fingerprint recognition, digital libraries, biodiversity data management, crime detection, military, medicine, Artificial Intelligence (AI), web image analysis and education, etc. Rani [4] pioneered a new neural network structure concept. This method is used to determine the efficacy of MLP (Multi-Layer Perception), specifically the RBF (Radial Basis Function), using CBIR. Ramya [5] proposed a hybrid solution that includes all image properties such as shape, texture, and color where Gabor filter moments and color moments are interpreted as local descriptors for texture and color features, respectively. Hence, images are retrieved by combining the abovementioned three image features to give a good set of features. Tadasare [6] defined a new algorithm for image retrieval that uses wavelet transforms to distinguish retinal fundus pictures. This algorithm is categorized into three steps: first, using the HSV transforms, preprocessing operations are carried out, then extraction method is applied using the wavelet transform, and finally, classification is done. Sharma [7] suggested a retrieval system for medical images. Due to the increased utilization of medical images, the medical image archive is rising with each day. So, here the author proposed a dual-tree complex wavelet transform algorithm that has impressive rotation invariant performance.

Wang et al. [8] reviewed CBIR techniques and using the color descriptors (CN), they depicted the color features to improve retrieval performance from huge databases. Yadaiah et al. [9] introduced the IFFS (Incremental Filtering

Feature Selection) algorithm, which uses a fuzzy rough collection to pick the best feature subset and to effectively group large amounts of data. Nagar [10] established a model for the medical field, where CBIR is used to query a huge image database using distance measures specified by the user. Rajkumar [11] introduced a Siamese neural network-based algorithm that combines shape and color features. Kaur and Singh [12] suggested an algorithm in which Gabor filtering is used to extract features, which are then optimized through lion optimization. Finally, for cuckoo search optimization, SVM is used, and for lion search optimization, the decision tree is used.

Jun et al. [13] described a mixed methodology for image retrieval as an integration of various global descriptors. This method generates a combined descriptor by concatenating multiple global descriptors in an end-to-end manner. The neural network is the main component of this process. Chu and Liu [14] focused on image retrieval's multi-integration functionality. Rather than object-oriented image retrieval, the suggested model retrieves images based on their similarity. When color and shape features are combined, it is easier to explain image content. Aiswarya et al. [15] proposed a CBIR technique for feature selection and dimensionality reduction that utilizes a multi-level stacked Autoencoder. This method is applied to mobile devices.

3 Feature Description

Images are best described with their features like texture, shape, and color. Figure 1 illustrates the classification of several CBIR methods. CBIR has many algorithms to extract images based on the different features of an image. This section describes each image feature one by one.

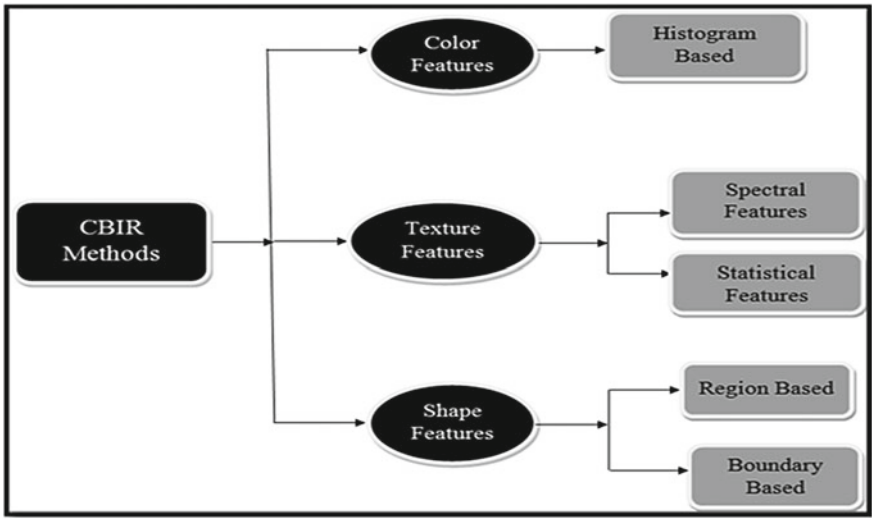


Fig. 1 Classification for CBIR

3.1 Color Feature

A low-level visual feature of a picture is its color feature. A histogram for a given image is used in image processing to explain color features [16]. In image processing, HSV (Hue, Saturation, Value), RGB (Red, Green, Blue), and YIQ are among the different types of channels used to describe an image. In image processing, the conversion of images from RGB to HSV channel is common as it is easier and more convenient for the process of feature extraction.

In CBIR techniques, color is mostly used to characterize an image's visual features. In order to represent an image, color space is used and RGB is the most commonly used color space. Following that, a primary color generated in the RGB space is represented by a histogram. Images in the database and the given query image are compared in the CBIR framework using a distance metric, which determines how similar the two histograms are. A graph that represents the frequency of occurrence of some event is a histogram. Typically, histograms have bars that show the frequency at which data occurs across the entire data set.

In Fig. 2, histogram's x-axis depicts the pixel value range. As it is an 8-bit per-pixel image, it has 256 levels of gray or shades of gray. As a result, the x-axis spectrum begins at 0 and ends at 255, with a 50-point difference. The y-axis shows the count of these intensities.

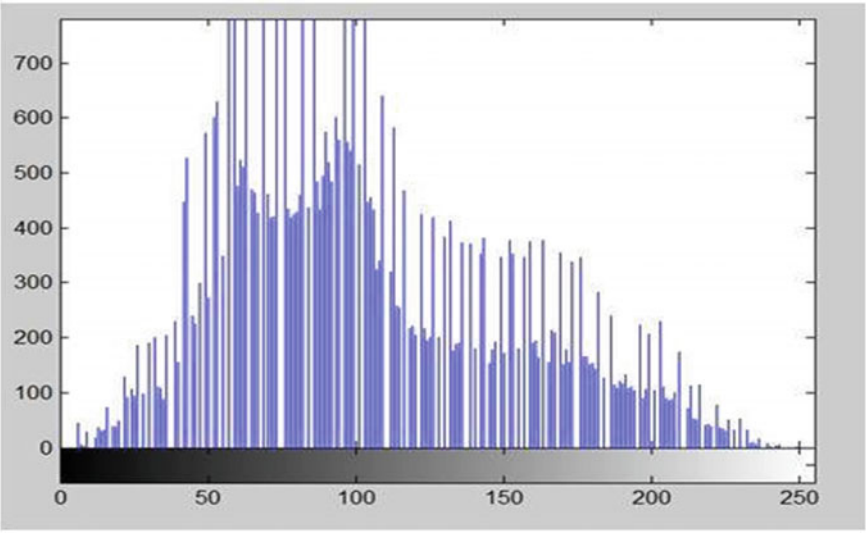


Fig. 2 Histogram for an Image

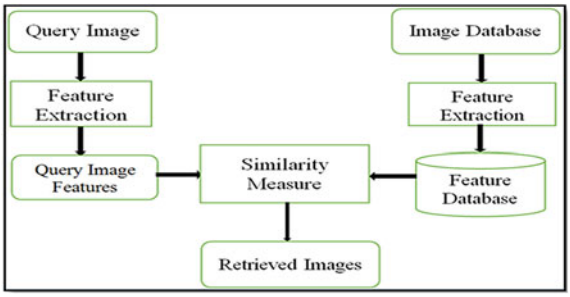


Fig. 3 Flowchart for CBIR system

3.2 Texture Feature

In the CBIR system, another feature to represent an image is the texture feature. A variety of low-level texture features are available that can be used in various image retrieval domains. Texture features are somewhat more effective unlike color features because they represent pixels group. Texture representation is further grouped into two categories: statistical methods and structural methods. For image identification, structural methods involve operators and graphs. Statistical approaches, on the other hand, use quantitative distributions of intensity of an image to identify texture. Zernike moment, Wavelet transform [17], co-occurrence matrices [18], and Fourier transform [19] are some of the texture feature-based algorithms.

3.3 Shape Feature

A low-level feature to represent an image is shape too. It also helps to identify real-world objects and shapes. Images from the database are extracted by the shape feature [20] based on their orientation or size, like contours and edges. Contour-based and region-based features are the two sorts of shape features. Polygonal approximation [21], one-dimensional function, moments, spatial interrelation feature [22], shape transform domain, and scale space methods [23] are among some of the methods for shape representation.

4 CBIR Overview

Feature extraction, feature matching, and retrieval system architecture are the three key components of the CBIR system. As shown in Fig. 3, CBIR technique consists of several steps: first, a database is generated wherein images are stored. So, given a query image as input, the system finds related images based on that query image. The next step is to extract features like texture, shape, and color from the images stored

in the database as well as from the input query image. In the final step, techniques for similarity measurement are applied to the database images and the input query image, and the system checks which features are identical among the query image given as input and database images, and then retrieves all of the closet images from the database.

5 CBIR Algorithms

There are numerous algorithms for CBIR (Content-Based Image Retrieval). The key features that characterize an image are texture, shape, and color. There are several algorithms for the extraction process. Here are a few examples of well-known algorithms:

5.1 Color Histogram

The color feature of an image is the most obvious and intuitive feature, and it is usually described using histograms [16]. Since the color histograms approach is fast, requires minimal memory space, and is unaffected by changes in image size or rotation, it attracts a lot of interest.

5.2 Zernike Moment

The rotation invariant Zernike moment is a shape feature. A complex no. moment identified by a group or set of orthogonal polynomials is known as Zernike moment. It is used in the unit circle as a whole orthogonal set. It has a lower level of information redundancy and higher rotation invariance. When compared to other shape features, it provides better outcomes. It is more effective for illustration process and more resistant to noise [24, 25].

The first step in calculating the Zernike moment is to define the regions of an image that are of interest, and then map those regions to the unit circle's origin. The pixels that are internal to the unit circle are included in the final stage of estimation, while the rest are ignored.

5.3 Curvelet Transform

Candes et al. [26] proposed a novel multiscale directional transform that allows for a non-adaptive sparse representation of objects with C^2 singularities that is optimal.

But, the above transform method lacks the object's geometric properties and thus the wavelet transform represents point singularities. Here, the curvelet transform solves these issues. Curvelet transforms have two characteristics: first, they are geometrical transforms, and second is their directionality.

5.4 Scale-Invariant Fourier Transform (SIFT)

CBIR can sometimes fail to capture certain local features that reflect detailed information and the degree of difference between scenes. As a result, SIFT (Scale Invariant Feature Transform) is used, as it is useful when used to retrieve images [27, 28]. Some previous image retrieval techniques may have failed to capture some local features reflecting scene information and nuances.

The main purpose of SIFT is to retrieve medical images. Web images can be retrieved using Content-based image retrieval with SIFT [29]. It can be used by web image search engines effectively.

6 Different Categories of CBIR System

The CBIR framework can handle a wide range of user experiences and implementation methods. CBIR is mainly divided into 9 categories. They are as follows.

6.1 Single Query Based CBIR

CBIR systems were focused on single query retrieval at just the start of image retrieval. The query is quoted by the user as text, which is then compared against the annotations of the database to generate results [30, 31].

6.2 Visual and Textual-Based CBIR

CBIR methods based on a single query are slower than methods based on annotations. The paper [32] describes a system for retrieving images that incorporates both textual and visual features. Images are segmented into clusters in this system, and each cluster is linked to a collection of terms derived from textual descriptors.

6.3 Sketch-Based CBIR

Since the environment has shifted from desktop to mobile devices in today's world, outline drawing is easier than text. The sketch-based CBIR system [33] is much relevant on querying large image databases.

6.4 Shape-Based CBIR

For CBIR system, image retrieval using the shape feature is a more relevant form. He et al. [34] proposed a method for CBIR based on shape feature using edge points based on contours in a binary image environment. Their method also deals with the rotation invariant feature which is a relevant property of shape feature of an image.

6.5 Region-Based CBIR

This is another category of CBIR system in which the user selects the Region of Interest (ROI) from the given query image and retrieval process is based on the query image's visual features. These types of CBIR applications boost image retrieval performance. A region-based retrieval method was proposed by Vinima and Poulas [35], which enables the user to define the query image's specific region. Texture and color features are used in the image retrieval process.

6.6 Barcode Annotation and Visual-Based CBIR

The results of image retrieval are faster when using an annotation-based system. The method for annotations based on barcode is presented by Tizhoosh [36], and the creation of small barcodes in order to embed them in medical DICOM files is also explained in the paper.

6.7 Visual Phrases-Based CBIR

The authors in [37] describe a procedure for scene text extraction from a given query image as well as a method for indexing relevant images. SURF is used for detection, and string matching is done.

6.8 *Multiple Query-Based CBIR*

Previously for image retrieval, methods based on a single query were used. In order to overcome the limitations of image retrieval based on single query, multiple query-based methods were suggested. CBIR technique based on multiple queries is discussed by Hsiao et al. [38], wherein corresponding to different image semantics, two query images are processed as query. These techniques have shown significantly improved recall and precision rates.

6.9 *Relevance Feedback-Based CBIR*

Precision and response time of CBIR systems are both problematic. To optimize information retrieval performance, relevance feedback can prove to be useful. The authors in [39] focused on using a Bayesian model feature subspace and progressive learning to create an image retrieval system based on relevance feedback.

7 CBIR Applications

Examples of CBIR applications [40] are:

- Security Check: Access privileges are determined by screening fingerprints or retinas. Security checks are largely focused on the individual's or company's criminal history, commercial records, and financial records.
- Crime Prevention: Any action by a person or organization, public or private, aimed at preventing or deterring crime until it occurs or leads to additional activity is referred to as crime prevention. Police departments use automatic facial recognition devices.
- Intellectual Property: Registration of trademark images, which compares a potential nominee mark to currently existing marks to ensure that there, is no chance of property ownership confusion.
- Medical Diagnosis: CBIR technology has been considered to help with health care, biomedical research, and education, in addition to managing exceptionally huge image collections [41]. Based on a study of the literature, we conclude that there is widespread appreciation of CBIR in the engineering community for academic researchers, but its application to realistic medical problems is still a long way off. Finding related past cases through CBIR using medical images in a medical database so as to aid diagnosis.
- Military (aerial, radar): CBIR is used in the military to recognize enemy aircrafts on radar screens, identify targets from satellite images, and so on. Most of the other methods used in crime prevention could be applicable in the military.

- Cultural (museums, art galleries): Research in digital cultural heritage is one of the many applications of CBIR that has been rising in importance and influence. Its growth has been driven by the specific needs among libraries and museums around the world for more historical cataloguing and extraction mechanisms on image data collections than that are currently available via keyword searches.
- Commerce (catalogue, fashion): Because of the accessibility of the internet, e-purchases are the latest trend across the world rather than conventional store purchases. When it comes to e-purchasing, people have a hard time deciding what they want to buy. The CBIR systems can help users make better clothing choices during e-purchases by assisting them with their clothing choices.
- CBIR research using Deep Learning: Machine learning is widely considered as a potential direction to overcome the semantic gap in the long run. Much progress in CBIR tasks can be achieved by exploring the state-of-the-art deep learning techniques for learning image feature representations and similarity measurement, driven by recent achievements of deep learning methods for computer vision and other applications [41].
- Face recognition: Face recognition is considered to be a Content-Based Image Retrieval (CBIR) problem [42]. For a given face, its arbitrary input image is interpreted as a sample for searching within a database, which contains a huge collection of images (i.e., depictions from a sufficient number of viewpoints) for each interested human face.
- Pattern Identification: In Pattern Recognition, experimentation is done to collect "relevant" data for an object and then these features (measurements) are used to classify it [43]. While they are similar in concept and use much of the same image evaluation methods, direct object recognition and CBIR are somewhat different techniques. Object recognition is essentially a statistical matching problem that uses an established object database. Object recognition, one might say, is a fairly well-defined subset of CBIR.
- Entertainment (personal album)
- Engineering and Architectural design.

8 CBIR Similarity Measures

For the retrieval method, distances are used in CBIR systems, which measure the distance between the database images and the query image and retrieve images similar to the given query image. In CBIR, the following distance measures are used:

8.1 Euclidean Distance

This distance is often used to determine the degree of similarity among the queried image and the most similar database images. Euclidean Distance is calculated as follows [44]:

$$D(T, J) = \left(\sum_{j=1}^{n} (J_j - T_j)^2 \right)^{\frac{1}{2}} \quad (1)$$

where $D(T, J)$ gives the measure for Euclidean distance between the query image's feature vector J and the feature vector of each database image T. J_j is the feature vector of the input query image and T_j is the feature vector of database images.

8.2 Canberra Distance

For similarity comparison, Canberra distance is also used. This measure is determined with the help of the following equation [45]:

$$CanbDist(a, b) = \sum_{i=1}^{d} \frac{|a_i - b_i|}{|a_{i|} + |b_{i|}} \quad (2)$$

where a is the feature vector of the given query image and b is the database image's feature vector, with dimension d.

9 Discussion

This paper has successfully discussed different feature descriptors, similarity measures, and 9 separations for the types of CBIR systems along with its applications and reviews of the associated latest papers in order to completely understand this image retrieval technique. Meaningful discussions on CBIR have been expressed through Fig. 4. It gives an overview of CBIR discussed in the paper and their best-related papers.

This paper is useful for the researchers willing to pursue their research in content-based image retrieval as it successfully explains the concerned technique including its different methods, categories, and applications all in one place. The paper also cites their respective best-related papers to enable the researchers to study these methods, categories, and applications in detail without much struggle.

10 Conclusion

Due to the beneficial searches with different images, CBIR has become an appealing area for everyday human life. The key components of CBIR are examined in this paper. With the size and number of image databases progressively increasing, we

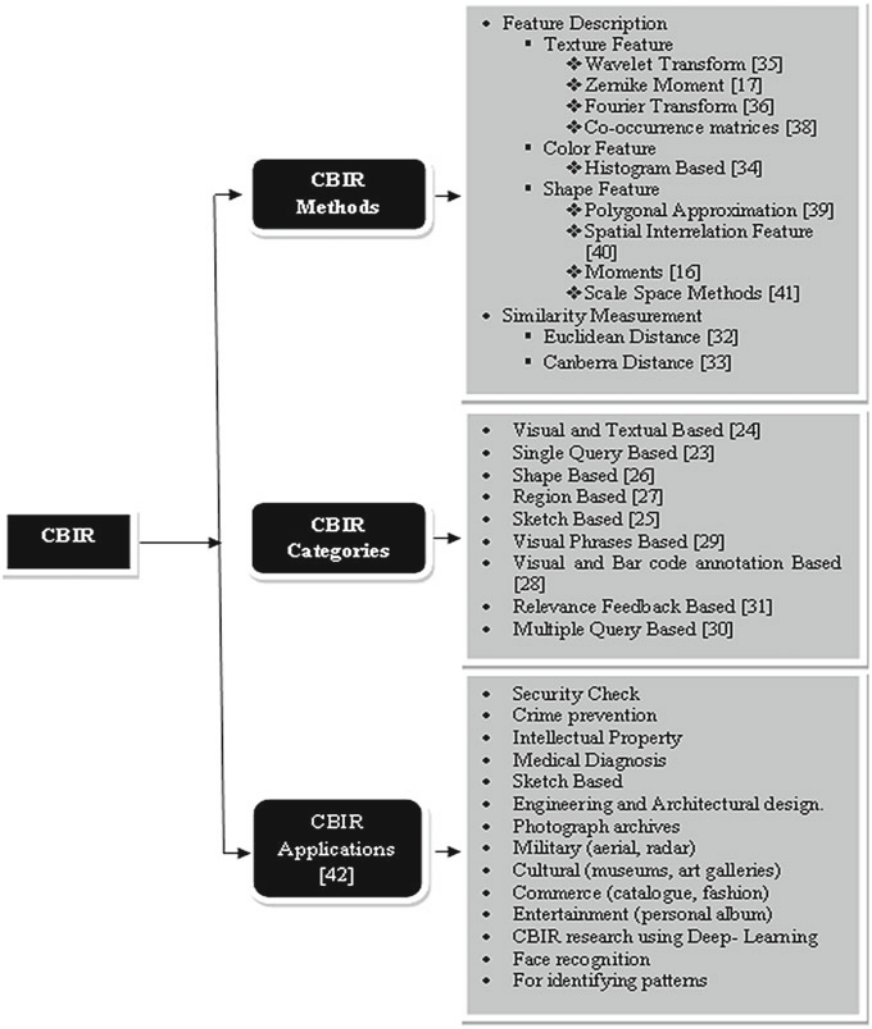


Fig. 4 An overview of CBIR methods, categories, and applications discussed in the paper

need a CBIR system that is accurate. The rotation invariant property of CBIR allows image retrieval easier. It is concerned with feature extraction and representation, which has a close connection to the perception of a human being.

The main components of a CBIR system have been reviewed in this paper, which include feature representation of an image, indexing, query processing, and similarity measurement as well as the current state of the art and key challenges and applications. This paper also summarizes nine CBIR categories along with their corresponding associated papers that can assist researchers in their research. In the CBIR system, there are some issues, one of which is rotation invariance. Some techniques alleviate

this issue to some extent, but not entirely. So, in our future research, we'll be looking for a technique that can overcome this problem effectively to obtain more accurate image retrieval results.

References

1. P.A. Dahake, S.S. Thakare, Content based image retrieval: a review. Int. Res. J. Eng. Technol. (IRJET) **05**(01) (2018)
2. S. Rubini, R. Divya, G. Divyalakshmi, T.M.S. Ganesan, Content based image retrieval (CBIR). Int. Res. J. Eng. Technol. (IRJET) **05**(03) (2018)
3. P. Kaur, R.K. Singh, Content based image retrieval using machine learning and soft computing techniques. Int. J. Sci. Technol. Res. **9**(1) (2020)
4. R.U. Rani, Image qualification learning technique through content based image retrieval. Int. J. Future Revolut. Comput. Sci. Commun. Eng. **4**(1) (2020)
5. V. Ramya, Content based image retrieval system using clustering with combined patterns. Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol. **3**(1), 1060–1063 (2018)
6. S.S. Tadasare, S.S. Pawar, Content based retinal image retrieval using lifting wavelet transform for classification of retinal fundus images. Int. J. Elect. Electron. Comput. Sci. Eng. **5**(1), 169–176 (2018)
7. D. Sharma, T. Gulati, R. Sharma, Rotation invariant content based image retrieval system for medical images. Int. J. Future Revolut. Comput. Sci. Commun. Eng. **4**(3), 171–174 (2018)
8. X.Y. Wang, H.Y. Yang, D.M. Li, A new content based image retrieval technique using color and texture information. Comput. Electr. Eng. **39**(3), 746–761 (2013)
9. V. Yadaiah, R. Vivekanandam, R. Jatothu, A Fuzzy logic based soft computing approach in CBIR system using incremental filtering feature selection to identify patterns. Int. J. Appl. Eng. Res. **13**, 2432–2442 (2018)
10. A. Rao Nagar, N.S. Sushmitha, M.K. Nalini, Content based image retrieval for medical domain. Int. J. Eng. Res. Technol. **8**(06), 1345–1351 (2019)
11. R. Rajkumar, M.V. Sudhamani, Content based image retrieval system using combination of color and shape features, and siamese neural network. Int. J. Innovat. Technol. Explor. Eng. **9**, 71–77 (2019)
12. P. Kaur, D. Kumar Singh, Content based image retrieval using machine learning and soft computing techniques. Int. J. Sci. Technol. Res. **9**(1), 1327–1332 (2020)
13. H. Jun, B. Ko, Y. Kim, L. Kim, J. Kim, Combination of multiple global descriptors for image retrieval (2020)
14. K. Chu, G. Hai Liu, Image retrieval based on a multi- integration features model (2020)
15. K.S. Aiswarya, N. Santhi, K. Ramar, Content based image retrieval for mobile devices using multi- stage autoencoders. J. Crit. Rev. (2020)
16. R. Brunelli, O. Mich, Histogram analysis for image retrieval. Pattern Recogn. **34**(8), 1625–1637 (2001)
17. M. Antonini, M. Barlaud, P. Mathieu, I. Daubechies, Image coding using wavelet transform. IEEE Trans. Image Process. **1**(2), 205–221 (1992)
18. C.C. Gotlieb, H.E. Kreyszig, Texture descriptors based on co-occurrence matrices. Comput. Vision Gr. Image Process. **51**(1), 70–86 (1990)
19. R.N. Bracewell, The fourier transform. Sci. Am. Div. Nat. Am. Inc. **260**(6), 86–95 (1989)
20. B. M. Mehtre, M.S. Kankanhalli, W. Foon Lee, Shape measures for content based image retrieval: a comparison. Inf. Process. Manag. **33**(3), 319–337 (1997)
21. S. Han, S. Yang, S, An invariant feature representation for shape retrieval. in *Proceedings of Sixth International Conference on Parallel and Distributed Computing, Applications and Technologies* (2005)

22. K. Orth, Estimation methods for models of spatial interaction. J. Am. Stat. Assoc **70**(349), 120–126 (1975)
23. A. Witkin, Scale-space filtering: a new approach to multi-scale description. in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 9 (1984). pp. 19–21
24. M.R. Teague, Image analysis via the general theory of moments. Opt. Soc. Am. J. **70**, 920–930 (1980)
25. W.Y. Kim, Y.S Kim, A region based shape descriptor using Zernike moments. Signal Process. Image Commun. **16**(1), 95–102 (2000)
26. E. Candes, L. Demanet, D. Donoho, L. ying, Fast discrete curvelet transforms. Soc. Indust. Appl. Math. **5**(3), 861–899 (2006)
27. J. Almeida, R.S. Torres, S. Goldenstein, SIFT applied to CBIR. Revista de Sistemas de Informacao da FSMA n. **4**, 41–48 (2009)
28. M. Kamath, D. Punjabi, T. Sabnis, D. Upadhyay, S. Shrawne, Improving content based image retrieval using scale invariant feature transform. Int. J. Eng. Adv. Technol. (IJEAT) **1**(5), 2249–8958 (2012)
29. Z. Wang, K. Jia, P. Liu, An effective web content-based image retrieval algorithm by using SIFT feature. In: *World Congress on Software Engineering, IEEE Computer Society* (2009).
30. R.A. Konstantinos, K.S. Ntalianis, D.D. Sourlas, S.D. Kollias, Mining user queries with Markov Chains: application to online image retrieval. IEEE Trans. Knowl. Data Eng. **25**(2) (2013)
31. S. Kulkarni, P. Kulkarni, Color image annotation using hybrid intelligent techniques for image retrieval. IEEE Int. Conf. (2012). 978-1-4673-5116-4/12
32. F. Debole, C. Gennaro, P. Savino, Enriching image feature description supporting effective content based retrieval and annotation. IEEE Int. Conf. (2014)
33. M. Eitz, M. Hildebrand, T. Boubekeur, M. Alexa, Sketch based image retrieval: benchmark and bag-of-features descriptors. IEEE Trans. Visualization Comput. Gr. **17**(11) (2011)
34. Y. He, L. Yang, Y. Zhang, X. Wu, Y. Chang, The binary image retrieval based on the improved shape context. in *7th International Conference on Image and Signal Processing* (2014)
35. E.R. Vinima, K.J. Poulas, Image retrieval using color and texture features of regions of interest. in *IEEE Conference* (2012)
36. H.R. Tizhoosh, Barcode annotations for medical image retrieval: a preliminary investigation. in *IEEE International Conference*. 978–1–4799-8339–1/15 (2015)
37. G. Schroth, S. Hilsenbeck, R. Huitl, F. Schweiger, E. Steinbach, Exploiting text-related features for content based image retrieval. in *IEEE International Symposium on Multimedia* (2011)
38. K.J. Hsiao, J. Calder, A.O. Hero, Pareto-depth for multiple query image retrieval. IEEE Trans. Image Process. **24**(2) (2015)
39. Z. Su, H. Zhang, S. Li, S. MA, Relevance feedback in content based image retrieval: bayesian framework, feature subspaces, and progressive learning. IEEE Trans. Image Process. **12**(8) (2003)
40. A. Khokher, R. Talwar, Content-based image retrieval: feature extraction techniques and applications. in *International Conference on Recent Advances and Future Trends in Information Technology* (2012) pp. 9–14
41. R.R Saritha, V. Paul, P.G. Kumar, Content based image retrieval using deep learning process. Clust. Comput. 1–14 (2018)
42. R. Gross, J. Shi, J. Cohn, *Quo Vadis Face Recognition: Third Workshop on Empirical Evaluation methods in Computer Vision, Pittsburg* (Carnegie Mellon University, USA, 2001)
43. D. Zhang, M.M. Islam, G. Lu, A review on automatic image annotation techniques. Pattern Recognit. **45**(1), 346–362 (2012)
44. L. Wang, Y. Zhang, J. Feng, On the Euclidean distance of images. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1334–1339
45. G. Jurman, S. Riccadonna, R. Visintainer, C. Furlanello, Canberra distance on ranked lists. in *Proceedings of Advances in Ranking NIPS 09 Workshop* (2009), pp. 22–27

Automatic Speech Emotion Recognition Using Cochleagram Features

Saumya Borwankar, Dhruv Shah, Jai Prakash Verma, and Sudeep Tanwar

Abstract Speech Emotion Recognition (SER) has become a popular field in recent years. The efficiency of the SER system is determined by how much useful information is present in the extracted features. Ongoing research has come close to achieve state-of-the-art results using neural networks like Convolutional Neural Networks (CNN) with the extracted features. Speech emotion detection deals with the recognition of the speaker's emotion from their speech sample. This detection helps us in recognizing the psychological and physical state. In this work, we have worked with two publicly available corpora—Surrey Audio-Visual Expressed Emotion (SAVEE) and Toronto Emotional Speech Set (TESS). In this paper, we approach this problem with the help of cochleagrams, we first extract cochleagram features from all the audio files and then classify different classes with the help of convolutional neural network architecture. Our proposed approach achieves an accuracy of about 97%. Cross-dataset evaluation is carried out and presented as well.

Keywords Speech emotion recognition · Cochleagram features · Deep learning · Machine learning · TESS · SAVEE

S. Borwankar · D. Shah · J. P. Verma (✉) · S. Tanwar
Nirma University, Ahmedabad, Gujarat, India
e-mail: jaiprakash.verma@nirmauni.ac.in

S. Borwankar
e-mail: 17bec095@nirmauni.ac.in

D. Shah
e-mail: 17bec097@nirmauni.ac.in

S. Tanwar
e-mail: sudeep.tanwar@nirmauni.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_34

1 Introduction

Deep learning applications have increased rapidly in solving classification problems. The easiest and common mode of communication is speech among humans. During this communication humans recognize emotions as well as a change in emotion based on speech which is an important area for Human–Machine Interaction (HMI). Teaching a machine to recognize emotions can help in many ways like in [2, 11] the author talks about using speech emotion detection in robotics, calling centers, automobiles, etc.

This task of emotion detection in speech has been the point of interest for a long time. This problem comes with its challenges like, for example, an audio sample can contain more than one emotion and different forms of audio can have different impacts as to what emotion the speaker is feeling. The feature selection is also an important aspect of this problem.

The previous approaches deal with the problem with the help of Support Vector Machines (SVM), Hidden Markov Models (HMMs) or neural networks. SVM are providing with good estimates with less effort while neural networks provide better results but require more computational power and more time to train. Elicited emotion speech corpus provides us with artificial emotion which may not be the correct and can be collected from actors, so the natural speech corpus has proved to be very suitable, which the authors talk about in [3]. So the research in this field requires more accurate and noise-free corpus to improve the results of the SER system.

In [6], the features of the speech sample are classified into prosodic features, spectrum features, and non-linear features. The search for best features has also attracted attention in the field of speech recognition. Features like Linear Prediction Cepstral Coefficients (LPCC), Mel-Frequency Cepstral Coefficients (MFCC), Log Frequency Power Coefficients (LFPC), etc. exhibit valuable and usable information. Mel spectrograms are gaining popularity in the field of speech recognition as they are a typical input to CNN architectures [15].

New methods need to be explored in SER to find the optimal and robust set of features that can determine the emotion of the speech sample. MFCC has proved to be the main feature in this problem and is used to build deep learning models [17, 25]. Recurrent neural networks have also been an attraction to the problem at hand, in [7] the authors proposed a RNN to extract mel features to create a layer for SER. In [23], Mirsamadi constructed Bidirectional Long Short-Term Memory Network (BLSTM) model on speech signal to make emotions more salient.

The key contributions of the paper are:

1. Novel feature extraction technique is looked at and discussed.
2. Discussion about the performance of the new feature.

The steps in our proposed approach are described briefly below:

1. Data Augmentation: The audio in the dataset needs to be augmented before further computation can be carried out on the audio sample. So the intensity of

the audio files is changed to augment the data. The audio samples are taken at 0dB, 5dB, 10dB, 20dB, 35dB, 100dB.
2. Cochleagram extraction: The cochleagram features are extracted from the audio files using Brian library.[1]
3. Dividing data: The audio files are then split into different classes (emotions), after which training and testing folders are created based on split size.
4. Classifier: Then the cochleagram features are passed to a classifier which classify them and the output layers classify the emotion.

The main aim of this paper is to improve the results of the previous approaches by selecting a new feature from the audio sample and classifying it with the help of convolutional neural networks. Section 2 explains the work relevant to SER systems. Different deep learning methods were looked upon in the previous research works. In Sect. 3, an explanation of the proposed feature is given. Section 4 gives a description of the dataset. Section 5 explains the implementation scenario. Finally, the experimental result achieved in our proposed approach and the conclusion that can be derived from the same has been explained in Sect. 5 and Sect. 6, respectively.

2 Relevant Work

Emotion detection has become one of the emerging fields in recent times and many researchers have worked to solve the problem with recognizing emotion in speech sample. Guo et al. [13] built a system model that used pre-processed speech signal, one feature fusion method with the help of heuristic discriminative features and spectrograms. Their results showed that the heuristic features can be used to make a contribution to SER system using a combination in CNN-ELM model. Their model reached an average accuracy of 93.3% and an F1-score of 92.50%.

Many experiments [5, 19, 26, 28] have used SVM classifiers to solve the issue of speech emotion detection. Since almost all of their appearances are nearly identical, only the first would be quickly mentioned. Three approaches to extending the simple SVM binary classification to the multi-class case are explored in this analysis. An SVM classifier is used to model each emotion in the first two methods, and it is conditioned against all other emotions. In the first solution, the class with the greatest distance from all classes is chosen. The SVM output distances are fed to a three-layer MLP classifier in the second approach, which generates the output. A hierarchical classification system was used in the third method. The three structures were put to the test using FERMUS III corpus utterances [27]. The classification accuracies for the first, second, and third methods, respectively, are 76.12%, 75.45%, and 81.29% for speaker-independent classification. The classification accuracies for the first, second, and third methods, respectively, are 92.95%, 88.7%, and 90.95% for speaker-dependent classification.

[1] https://github.com/brian-team/brian2/tree/a9766792f9a0c3f678bf481fc0958c735cf99afc.

Following the progress of LSTM in natural language processing [1, 20, 22], it has now been applied to speech emotion detection. Wöllmer [29] used LSTM for the first time to recognize continuous emotions, extracting 4843 features for each utterance as the LSTM's feedback. The static features were used as the feedback of a bidirectional LSTM (BLSTM) to predict the emotional expression of a spoken utterance in his subsequent work [15]. Since global statistics ignore the temporal structure of speech, the temporal information in speech is not completely used in terms of features [24]. The time window is fed frame by frame into a recurrent layer in order to improve the features.

Liu et al. [21] applied a technique called emotional feature fusion which combines prosodic features and spectral features. The authors used Deep Neural Networks (DNN) and CNN to build their model. They have worked on the Chinese emotion dataset (CASIA) which included six emotions: fear, anger, happy, sad, surprise, and neutral. The results improved as they used both prosodic and spectral features and this method has never been used before.

Wunarso et al. [30] have proposed an approach wherein the authors extracted the speech duration, approximate coefficients and amplitude from the Indonesian speech database (I-SpeED) , the dataset is in their native language and they used support vector machine as a classification method. After their evaluation and analysis, they were able to get a mean accuracy of 76.84% .

Zamil et al. [31] have proposed MFCC highlight investigation on the discourse sign to recognize the fundamental feelings. The classifier utilized was Logistic Model Tree (LMT) to group between various classes of feeling. The creator utilized a democratic calculation on the casings that were arranged to identify the feelings. RAVDESS and Emo-DB were the two datasets that were utilized to assess their calculation. The presentation of the best feeling was around 70%.

Han et al. [14] generated a probability distribution for various emotions given per section using a deep neural network. They've also used a Shallow single neural network to classify emotions from utterance attributes, as well as their single hidden layer network, ELM (extreme learning machine), which can produce excellent classification results even with a limited training collection. To be able to identify emotions, they merged their DNN section level distributions to establish utterance level functions, which they then gave to ELM to describe the emotions. They used a standard HMM-based recognizer and a standard SVM-based framework to equate their suggested approach to others.

For speech emotion detection, the experiment used the IITKGP-SEC and IITKGP-SEHSC speech corpus databases. GMM-HMM, AANN-GMM, and HMM-AANN were used to suggest different models for feature extraction and deployment. The AANN performed well when paired with GMM and HMM, according to the findings. GMM-AANN had a 44.52% accuracy, while HMM-AANN had a 39% accuracy [4]. The proposed framework recognizes emotions from speech using an English database of five different emotional states for a human. MFCC's conventional approach to feature extraction was used. The classification is based on the frequency spectrum by using KNN. Overall, the success rate was 63.63%. Males had a success rate of 72.72%, while females had a success rate of 54.54% [8]. A dictionary of acted and

Table 1 A comparative analysis for better understanding of approaches and the classification done by researchers

References	Objective	Approach	Dataset	Finding	Accuracy	Limitations
Guo et al. [13]	Classify four emotions	3-D attention-based convolutional recurrent neural networks (ACRNN) for SER	IEMOCAP and EMO-DB	The use of log-Mels (static, deltas, and delta-deltas) with ACRNN helped classify emotions properly	82.82% IEMOCAP and 64.74% EMO-DB	They were only able to classify four emotions
Fayek et al. [10]	Classify all emotions in both datasets	Deep Neural Network (DNN)	eINTERFACE and SAVEE	Use of DNN for the task of SER	60.53% on eNTERFACE 59.7% on SAVEE	The accuracy was low
Liu et al. [21]	Feature fusion technique using deep learning	Combining hyper-prosodic features and spectrogram features for SER	CASIA	Speech emotion can be reflected in frequency domain as well	99% precision	Other features for fusing time and frequency domain were not looked at
Han et al. [14]	They proposed a DNN to estimate emotion states for each speech segment	Using a ELM to classify emotions	IEMOCAP	This approach substantially boosts the performance of emotion recognition	60%	The use of neural networks can be looked at
Zamil et al. [31]	MFCC highlight investigation	The classifier utilized was Logistic Model Tree (LMT) to group between various classes of feeling	RAVDESS and Emo-DB	The use of MFCC increased the accuracy	70%	More experiments can be set up with other features

actual emotions was used to recognize expression. Feature engineering for emotion GMM and MFCC. MFCC was used to achieve a total of 21 features, and GMM was used to assemble 16 components with a 95% accuracy [18].

Fayek et al. [10] explain the use of deep neural networks on the SER method, claiming that they can achieve an accuracy of 59.7% on the SAVEE dataset and 60.53% on the interface dataset using deep neural networks. In citecomp1, Eric et al. [12] developed a new multi-time scale convolution for the SER method and tested it using separate subsets of four datasets. On the EMO-DB dataset, they were able to achieve the best accuracy score of about 70.97% (Table 1).

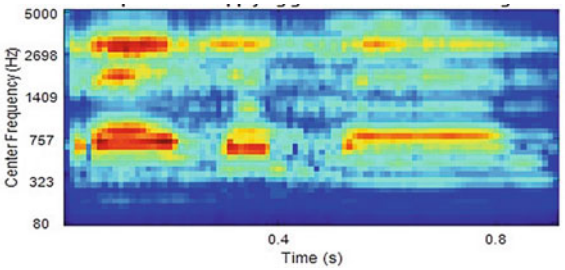


Fig. 1 Cochleagram representation

3 Cochleagram

Gammatone filter banks, which are a series of bandpass filters with the impulse response [4], are needed for the formulation of the cochleagram.

$$g(r) = Ar^{j-1}e^{-2\pi Br}\cos(2\pi f_c r + \phi), \tag{1}$$

where A is the amplitude, j is the order of the filter, B is the duration of the impulse response or filter bandwidth, Fc is the center frequency of the filter, theta is the phase, and r is the time. Thousands of hair cells make up the human cochlea, each of which resonates at a specific frequency and bandwidth. The mapping between center frequency and cochlea location in [5] is calculated by combining the reciprocal of (5) with a phase factor parameter to show filter overlap. The mapping between filter index and center frequency is then done. The cochleagram is a type of spectrogram. It makes use of a gammatone filter and has been shown to expose more spectral data (Fig. 1).

4 Dataset Selection

We used the TESS dataset to analyze our proposed solution; the dataset was selected so that we could compare our approach to other approaches. As a result, we tested our model by splitting the dataset into three parts training, testing, and validation. Table 2 shows the results of the comparison. Only five emotion labels were selected to remove the dataset imbalance problem, namely, sad, angry, happy, neutral, and fear.

4.1 TESS

Toronto Emotional Speech Set [9] is a publicly available dataset which consists of 2 English speakers having 280 utterances which comes to a total of 1 h 36 minutes.

Table 2 Dataset details

References	Type	Size (# files)	Anger	Bored	Disgust	Fear	Happy	Sad	Neutral	Surprise	Advantages	Limitation
TESS [9]	Audio	2800	✓		✓	✓	✓	✓	✓	✓	Dataset is female only and is of very high quality audio	No male voice

There are 7 emotion labels in the dataset, namely, happy, sad, angry, disgusted, neutral, surprise, and fear. The audio data is augmented and the audio samples are taken at various intensities like 0dB, 5dB, 10dB, 15dB, 20dB, 25dB, 30dB, and 100dB are taken after which the cochleagrams of each are calculated. After augmentation, every emotion label has around 3600 audio files. TESS dataset has been subjected to various researches including Guizzo et al. [12], which is why it was selected as part of our proposed research work.

4.2 SAVEE

Surrey Audio-Visual Expressed Emotion (SAVEE) [16] dataset has four English-speaking males at the University of Surrey aged between 27 and 31 years. There are six emotion labels in the dataset: surprise, sadness, happiness, anxiety, disgust, and rage. There are 480 annotated utterances in total. The data is open to the public. Each emotion mark contains 60 audio files. The SAVEE dataset has been used by a number of scholars, including Fayek et al. [10], which is why it was chosen.

5 Implementation and Execution

The proposed work was carefully pre-processed and provided to the CNN architecture to perform classification. The detailed results are shown in the next section. Taking into account the class imbalance only five emotion labels were taken for evaluation. The emotions were considered taking into account previous approaches and the class imbalance, so only five emotion labels were considered for evaluation. The final feature set was divided into training, testing, and validation data. The split size was taken as 70%, 20%, and 10%, respectively.

5.1 Algorithm

The implementation of the algorithm is carried out on a machine having 8Gb DDR4 RAM, a 4Gb Nvidia GTX1050 Graphics Processing Unit (GPU), and an Intel Core i7-7700HQ Central Processing Unit (CPU) 2.80 GHz which is a 64-bit processor. The entire algorithm can be visualized as below as mentioned in Algorithm 1.

Algorithm 1 Model creation

Require: Read file (.wav)
procedure READING(audio files) ▷ Saving the audio file as a variable
end procedure
procedure MATRIX(audio files) ▷ Calculates Cochleagram from audio files
  while audio file in directory **do** ▷ Loops over every file in directory
    $a \leftarrow MFCC, Melspectrogram, ChromaCENS$ ▷ Extract Cochleagram feature
    $b \leftarrow Name$ ▷ Get the label of the file
  end while
  $a, b \leftarrow a.reshape, b.reshape$ ▷ Converting into proper format
  return a,b
end procedure
procedure MAKE MODEL ▷ Creation of Resnet model
  while layers **do**
    $model \leftarrow model.layers$ ▷ Defining model architecture
  end while
  $model \leftarrow model.compile$ ▷ Compiling the model
  return model
end procedure
procedure DIVIDE DATA ▷ Train, test and validation data
  $[training + testing], validationfeature \leftarrow [a, b], c$ ▷ Splitting data 7:2:1
  $x_train, x_test, y_train, y_test \leftarrow a, b$ ▷ Diving the data using scikit-learn
  return x_train,x_test,y_train,y_test
end procedure
for <number of epochs> **do**
  <Pass train and test data into model>
end for
save $model.h5$ ▷ This model file can be used to make predictions

5.2 Execution

The Resnet architecture is implemented in steps, with each stage containing a series of predefined layers and each residual module containing a certain number of residual modules. As a result, the Resnet in our proposed solution has three phases, each with three, four, or six residual blocks. The first stage has three sets of independent residual units, each with three Conv layers. With dimensionality reduction added, this stage will learn 32, 64, and 128 filers in that order. Four sets of residual modules are included in the next stage. Each residual module has three Conv layers, each with 64, 64, and 256 filters in that order, with dimensionality reduction added. The final stage has six of these residual modules, each with three Conv layers that learn 128, 128, and 512 filers. Batch normalization is also available in each residual module. Following that, an average pooling layer is added, followed by a Softmax feature layer. With a momentum of 0.9 and a learning rate of 0.1, a stochastic gradient descent optimizer was used. Since there are five emotion groups to classify, the loss function is set to categorical cross-entropy for model compilation, and the device parameters

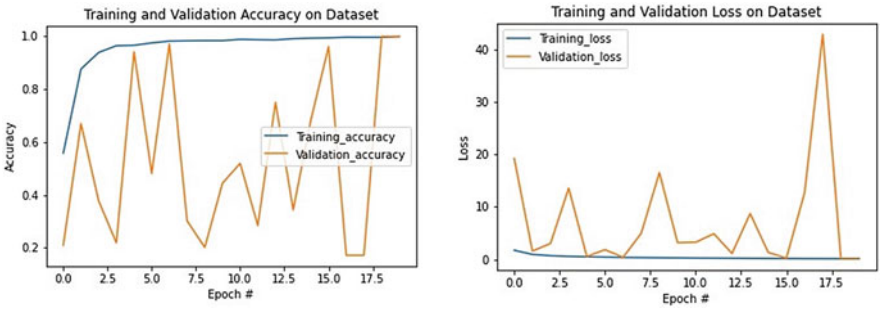


Fig. 2 Training and validation **a** Accuracy and **b** Loss on TESS dataset

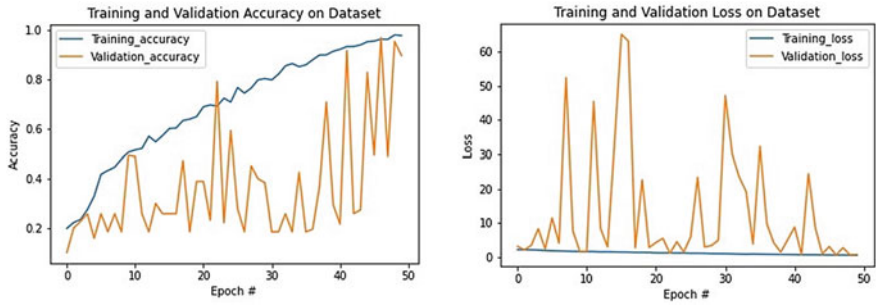


Fig. 3 Training and validation **a** Accuracy and **b** Loss on SAVEE dataset

are modified after each epoch. The model was trained for 35 epochs and a batch size of 64 was chosen. For the SAVEE model training, the model was run for 50 epochs.

6 Results

So as the aim of the paper is to evaluate the new method for feature selection the classification was done on TESS dataset, SAVEE dataset and fed to the resnet for the classification. The results of the approach are shown in Table 3. The evaluation parameters such as accuracy, recall, and F1 are taken into consideration. Figure 4 shows us the training and validation accuracy and loss for all datasets combined. Figure 5 shows us the training and validation accuracy and loss for the TESS dataset (Figs. 2 and 3).

Table 3 shows that when the model was given more training results, it was able to do better. As compared to previous methods, the model achieved a 97% accuracy rate, which is considered good. Table 3 also shows how various techniques work on

Table 3 Comparison of different methods

Method	Dataset	Classes	Accuracy (%)	F1	Recall
[12] Multi-time scale convolution	TESS	5	53.05	–	–
[10] Deep neural networks	SAVEE	7	59.7	–	–
Our approach	TESS	5	97	0.98	0.97
Our approach	SAVEE	7	92	0.92	0.93

Table 4 Cross-dataset performance

Model trained on	Model evaluated on	Accuracy (%)	Precision	Recall	F1-score
SAVEE	TESS	98	0.99	0.98	0.98
TESS	SAVEE	30	0.22	0.30	0.29

various datasets. The findings clearly demonstrate that using the cochleagram as a feature for the SER challenge will aid in obtaining precise and reliable data.

6.1 Cross-Dataset Performance

A series of experiments were conducted to check the model performance on other publicly available datasets. Both the models that were trained were used to evaluate the performance on opposite datasets. TESS model was evaluated on SAVEE dataset and SAVEE model was evaluated on TESS dataset.

From Table 4, we are able to observe that the model was able to generalize better on SAVEE dataset rather than TESS dataset. The reason might have been the audio samples present in SAVEE dataset. More analysis for this is needed to be done.

7 Discussion

The cochleagram feature will assist us in determining the value of cochleagram dimension in the audio domain. The SER challenge has received a lot of attention, and several research papers have been published in this area. There are different features in an audio file that can be used to identify or recognize it, but these features lack the sophistication that the cochleagram dimension can offer, which aids in

correctly recognizing emotions. The developed SER method has numerous uses in various fields such as health care, where a counsellor can take control of a patient's feelings, and it can also aid in user authentication. As the need for understanding emotions has grown for a variety of causes, including the diagnosis of mental illness and diseases, the SER mechanism has become an important part of science.

8 Conclusion

For the role of speech emotion recognition in previous methods, different features were used. We suggested a new function vector, the cochleagram feature vector, in this article, and evaluated it on the TESS dataset, finding that it outperforms previous functions. In addition, potential studies will enhance the classification of cochleagram features, and other classification architectures can be investigated. The results of the study revealed that cochleagram features are successful at identifying emotions from speech details. The suggested method achieved a 97% accuracy rate on TESS dataset and 92% on SAVEE dataset. Cross-dataset evaluation tells us that the model generalizes on a specific type of data. So the results showed 98% accuracy on evaluation on TESS dataset using SAVEE model and 30% accuracy on SAVEE dataset using TESS model. To increase precision and consider what suits best with the cochleagram component of the audio, we should look at various Resnet architectures as well as other classification architectures in the future, along with the performance of model on cross-dataset evaluation.

References

1. B. Athiwaratkun, J.W. Stokes, Malware classification with lstm and gru language models and a character-level cnn. *In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2017), pp. 2482–2486
2. S. Basu, A. Bag, M. Mahadevappa, J. Mukherjee, R. Guha, Affect detection in normal groups with the help of biological markers. *In 2015 Annual IEEE India Conference (INDICON)* (IEEE, 2015), pp. 1–6
3. S. Basu, J. Chakraborty, A. Bag, Md. Aftabuddin, A review on emotion recognition using speech. In *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)* (IEEE, 2017), pp. 109–114
4. M. Bhaykar, J. Yadav, K. Sreenivasa Rao, Speaker dependent, speaker independent and cross language emotion recognition from speech using gmm and hmm. in *2013 National Conference on Communications (NCC)* (IEEE, 2013), pp. 1–5
5. K. Chan, J. Hao, T. Lee, O. Wook Kwon, Emotion recognition by speech signals. In *Proceedings of International Conference EUROSPEECH, Citeseer* (2003)
6. P. Chandrasekar, S. Chapaneri, D. Jayaswal, Automatic speech emotion recognition: a survey. *In 2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA)* (IEEE, 2014), pp. 341–346
7. M. Chen, X. He, J. Yang, H. Zhang, 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Process. Lett. **25**(10), 1440–1444 (2018)

8. M. Chourasia, S. Haral, S. Bhatkar, S. Kulkarni, Emotion recognition from speech signal using deep learning. *in Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2020* (Springer Singapore, 2021), pp. 471–481
9. K. Dupuis, M. Kathleen Pichora-Fuller, *Toronto Emotional Speech Set (TESS)*. (University of Toronto, Psychology Department 2010)
10. H.M Fayek, M. Lech, L. Cavedon, Towards real-time speech emotion recognition using deep neural networks. in *2015 9th International Conference on Signal Processing and Communication Systems (ICSPCS)* (IEEE, 2015), pp. 1–5
11. D.J. France, R.G. Shiavi, S. Silverman, M. Silverman, M. Wilkes, Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans. Biomed. Eng. **47**(7), 829–837 (2000)
12. E. Guizzo, T. Weyde, J.B. Leveson, Multi-time-scale convolution for emotion recognition from speech audio signals. in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2020), pp. 6489–6493
13. L. Guo, L. Wang, J. Dang, L. Zhang, H. Guan, A feature fusion method based on extreme learning machine for speech emotion recognition. in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2018), pp. 2666–2670
14. K. Han, D. Yu, I. Tashev, Speech emotion recognition using deep neural network and extreme learning machine. in *Fifteenth Annual Conference of the International Speech Communication Association* (2014)
15. Z. Huang, M. Dong, Q. Mao, Y. Zhan, Speech emotion recognition using cnn. in *Proceedings of the 22nd ACM International Conference on Multimedia* (2014), pp. 801–804
16. P. Jackson, S. Haq, *Surrey Audio-Visual Expressed Emotion (savee) Database* (University of Surrey, Guildford, UK, 2014)
17. Q. Kong, I. Sobieraj, W. Wang, M. Plumbley, Deep neural network baseline for dcase challenge 2016. *Proceedings of DCASE 2016* (2016)
18. S. Lalitha, A. Mudupu, B.V. Nandyala, R. Munagala, Speech emotion recognition using dwt. in *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)* (IEEE, 2015), pp. 1–4
19. C.M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, S. Narayanan, Emotion recognition based on phoneme classes. in *Eighth International Conference on Spoken Language Processing* (2004)
20. W. Li, B. Mak, Derivation of document vectors from adaptation of lstm language model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (2017), pp. 456–461
21. G. Liu, W. He, B. Jin, Feature fusion of speech emotion recognition based on deep learning. in *2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC)* (IEEE, 2018), pp. 193–197
22. S. Merity, N.S. Keskar, R. Socher, Regularizing and optimizing lstm language models. arXiv:1708.02182 (2017)
23. S. Mirsamadi, E. Barsoum, C. Zhang, Automatic speech emotion recognition using recurrent neural networks with local attention. in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2017), pp. 2227–2231
24. A. Nogueiras, A. Moreno, A. Bonafonte, J.B. Mariño, Speech emotion recognition using hidden markov models. in *Seventh European Conference on Speech Communication and Technology* (2001)
25. Huy Phan, Lars Hertel, Marco Maass, Philipp Koch, Radoslaw Mazur, Alfred Mertins, Improved audio scene classification based on label-tree embeddings and convolutional neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(6), 1278–1290 (2017)
26. Oudeyer Pierre-Yves, The production and recognition of emotions in speech: features and algorithms. Int. J. Human-Comput. Stud. **59**(1–2), 157–183 (2003)
27. B. Schuller, Towards intuitive speech interaction by the integration of emotional aspects. in *IEEE International Conference on Systems, Man and Cybernetics* (vol. 6, IEEE, 2002), p. 6

28. B. Schuller, G. Rigoll, M. Lang, Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing* (vol. 1, IEEE, 2004), pp. I–577
29. M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E.D. Cowie, R. Cowie, Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. in *Proc. 9th Interspeech 2008 incorp. 12th Australasian Int. Conf. on Speech Science and Technology SST 2008, Brisbane, Australia* (2008), pp. 597–600
30. N.B. Wunarso, Y.E. Soelistio, Towards indonesian speech-emotion automatic recognition (i-spear). in *2017 4th International Conference on New Media Studies (CONMEDIA)* (IEEE, 2017), pp. 98–101
31. A.A.A. Zamil, S. Hasan, S.M.D.J. Baki, J.M.D. Adam, I. Zaman, Emotion detection from speech signals using voting mechanism on classified frames. in *2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)* (IEEE, 2019), pp. 281–285

An Analysis of Various Machine Learning Techniques Used for Diseases Prediction: A Review

Mudasir Hamid Sheikh, Sonu Mittal, and Rumaan Bashir

Abstract In recent past, Machine Learning has got more attention in the medical fields to diagnose diseases. To minimise the error rates of computer-aided diagnostic systems, several types of enhancements were made in these systems to get the proper treatment at the right time. Machine Learning is an essential computer-aided diagnosis, and it promises the improved accuracy of diagnosis of any disease. The analysis of high-dimensional biomedical data is done with classy and automatic algorithms provided by machine learning. Our study's primary purpose was to analyse the existing Machine Learning classification algorithms used to predict various diseases. This paper has made a precise analysis of support vector machines, Naïve Bayes algorithm, and random forest algorithms used to predict and classify different types of diseases.

1 Introduction

With the advancements in computer science technologies for the recognition of various diseases in healthcare centres, Machine Learning has played a phenomenal role. Machine Learning (ML) is a sub-area of Artificial Intelligence (AI). In Machine Learning, the machine is being trained on some datasets and later on it is used to test whether the machine is giving correct results or not on some test datasets as shown in Fig. 1. One of the main differences between the traditional programming and ML is that, in traditional programming models, the data and programme are fed to the machine and the desired output is obtained but in ML models, the data and output are fed to the machine and a desired programme is obtained. Therefore,

M. H. Sheikh (✉) · S. Mittal
School of Engineering and Technology, Jaipur National University, Jaipur 302017, Jagatpura, India

S. Mittal
e-mail: dr.sonumittal@jnujaipur.ac.in

R. Bashir
Department of Computer Science, IUST, Awantipora, Kashmir, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_35

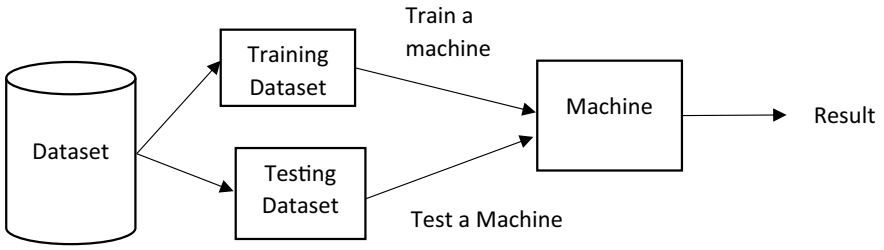


Fig. 1 Machine learning using some data

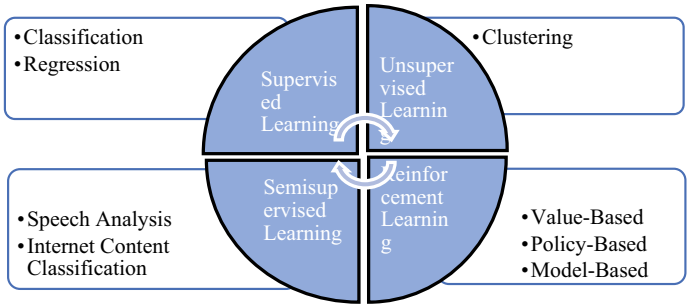


Fig. 2 Machine learning techniques

ML is preferred in each field rather than the other approaches of problem-solving in computer science. The seven stages of ML, namely gathering data, preparing that data, choosing a model, training, evaluation, hyperparameter tuning and prediction.

There are several ML techniques as shown in Fig. 2 which we can use for the prediction of various diseases. Some of the techniques are defined as:

1.1 Supervised Learning [1]

In this type of learning, we have a defined set of outputs. Suppose we have an input variable n and its corresponding output m. Then, we can generalise the supervised learning mathematically as:

$$m = f(n) \tag{1}$$

From equation I, it is clear that for the given input, we will get a defined output. While making a Machine Learning with a given dataset, we train the algorithm with it and train the machine until it produces the correct result. In this technique, we can assume that there is basically a teacher who compares the current result produced with the targeted one and recompiles the code until we get the targeted one. Classification

and regression are the two main categories of supervised learning. In the paper [1], the authors have discussed about the various supervised learning techniques and have expressed them in the mathematical manner also.

1.2 Unsupervised Learning [2]

In this type of learning, while training a machine, the model chosen for training the machine chooses the labels itself and produces the result. In this learning, the target vector is not specified. The machine is continuously trained until it produces a reliable result for the given training dataset. For the processing of video data, unsupervised learning plays an important role [2]. Clustering is one of the very important types of unsupervised learning.

1.3 Semi-supervised Learning [3]

This learning domain of machines comes in between supervised and unsupervised learning techniques. This is one of the best ways for exploiting the unlabelled data. In this type of learning, the machine is trained with both labelled and unlabelled data. The labelled data given to the machine is less as compared to the unlabelled data, i.e. the machine is fed by a large amount of unlabelled data for learning purposes.

1.4 Reinforcement Learning

In this type of ML technique, feedback is very important. The feedback is given back to the machine till the result is obtained. The feedback which is propagated back to the input layers is basically an error. This error is calculated by:

$$\text{Error (e)} = \text{Actual Value} - \text{calculated value} \tag{2}$$

The reinforcement algorithms are not informed by the correct result and we make the machine explore each and every possibility of getting the actual result.

2 Use of Machine Learning Algorithms in Healthcare for the Prediction of Various Types of Diseases

2.1 Naïve Bayes Algorithm

Naïve Bayes algorithm is one of the supervised ML techniques. It is a statistical classifier which uses the concept of conditional probability. Several researchers have used this conditional probability-based algorithm for the classification of various types of diseases. Some of them have used this approach for the classification of heart diseases, lung cancer, pneumonia, etc. The accuracy obtained by using this approach for the disease's classification is shown in Table 1. Shanthi et al. have proposed the early detection of lung cancer as lung cancer is one of the dangerous deaths causing diseases. The authors in paper [4] have used the Naïve algorithm for the prediction of lung cancer and the authors have tested their proposed work on 140 normal and 130 abnormal records present in the dataset. The authors have also computed the accuracy for different proposed methods.

The authors in paper [5] have used Naïve Bayes classifier for the classification of 209 records of patients. The authors have achieved an accuracy of 95.24% and the model has performed well with less complexity.

Sanchayita et al. in paper [6] have used the hybrid approach for the diseases prediction and the authors have used the Naïves algorithm technique on the dataset consisting of 209 records and have achieved 81.33% accuracy.

In year 2018, M.S. Amin et al. have also used several data mining techniques for the prediction of heart diseases. by applying the Naïve algorithm on 303 patient records for the diagnosis of said diseases and have achieved 85.86% of accuracy [7].

In paper [8], Vembandasamy et al. have proposed a heart disease prediction system. The authors have used several data mining approaches and have obtained the 86.4198% of accuracy with minimum time by using the Naïve Bayes algorithm. This algorithm was used to classify the dataset for its diagnosis.

Table 1 Accuracy of various diseases predicted with the help of Naïve Bayes algorithm

Name of the diseases	Dataset	Year	Accuracy	References
Lung cancer	140 normal images 130 abnormal images	2020	–	[4]
Colon cancer	209 patient records	2019	95.24%	[5]
Heart diseases	209 records of patients	2018	81.33%	[6]
Heart diseases	303 records	2018	85.86%	[7]
Heart diseases	Clinical dataset of 500 patients	2015	86.4198%	[8]
Liver diseases	369 images	2015	61.28%	[9]

2.2 Support Vector Machine (SVM)

SVM is also one of the supervised learning techniques. SVM is used for both classification and regression problems. Most of the researchers in the literature have used this algo for the classification of various diseases and by varying the number of records in the dataset, variation in the accuracy is also found. SVM basically creates the hyperplanes in the n-dimensional space and these hyperplanes act as the decision boundaries as shown in Fig. 3.

The authors in paper [9] have used SVM and Navïe algorithms for the prediction of liver diseases. Different accuracies have been obtained for these two different algorithms. The authors have thoroughly compared the SVM and Naïve algorithms on the basis of accuracy and execution time. The Naïve algorithm requires the minimum execution time but the accuracy of this algorithm is lesser than the SVM.

Ali et al. [10] have proposed a new hybrid intelligent system in which three algorithms are used, one for dimensionality reduction, second for classification, and third one for optimisation. The authors have tested the model on 165 patient record dataset and have achieved the accuracy of 90.30%.

In 2020, Sethy et al. in their paper [11], have used one deep learning algorithm for extracting the deep features from the images of the dataset and these features were fed as an input to the SVM classifier for making the right choice for prediction of the coronavirus.

For early detection of breast cancer in patients, the authors in paper [12] have used the mammography images. The important features for the prediction of cancerous cells in breast are detected with the help of Hough transformation and then SVM were used for its classification and this system has achieved the good accuracy of 94.0%.

In paper [13], Mythili et al. have proposed a rule-based model for the prediction of heart diseases using SVM. The authors have tested the proposed model on various already present algorithms in the literature. For accurate prediction of the heart diseases, the authors have combined SVM, logistic regression and decision trees.

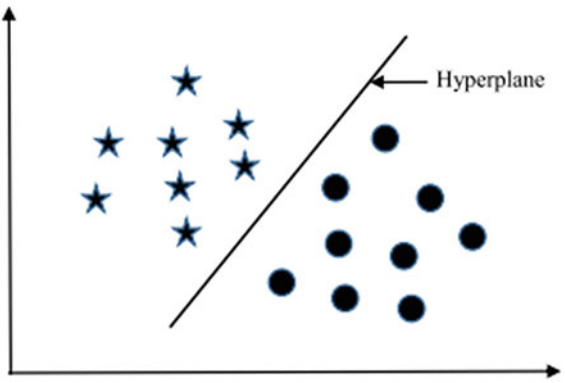


Fig. 3 Hyperplane classifying two data classes in 2-Dimensional Space

Table 2 Accuracy of various diseases predicted with the help of SVM

Name of the diseases	Dataset	Year	Accuracy	References
Hepatocellular carcinoma	165 patient records with49 feature values	2020	90.30%	[10]
Corona virus diseases	127 patient records	2020	95.33%	[11]
Breast cancer	95 mammogram images	2019	94.0%	[12]
Liver diseases	312	2015	79.66%	[9]
Chronic kidney diseases	CKD dataset	2015	0.7375	[14]
Heart disease	303 patient records	2013	–	[13]

The accuracy obtained by applying the SVM for the classification of various diseases is shown in Table 2.

2.3 Random Forest

Random Forest is one of the popular techniques of supervised machine learning [15]. It uses the concept of ensemble learning for solving the classification of various diseases. This is used for solving a large complex problem and it improves the efficiency of the model. Most of the researchers have used this classifier for the classification of various types of diseases and some of these diseases have been listed in Table 3.

In paper [16], the authors have used random forest classifier for the prediction of thyroid disorder. The thyroid is found in almost all the families across the globe because of the differences in lifestyle. Thyroid is of two types and the authors have worked on hypothyroidism and have found that, diseases are predicted with good accuracy by using the random forest classifier as compared to the other applied classifiers. The authors have recorded the accuracy of 0.994 for this classifier.

Table 3 Accuracy of various diseases predicted with the help of Random Forest Classifier

Name of the diseases	Dataset	Year	Accuracy	References
Thyroid disorder	UCI repository with 3772 instances	2019	0.994	[16]
Chronic kidney diseases	400 instances	2019	99.84%	[17]
Osteoarthritis diseases	33 patient data	2018	86.96%	[18]
Lymph diseases	Lymphography Database with 148 instances	2014	92.2%	[19]
Liver, cancer and heart diseases	Few records of patients	2013	57.97%, 75% and 55.55%	[20]

Devika et al. in 2019 have done the comparative analysis of various types of classifiers for the prediction of chronic kidney diseases and have achieved 99.84% of accuracy for random forest classification upon 400 instances of the dataset of kidney diseases patient data [17].

In 2018, Ulfah et al. have applied random classifier on the dataset containing 33 patient records and have divided this into three classes. The authors have achieved an accuracy of 86.96% for predicting whether the patient is suffering from osteoarthritis or not [18].

In paper [19], the authors have used genetic algorithm for the selection of features from the dataset of lymph patients and then applied the classifier for the prediction of lymph diseases. The model has shown the accuracy of 92.2% for its classification.

Ashfaq et al. in 2013 have done a comparative analysis for the prediction of liver, cancer and heart diseases by applying the same classifier on these and have achieved different accuracies for the different types of diseases [20].

3 Analysis and Diagnosis of Various Types of Diseases with the Help of Naïve Bayes Algorithm, Support Vector Machine (SVM) and Random Forest

After proper analysis of SVM, Naïve and all other ML techniques present in the literature which are used for the prediction and analysis of various diseases, it has been observed that the accuracy and computation speed varies from technique to technique. There are advantages as well as disadvantages associated with each of the ML techniques. Some of the advantages of SVM over the other existing ML techniques are:

1. Clear separation line between classes.
2. Memory efficient algorithm.
3. Non-linear datasets are handled efficiently
4. It is used for solving both the classification and regression problems.

Several tools and softwares were identified for the prediction of VAP. In the literature, several authors have used weka and all other statistical tools for the prediction systems. Among these tools, the python libraries are very useful for ML engineers.

3.1 Performance Metrics Used for Classification Models of Machine Learning

Confusion matrix is used for calculating the various performance metrics of the ML model. This matrix has the size of n*n for ‘n’ classes. The matrix is divided into

Table 4 Confusion matrix

Total predictions(n)	Actual [yes]	Actual [no]
Predicted [yes]	True positive	False negative
Predicted [no]	False positive	True negative

two dimensions consisting of predicted values and actual values along with the total number of predictions. The confusion matrix is shown in Table 4.

Using confusion matrix, several parameters are calculated. These are explained below:

Accuracy (A): For determining the accuracy of the classification problems, it is one of the important parameters. The accuracy of any classification problem is calculated by taking the sum of True Positive (TP) and True Negative (TN) divided by sum of TP, TN, False Positive (FP) and False Negative (FN). Mathematically, accuracy is given by:

$$A = \frac{TP + TN}{(TP + TN + FP + FN)} \tag{3}$$

Misclassification Rate (MR): It is also known as error rate. It is used to determine the wrong predictions that are made by the model. It is calculated by dividing the sum of false positive and false negative with the sum of all the attributes of the confusion matrix. Mathematically, it is given as:

$$MR = \frac{FP + FN}{TP + TN + FP + FN} \tag{4}$$

Precision (P): Precision of any model can be calculated by dividing the true positive values obtained by the sum of true positive and false positive values. Mathematically, it is represented as:

$$P = \frac{TP}{TP + FP} \tag{5}$$

Recall (R): When true positive value is divided by the sum of true positive and false negative, recall is obtained. The recall value for any model must be high. Mathematically, it is given as:

$$R = \frac{TP}{TP + FN} \tag{6}$$

4 Conclusion

In this paper, we have done the analysis of various existing ML techniques for the prediction of various types of diseases. After proper analysis of these various techniques of ML, we have observed that SVM is most popular, because of its unique advantages from the other algorithms. It has also been found that for achieving better results and improved decision-making for the prediction of diseases, SVM classifier is best among all. We have also defined several performance metrics that the researchers can use for the evaluation of the prediction models.

References

1. M. Moshkov, B. Zielosko, Supervised learning. pp. 113–126 (2011). https://doi.org/10.1007/978-3-642-20995-6_7
2. D. Greene, P. Cunningham, R. Mayer, Unsupervised learning and clustering, no. February 2017. (2008)
3. G. Kostopoulos, S. Karlos, S. Kotsiantis, O. Ragos, Semi-supervised regression: a recent review. J. Intell. Fuzzy Syst. **35**(2), 1483–1500 (2018). https://doi.org/10.3233/JIFS-169689
4. S. Shanthi, N. Rajkumar, Lung cancer prediction using stochastic diffusion search (SDS) based feature selection and machine learning methods. Neural Process. Lett., 0123456789 (2020). https://doi.org/10.1007/s11063-020-10192-0
5. N. Salmi, Z. Rustam, Naïve bayes classifier models for predicting the colon cancer. IOP Conf. Ser. Mater. Sci. Eng. **546**, 5 (2019). https://doi.org/10.1088/1757-899X/546/5/052068
6. C. Science, C. Science, C. Science, C. Science, C. Science, A hybrid machine learning approach for prediction of heart diseases, pp. 1–6 (2018)
7. M.S. Amin, Y.K. Chiam, K.D. Varathan, Identification of significant features and data mining techniques in predicting heart disease. Telemat. Informatics **36**, 82–93 (2019). https://doi.org/10.1016/j.tele.2018.11.007
8. K. Vembandasamy, R. Sasipriya, E. Deepa, Heart diseases detection using naive bayes algorithm. Int. J. Innov. Sci. Eng. Technol. **2**(9), 441–444 (2015)
9. S. Vijayarani, S. Dhayanand, Liver disease prediction using SVM and Naïve Bayes algorithms. Int. J. Sci. Eng. Technol. Res. **4**, 4, pp. 816–820 (2015)
10. L. Ali, I. Wajahat, N. Amiri Golilarz, F. Keshtkar, S.A.C. Bukhari, LDA–GA–SVM: improved hepatocellular carcinoma prediction through dimensionality reduction and genetically optimized support vector machine. Neural Comput. Appl. 2 (2020). https://doi.org/10.1007/s00521-020-05157-2
11. P.K. Sethy, S.K. Behera, P.K. Ratha, P. Biswas, Detection of coronavirus disease (COVID-19) based on deep features and support vector machine. Int. J. Math. Eng. Manag. Sci. **5**(4), 643–651 (2020). https://doi.org/10.33889/IJMEMS.2020.5.4.052
12. R. Vijayarajeswari, P. Parthasarathy, S. Vivekanandan, A.A. Basha, Classification of mammogram for early detection of breast cancer using SVM classifier and Hough transform. Meas. J. Int. Meas. Confed. **146**, 800–805 (2019). https://doi.org/10.1016/j.measurement.2019.05.083
13. M.T. D. Mukherji, N. Padalia, A. Naidu, A heart disease prediction model using SVM-decision trees-logistic regression (SDL). Int. J. Comput. Appl., **68**, 16, pp. 11–15 (2013). https://doi.org/10.5120/11662-7250
14. P. Sinha, P. Sinha, Comparative study of chronic kidney disease prediction using KNN and SVM. Int. J. Eng. Res. **V4**(12), 608–612 (2015). https://doi.org/10.17577/ijertv4is120622
15. L. Breiman, Random forests. Mach. Learn. **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324

16. D. Mishra, R. Buyya, P. Mohapatra, S. Patnaik, *Intelligent and Cloud Computing*, vol. 2. (2019)
17. R. Devika, S.V. Avilala, V. Subramaniyaswamy, Comparative study of classifier for chronic kidney disease prediction using naive bayes, KNN and random forest. Proc. 3rd Int. Conf. Comput. Methodol. Commun. ICCMC 2019, no. Iccmc, pp. 679–684 (2019). https://doi.org/10.1109/ICCMC.2019.8819654
18. U. Aprilliani, Z. Rustam, Osteoarthritis disease prediction based on random forest. 2018 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2018, pp. 237–240 (2019). https://doi.org/10.1109/ICACSIS.2018.8618166
19. A.T. Azar, H.I. Elshazly, A.E. Hassanien, A.M. Elkorany, A random forest classifier for lymph diseases. Comput. Methods Programs Biomed. **113**(2), 465–473 (2014). https://doi.org/10.1016/j.cmpb.2013.11.004
20. A. AhmedK, S. Aljahdali, S. Naimatullah Hussain, Comparative prediction performance with support vector machine and random forest classification techniques. Int. J. Comput. Appl. **69**, 11, pp. 12–16 (2013). https://doi.org/10.5120/11885-7922

Credit Card Fraud Transaction Classification Using Improved Class Balancing and Support Vector Machines

Pradeep Verma and Poornima Tyagi

Abstract Artificial intelligence and machine learning-based solutions are being implemented in the field of credit card fraud transaction detection. However, there are some practical implications for their implementation considering the nature of the problem. Generally, there exists high-class imbalance in available datasets for credit card fraud detection. The machine learning models trained on such data tend to get biased toward the majority class. Although, these models exhibit high train and validation accuracy, yet they have unacceptably high false positives or false negatives. In such scenarios, missing even a few instances of fraud transactions may be disastrous. A system would be of no use that fails to detect the fraud transactions. No matter what ratio or intervals they occur. That system will be preferred which is able to detect even the rare incidences of fraud transactions. These characteristics of credit card fraud transaction detection make the machine learning training more challenging and special for such systems. The paper proposes a technique for credit card fraud transaction detection by classifying them into two categories fraudulent & legitimate. A technique of class balancing is applied to the dataset. Subsequently, a classifier is trained on the dataset using a nonlinear variant of support vector machines.

Keywords Credit Card Fraud Detection · Machine Learning · Imbalance Dataset · Support Vector Machine · Supervised Learning

1 Introduction

This is the time when the world is on the verge of digital transformation. The ongoing pandemic crisis has fostered the dependency on digital means of payment. They have become more preferable than ever before. Digital transactions are on increase. However, it comes with some challenges as far as establishing security on them is concerned. In 2016, the data of one million SBI card users was breached. There have been many such news before and after that. It is a challenging task to discriminate between a legitimate and a fraud transaction, done using a credit card. Detection of

P. Verma (✉) · P. Tyagi
Himalayan Garhwal University, Pauri, Uttrakhand, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_36

fraudulent credit card transactions has become imperative due to several reasons. There have been many high-profile incidents of data breach that shook the faith of stakeholders in digital payment systems such as credit cards. Now, the companies, monitoring bodies, and even individuals have started looking for some innovative solutions for identifying and preventing these fraud transactions.

This problem has drawn the attention of many researchers in recent times due to its many potential commercial applications. One such application is the development of an intelligent transaction monitoring system that can track the transactions in real time for its legitimacy. If a transaction is found to be suspicious or fraud, then a flag may be raised and intimation can be sent to the user. The researchers have been striving to solve this problem. However, there are many practical challenges on the ground. The real challenge is to make such systems usable in reality.

Most of the systems work well on a constrained environment and with predefined conditions. However, when placed in action to work with the real data, they fail to give satisfactory outcomes. This problem is due to the higher false negatives for the minority class. Since, in most of the datasets available for credit card fraud detection, the fraud transaction class happens to be the minority class. Therefore, even a few instances of false negatives for this class may turn to be unaffordable and unacceptable. Rest of the paper has been organized as follows. Section 2 describes the survey of the related work. Section 3 describes the proposed methodology. Experimental results and detailed description has been provided in Sect. 4. Section 5 gives the conclusion and at last references have been listed.

2 Literature Survey of the Related Work

E-commerce applications are becoming widespread and ever-increasing. People across the globe are preferring the digital means of payment for their various needs ranging from purchase of groceries to payment of health insurance [1, 2]. Credit card has become one of the most preferred ways of doing these transactions. It is popular among online and offline users. The impetus is attributed to several reasons such as credit card users have flexible extended deadlines, a user can purchase something when not having enough cash in a bank account, the repayment can be done in easy equated monthly installments, convenience of use, etc. As of 2020, there are over 55 million credit card users in India alone. Billions of credit card transactions take place every month.

The fraudulent transactions are scattered across this large volume of all transactions. It is difficult to catch them through rule-based techniques [3–6]. However, machine learning-based approaches may be effective to discriminate and distinguish a fraud transaction from a volume of all transactions. In practical scenarios, most of the time the data has missing values, is in inappropriate form, has very large or less attributes, repetitive records, imbalances, and also inaccurate [5]. In order to make the most of the machine learning models, the data needs proper preprocessing and cleansing steps. Data mining approaches[2, 3], dimensionality reduction [4],

feature engineering [4, 5], association rules [6], machine learning [7–19], and optimization & class balancing techniques [20–25], all play a critical role in the task of labelling and instance of credit card transaction into one of the available classes, i.e., legitimate or fraudulent. The design of the fraud detection systems majorly rely on the performance of underlying machine learning algorithms.

There are many predicaments for the accuracy and effectiveness of these algorithms such as imbalance of data belonging to each class, non-availability of data due to various privacy laws, and continuous streams of data [1]. Andrea et al. [1], in their works try to answer some of the very pertinent questions related to implementation of a credit card fraud detection system. Some of these questions include what are the best performing approaches to work on a real dataset, what volume of tractions is sufficient to train the model, and how frequently should the model be updated to stay effective & relevant.

2.1 Machine Learning Approaches for Credit Card Fraud Detection

Researchers are used various machine learning-based credit card fraud transaction detection algorithms such as support vector machines [3], Hidden Markov models (HMM) [7], Bayesian network classifiers [9–11], decision tree classifiers [11], artificial neural networks [12], self-organizing maps [14], linear discriminant analysis, deep learning approaches [10, 13], and others [15, 16]. The machine learning approaches can be divided into two categories, namely—supervised machine learning approaches and unsupervised learning approaches.

Supervised machine learning approaches need the labelled training dataset. These algorithms learn the underlying discriminant features in data and try to fit a model that can be used to classify the unseen data in one of the classes of the data [3]. The supervised machine learning approaches can be either end to end machine learning approaches or feature-based approaches. The end to end approaches are those which can be trained on any type of data. These techniques are more robust and versatile [10]. Deep learning approaches based on neural networks, convolutional neural networks (CNN), and other variants such as Recurrent Neural Networks (RNN) and Long Short Term Memory Networks (LSTM) fall under this category of end to end machine learning [13]. Another category of supervised machine learning contains feature-based approaches. These methods can work only on the numerical representation of the data. If the data is presented in some other form, it must be first converted into the numerical representation with a preceding step called feature extraction.

Some of the popular feature-based machine learning algorithms are k-nearest neighbors, support vector machines, decision trees, logistic regression, random forests, etc. On the other hand, unsupervised machine learning algorithms do not require the labelled data. These algorithms can work on the unlabelled data. They

may deduce the patterns from the data without any supervisory information or examples provided in supervised counterparts. Unsupervised algorithms can find the clusters in the data and even can be used to establish & explore the association rules in existing data. Furthermore, these techniques can be used to augment data and to generate new samples. Unsupervised techniques may learn the distributions from given data and produce newer samples.

3 Methodology

Here, we are using a variant of a support vector machine. The rationale behind choosing support vector machines is their relatively better performance for binary classification problems. Present problem of credit card fraud detection happens to be a binary classification task. Therefore, we have chosen here a variant of support vector machine due to its more suitability and better empirical performance on binary classification.

The methodology is being explained in three steps, namely—feature scaling, hybrid class balancing, and support vector classification.

3.1 Feature Scaling

On numerical dataset, some features (attributes) may have very large or small values. It may affect the performance of the classifier and it may skew toward the smaller range features. Here, a simple scaling mechanism is applied to normalize the values uniformly. From each feature, the mean value is subtracted and scaled to unit variance. Unit variance is achieved with division by the standard deviation. It results in the dataset to convert into a distribution having standard deviation as 1.

3.2 Hybrid Class Balancing

There exist many techniques for balancing an unbalanced dataset such as oversampling the minority class, synthetic minority oversampling technique (SMOTE), undersampling the majority class, etc. We use a hybrid balancing technique having undersampling and a borderline SMOTE. Following two steps are performed for resampling:

I. Upsampling is performed by generating some synthetic samples of minority class in the dataset. Instead of doing this blindly, the samples are created for the minority class along the decision boundary existing between the majority and the minority classes.

II. The downsampling step involves reducing a fraction of samples for the majority class. It is done randomly for different proportions of the dataset.

Above two steps are performed on the dataset to find the most suitable threshold values for upsampling and undersampling that optimize the overall performance of the method.

3.3 Support Vector Machine Model

Support vector machines can be put into two categories: nonlinear and linear SVM. The linear SVM can work only in case of linearly separable data. Below is Figure 1 which shows discrimination of legitimate and fraudulent transactions separated by hyperplane using Support Vector Machine.

For a given sets S0 and S1 in an n-dimensional Euclidean space, the two sets are called linearly separable if are n + 1 real numbers ×1, ×2,…, xn, k; satisfying the following condition:

$$\forall_S \in S0, \sum_{i-1}^{n} si\ xi > k \text{ and } \forall_S \in S1, \sum_{i-1}^{n} si\ xi < k \tag{1}$$

If the data does not satisfy the condition in Eq. (1), then linear SVM cannot work. Present situation happens to be a case of nonlinearly separable data. In order to classify the nonlinearly separable data, the data first needs to be projected to a higher dimensionality space. This approach is called "kernel trick."

This step helps in drawing a hyperplane that can easily learn the discriminative features from the data. There are various kernel functions such as linear kernel, Gaussian kernel, Bessel function kernel, etc.

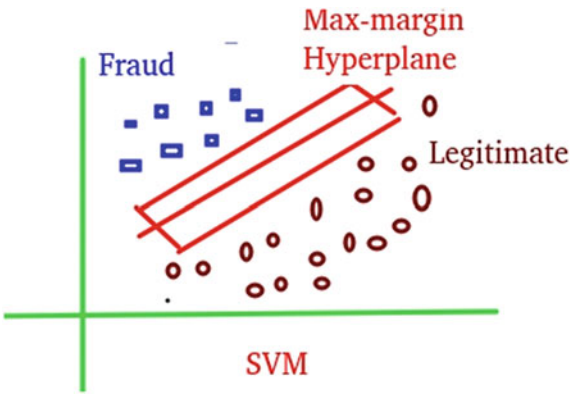


Fig. 1 Support vector machine linear feature discrimination using hyperplane

Experimental results reveal that the Gaussian kernel method is the most suited kernel for the given problem of credit card fraud detection.

$$G(x, y) = \exp\left(-\left(||x - y||^2/\sigma^2\right)\right) \tag{2}$$

The Gaussian kernel is calculated using the formula given in equation no. (2). Here, x and y represent the data point and σ^2 is the variance.

4 Experimental Results

Objective here is to train the machine learning classifier for classifying a transaction instance into one of the category classes viz. 0 (legitimate) and fraud (1). The WMLG credit card transactions dataset has been used for this purpose [26]. The dataset contains a total of 284,807 credit card transactions made by European customers in September 2013.

The given dataset happens to be highly imbalanced. There are only 0.172% fraud transactions (492 instances) of all the transactions in the dataset. There are a total of 31 attribute columns in the dataset.

The last column represents the class label, having values 0 (for legitimate transaction) and 1 (for fraud transaction). Except two attributes, namely, Time and Amount, the rest of the 28 non-class columns have masked column names due to some secrecy concerns. These columns have names V1, V2, …, V28. A few samples have been taken from the dataset and are shown in Table 1. The data has been transposed to keep the attributes in column 1 and samples in the rest of the columns. It is evident from the result that all the values in the dataset happen to be numerical values.

4.1 Exploratory Analysis on the Dataset

The exploratory data analysis on credit card dataset reveals that there are no missing values in the dataset. All the columns have numeric values and dataset class labels are highly imbalanced.

Figure 2a shows the pie chart of samples belonging to each category. It shows a very high imbalance in the dataset. Figure 2b shows the distribution plot for values to each of the attributes V1–V28. Figure 3 shows the correlation of each attribute with every other attribute in the dataset. For instance, attribute Time has a high positive correlation with V3, V11, and V25 attributes.

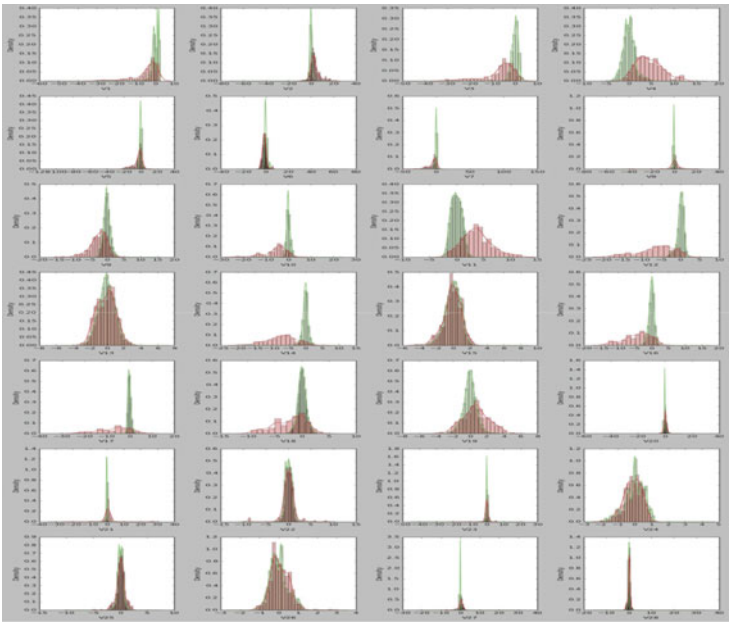
Table 1 Feature attributes and some sample records from dataset

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
Time	0.000000	0.000000	1.000000	1.000000	2.000000
V1	1.191857	−1.359807	−0.966272	−1.358354	−1.158233
V2	0.266151	−0.072781	− 0.185226	−1.340163	0.877737
V3	0.166480	2.536347	1.792993	1.773209	1.548718
V4	0.448154	1.378155	−0.863291	0.379780	0.403034
V5	0.060018	−0.338321	−0.010309	− 0.503198	−0.407193
V6	−0.082361	0.462388	1.247203	1.800499	0.095921
V7	−0.078803	0.239599	0.237609	0.791461	0.592941
V8	0.085102	0.098698	0.377436	0.247676	−0.270533
V9	−0.255425	0.363787	−1.387024	−1.514654	0.817739
V10	−0.166974	0.090794	−0.054952	0.207643	0.753074
V11	1.612727	−0.551600	−0.226487	0.624501	−0.822843
V12	1.065235	−0.617801	0.178228	0.066084	0.538196
V13	0.489095	−0.991390	0.507757	0.717293	1.345852
V14	−0.143772	−0.311169	−0.287924	−0.165946	−1.119670
V15	0.635558	1.468177	−0.631418	2.345865	0.175121
V16	0.463917	−0.470401	−1.059647	−2.890083	−0.451449
V17	−0.114805	0.207971	−0.684093	1.109969	−0.237033
V18	−0.183361	0.025791	1.965775	−0.121359	−0.038195
V19	−0.145783	0.403993	−1.232622	−2.261857	0.803487
V20	− 0.069083	0.251412	−0.208038	0.524980	0.408542
V21	−0.225775	−0.018307	−0.108300	0.247998	−0.009431
V22	−0.638672	0.277838	0.005274	0.771679	0.798278
V23	0.101288	−0.110474	−0.190321	0.909412	−0.137458
V24	−0.339846	0.066928	−1.175575	−0.689281	0.141267
V25	0.167170	0.128539	0.647376	−0.327642	−0.206010
V26	0.125895	−0.189115	−0.221929	−0.139097	0.502292
V27	−0.008983	0.133558	0.062723	−0.055353	0.219422
V28	0.014724	−0.021053	0.061458	−0.059752	0.215153
Amount	2.690000	149.620000	123.500000	378.660000	69.990000
Class	0.000000	0.000000	0.000000	0.000000	0.00000

4.2 Performance Analysis of the Proposed Method

Before applying the training to the classifier, the outliers are removed from the data and hybrid class balancing technique, as discussed in Sect. 3.2, is applied. After these steps, the majority class has 62% samples and the minority class has 38% samples.

(a)



(b)

Fig. 2 **a** Distribution of class labels in dataset, **b** Distribution of masked attributes V1–V28

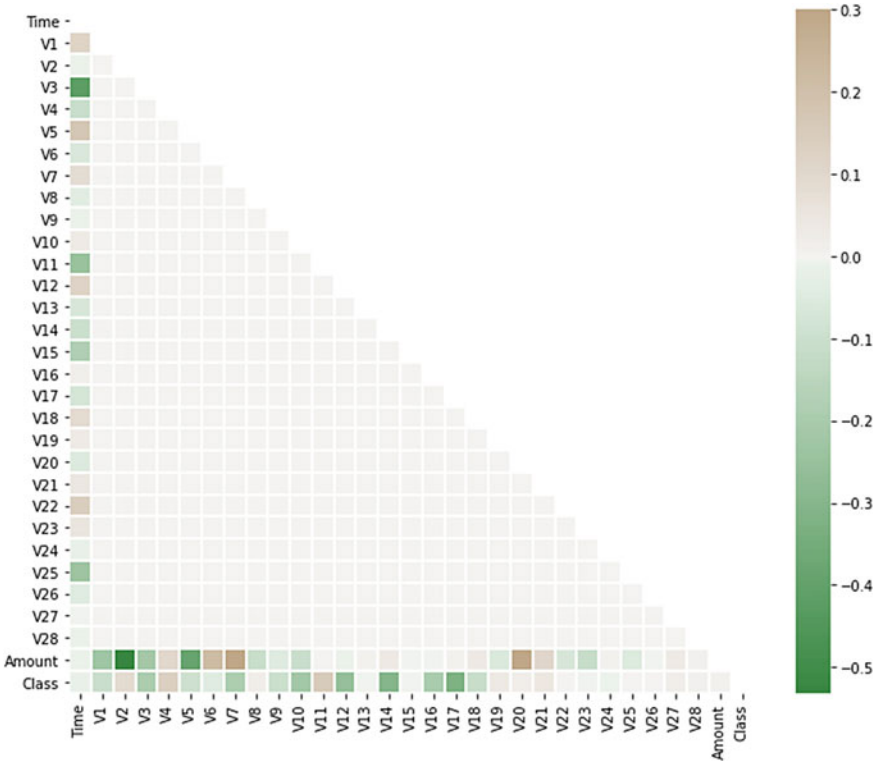


Fig. 3 Attribute correlation heat map

The proposed technique is able to achieve an accuracy of 93.76%. However, being an imbalanced classification problem, mere accuracy cannot be a sufficient performance parameter for assessing the classifier's performance.

Therefore, for a more rational assessment of the proposed method AUC-ROC curve has been used. It is a curve drawn between true positive rate and false positive rate of the method at multiple threshold values. Figure 4 shows the given curve. It is evident from the figure that the AUC score of the method happens to be 0.7625 (76.25%) which is quite high given the difficulty level of the problem.

On the contrary, looking at Fig. 5, we can see that the AUC percentage of simple SVM classifier and SVM with SMOTE classifier are 0.583 and 0.635, respectively. Further, the classification accuracy of the proposed method is 93.76%, which is higher than SVM and SVM + SMOTE methods. Therefore, the proposed method performs well on the given problem of credit card fraud classification.

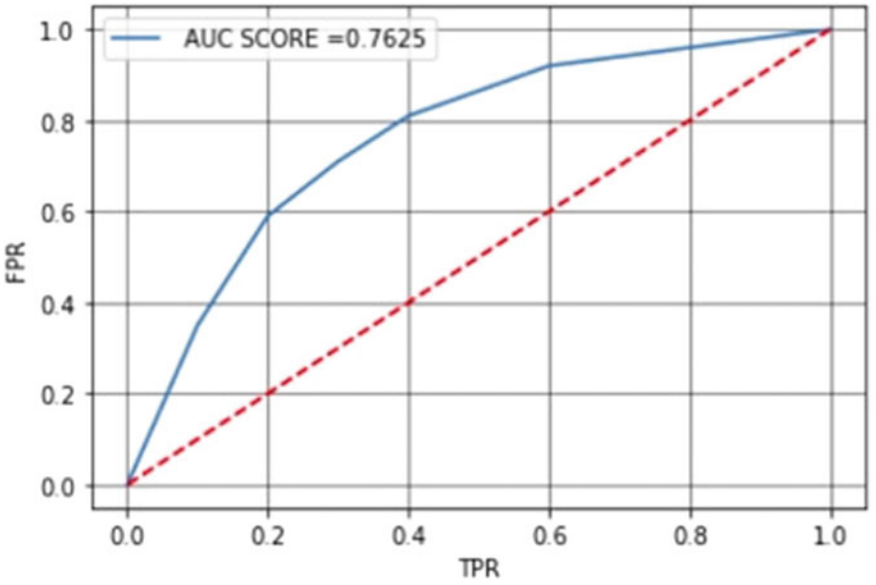


Fig. 4 AUC-ROC curve for the proposed method

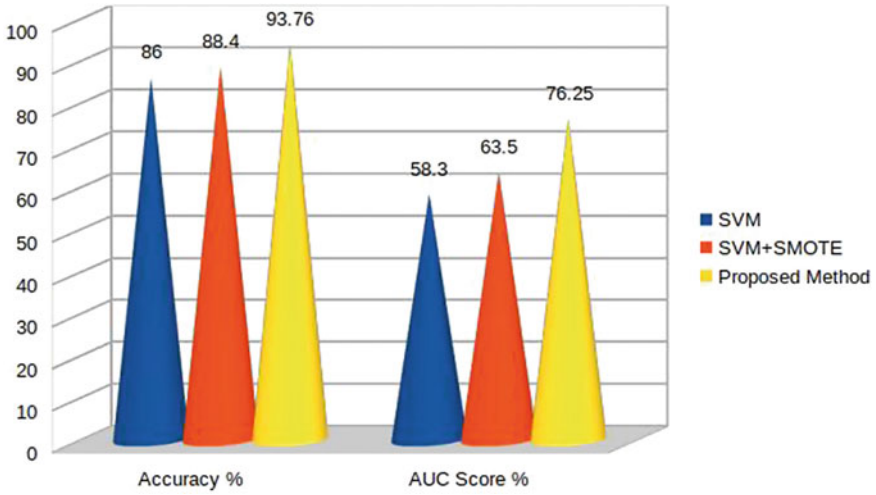


Fig. 5 Performance comparison with other methods

5 Conclusion and Discussion

Credit card fraud detection happens to be a challenging task due to high stakes of missing even some instances of fraud. Further, there is high-class imbalance in available datasets for credit card fraud detection. It makes the trained models get

biased toward the majority class and makes them exhibit high train and validation accuracies, yet they have unacceptably high false positives or false negatives. The paper addresses this issue by applying a hybrid technique for class imbalance. Further, a variant of support vector machines has been employed to classify the transactions into two classes. Experimental results show that the proposed technique is accurate and can detect the instances of fraud transactions with better precision and recall as compared with simple support vector machine approach.

References

1. A. Dal Pozzolo, O. Caelen, Y.A. Le Borgne, S. Waterschoot, G. Bontempi, Learned lessons in credit card fraud detection from a practitioner perspective. Expert Syst. Appl. **41**(10), 4915–4928 (2014)
2. C. Whitrow, D.J. Hand, P. Juszczak, D. Weston, N.M. Adams, Transaction aggregation as a strategy for credit card fraud detection. Data Min. Knowl. Disc. **18**(1), 30–55 (2009)
3. S. Bhattacharyya, S. Jha, K. Tharakunnel, J.C. Westland, Data mining for credit card fraud: a comparative study. Decis. Support Syst. **50**(3), 602–613 (2011)
4. A.C. Bahnsen, D. Aouada, A. Stojanovic, B. Ottersten, Feature engineering strategies for credit card fraud detection. Expert Syst. Appl. **51**, 134–142 (2016)
5. Y. Lucas, P.E. Portier, L. Laporte, L. He-Guelton, O. Caelen, M. Granitzer, S. Calabretto, Towards automated feature engineering for credit card fraud detection using multi-perspective HMMs. Future Gener. Comput. Syst. **102**, 393–402 (2020)
6. D. Sánchez, M.A. Vila, L. Cerda, J.M. Serrano, Association rules applied to credit card fraud detection. Expert Syst. Appl. **36**(2), 3630–3640 (2009)
7. A. Srivastava, A. Kundu, S. Sural, A. Majumdar, Credit card fraud detection using hidden Markov model. IEEE Trans. Dependable Secur. Comput. (2008)
8. K. RamaKalyani, D. UmaDevi, Fraud detection of credit card payment system by genetic algorithm. Int. J. Sci. Eng. Res. **3**(7), 1–6 (2012)
9. S. Panigrahi, A. Kundu, S. Sural, A.K. Majumdar, Credit card fraud detection: a fusion approach using Dempster-Shafer theory and Bayesian learning. Inf. Fusion **10**(4), 354–363 (2009)
10. Z. Li, G. Liu, C. Jiang, Deep representation learning with full center loss for credit card fraud detection. IEEE Trans. Comput. Soc. Syst. **7**(2), 569–579 (2020)
11. A. Husejinovic, Credit card fraud detection using naive Bayesian and c4. 5 decision tree classifiers. Husejinovic, A. (2020). Credit Card Fraud Detect. Using Naive Bayesian C **4**, 1–5 (2020)
12. S. Ghosh, D.L. Reilly, Credit card fraud detection with a neural-network. in *System Sciences, 1994. Proceedings of the Twenty-Seventh Hawaii International Conference on* (vol. 3). IEEE, pp. 621–630 (1994)
13. A. Roy, J. Sun, R. Mahoney, L. Alonzi, S. Adams, P. Beling, Deep learning detecting fraud in credit card transactions. in *2018 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, pp. 129–134 (2018)
14. V. Zaslavsky, A. Strizhak, Credit card fraud detection using self-organizing maps. Inf. Secur. **18**, 48 (2006)
15. F. Carcillo, Y.A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, G. Bontempi, Combining unsupervised and supervised learning in credit card fraud detection. Inf. Sci. (2019)
16. R.J. Bolton, D.J. Hand, Unsupervised profiling methods for fraud detection. Credit Scoring Credit Control **VII**, 235–255 (2001)
17. J. Jurgovsky, M. Granitzer, K. Ziegler, S. Calabretto, P.E. Portier, L. He-Guelton, O. Caelen, Sequence classification for credit-card fraud detection. Expert Syst. Appl. **100**, 234–245 (2018)

18. H. Zhu, G. Liu, M. Zhou, Y. Xie, A. Abusorrah, Q. Kang, Optimizing weighted extreme learning machines for imbalanced classification and application to credit card fraud detection. Neurocomputing **407**, 50–62 (2020)
19. A.A. Taha, S.J. Malebary, An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. IEEE Access **8**, 25579–25587 (2020)
20. Y. Jian, M. Ye, Y. Min, L. Tian, G. Wang, FORF-S: a novel classification technique for class imbalance problem. IEEE Access **8**, 218720–218728 (2020)
21. S. Ancy, D. Paulraj, Handling imbalanced data with concept drift by applying dynamic sampling and ensemble classification model. Comput. Commun. **153**, 553–560 (2020)
22. F. Kamalov, D. Denisov, Gamma distribution-based sampling for imbalanced data. Knowl. Based Syst. **207**, 106368 (2020)
23. X. Wang, J. Xu, T. Zeng, L. Jing, Local distribution-based adaptive minority oversampling for imbalanced data classification. Neurocomputing **422**, 200–213 (2021)
24. O.D. Myers, S.J. Sumner, S. Li, S. Barnes, X. Du, One step forward for reducing false positive and false negative compound identifications from mass spectrometry metabolomics data: new algorithms for constructing extracted ion chromatograms and detecting chromatographic peaks. Anal. Chem. **89**(17), 8696–8703 (2017)
25. A.A. Renshaw, E.W. Gould, Reducing false-negative and false-positive diagnoses in anatomic pathology consultation material. Arch. Pathol. Lab. Med. **137**(12), 1770–1773 (2013)
26. Worldline and the machine learning group, credit card fraud detection dataset, (2013). https://www.kaggle.com/mlg-ulb/creditcardfraud. Last accessed 20 Mar 2020

An Improved Lossless Algorithm for Text Compression

Kartik Bhatia, Anupam Singh, Anamol Verma, and Dipansh Mittal

Abstract This paper provides a method of lossless, adaptive, and asymmetric text compression at bit-level and comparison among text compression techniques. To increase efficiency, it can be paired with other statistical compression techniques. This is a mathematical approach, where text data is first converted to binary and afterward, utilizing a character word length of 8-bits it additionally compresses the binary codes obtained. Our proposed algorithm, reckons an ideal character word length b holding a condition that b must be greater than 8, which raises the ratio of compression by a factor of, [b/8]. To espouse this algorithm, it is exercised as complementary with the implementation of the Huffman Algorithm to compress a source text file.

Keywords Data compression · Lossless · Huffman coding · Binary codes

1 Introduction

Data compression [1, 2] uses fewer bits to represent the original text data. It can be implemented anywhere on various file types. Before data compression, a 1000-word text file generally covered 100 KB size, while a 10,000-word text file reached MB size. Later data compression approaches shrink large files into much smaller ones. It tried to help in conserving storage capacity, accelerating file transmission, and reducing costs for hardware storage and network capacity operations.

Figure 1 depicts a basic diagram, where when data compression techniques are applied to the original file, a compressed file is obtained. Similarly, when data decompression techniques are applied to the compressed file, the original file is obtained.

K. Bhatia · A. Singh (✉) · A. Verma · D. Mittal
School of Computer Science, UPES, Dehradun, Uttarakhand, India
e-mail: anupam.singh@ddn.upes.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_37

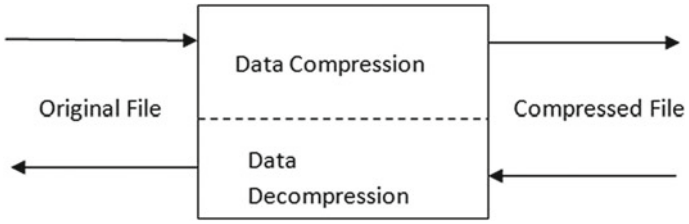


Fig. 1 Data compression

2 Background

2.1 Data Compression Methods

Data compression is used to minimize the file size. There are two types of compression: Lossless [3, 4] and Lossy [5]. These approaches are identified for specific purposes. We can not apply lossy compression on text data as the meaning of information will be changed. Lossy compression is the one that loses some data to achieve compression, while Lossless compression is the one that keeps all the data. The key here is to find a smarter technique to encode the data.

Lossless compression sanctions the file to return to its original file size, without losing a single bit of data, after decompression. Lossless compression is helpful where the loss of a single bit would change the meaning of information. It can compress the data at whatever point repetition is available. Hence, lossless compression exploits data repetition.

Lossy compression eliminates insignificant data, permanently. Lossy compression is suitable in cases where removing some data bits does not affect the appearance of the content. In lossy compression, messages become more proficient by disposing of undesirable data. Lossy compression lessens the size of the data while preserving information.

2.2 Types of Data Compression

There are three types of compression methods: Substitution, Statistical, and Dictionary-based compression.

The substitution data compression technique [6] swaps the imitating characters by a shorter character depiction.

The Statistical data compression technique generates shorter binary codes based on the possibility of the rate of recurrence of the characters in the given text file. Shannon–Fano coding and adaptive/static Huffman coding come under this technique.

The Dictionary data compression technique substitutes the sub-strings of the text according to its indices by referring to the dictionary of the sub-strings.

In this paper, we are involved with the statistical data compression technique that generates the briefest binary code for each character according to their rate of recurrence in the text file.

Huffman coding [7–9] is a lossless data compression algorithm, industrialized by David A. Huffman, in 1951. It uses the information of the frequency of characters to assign variable-length codes to each character. More frequent characters are assigned shorter binary codes, thus reducing the length of the binary sequence. The initial step in producing a Huffman code is to arrange the characters in descending order according to their rate of occurrence in the text file. The next step is to compose a binary tree structure.

To compose a binary tree, we group the two least frequent characters and add their frequencies. Similarly, pick the next two elements with the lowest rate of recurrence, group them and add their rate of recurrence, until one element remains.

Shannon–Fano [10–12] coding is named after Claude Elwood Shannon and Robert Fano. It is the predecessor of Huffman coding. It uses a similar algorithm where it generates shorter binary codes for more frequently occurring characters and comparatively longer binary codes for less frequently occurring characters and the prefix principle still applies. However, the difference lies in the approach used by the algorithms. Huffman coding uses a bottom-up approach of determining the binary codes for each symbol, whereas Shannon-Fano uses a top-down approach.

3 Proposed Method

In statistical compression algorithms [13–16], first, the possibilities (frequencies) for all characters in the text file is computed. The characters incorporate all letters, digits, and accentuations. These characters are to be placed in descending order, i.e., from highest to lowest frequencies. The next step is to locate the equivalent binary code for every novel character according to the statistical data compression technique. Here Huffman coding is used, which relegates the most widely recognized characters in the text file to the briefest binary codes, and the least character the longest. At that point, these binary codes are utilized to change overall characters in the text file to a binary code. In general, in all statistical algorithms, a character word length of 8 bits is used, where the corresponding decimal value for, each 8-bits (0–255) is reckoned and converted to a character that is written to the output file.

For an 8-bits character word length, there are 0–255 potential decimal values, i.e., 256 probabilities. In most cases, not all 256 potential decimal values are used. So, in this new methodology, we try to use more than 8-bits to store a single character. This is not a concern; what is significant is the quantity of the distinctive decimal values (conceivable outcomes) that have been determined, and if they are under 256 prospects, it is feasible to utilize a character word length of b-bits to store the

character instead of 8-bits. It is crucial to track down an ideal incentive for b which produces decimal values that are less than or equal to 256 probabilities.

In this new approach, a coder will go through the data file, convert each character into its suitable binary code, and compose the subsequent bits to the output file in the wake of improving the compressed character word length. All together not to get the codes stirred up in the decoding process, all information is stored in the compressed file header.

3.1 Proposed Algorithm

The proposed algorithm for the compression and decompression of the file is presented as follows.

Novel Compression

1. Open character file (.txt) as input in reading mode.
2. Calculate the rate of recurrence for every novel character.
3. Arrange the character frequencies in sequence from the highest rate of recurrence to the lowest rate of recurrence.
4. Coin a leaf node for every novel character and construct a min-heap for all leaf nodes.
5. Drag two nodes with the lowest rate of recurrence from a min-heap.
6. Coin a new internal node with a rate of recurrence equal to the sum of two node's rate of recurrence. Sort the first extracted node as its left child and another extracted node as its right child. Insert node to a min-heap.
7. Rehash steps 5 and 6 until the heap contains just a single node. The excess node is the root node, and the tree is complete. Develop the binary code for the original data file utilizing Huffman coding.
8. Store the binary codes in a file and shift it's current pointer position with the last binary digit.
9. Calculate the decimal equivalent of the current binary digit and check if it is greater than 127. If not, then shift the pointer to its left until summation sums up to 127.
10. If the decimal value is greater than 127, add 33 to it and store it in a compressed file with the addition of constant value 33 in it because decimal values between 0 and 32 are not printable. Store the block information in its header file and set the block pointer and decimal variable to 0.
11. Repeat Step 9 and Step 10 till the first character of the file is reached.
12. The compressed file of the original file will be in "compressed.txt."

Decompression

1. Open compressed file in reading mode.
2. Read a character from the "compressed.txt" file.
3. Convert character to its decimal equivalent and subtract 33 from it.

4. Write equivalent binary digits to “binary.dat”
5. Repeat steps 2–4 till the end of the file is reached.
6. Open “binary.dat” in reading mode and move the file pointer to the end.
7. Read a bit from “binary.dat.”
8. If it matches any Huffman code write the character to its uncompressed file.
9. If it does not match, then include more bit.
10. Repeat Step 7,8 till no more bits remains.
11. Close the uncompressed file and delete all intermediate files.
12. END.

4 Studies and Findings

Unlike Huffman coding, the Shannon–Fano [13, 17] method has a lower probability to accomplish the lowest possible estimated codeword length. Both Huffman coding and Shannon–Fano generate prefix codes. However, the codes that are produced from Shannon–Fano coding are not optimal, while the codes generated from Huffman coding produces optimal results. Contingent upon the area of application, there are many ways to make sure the performance of compression algorithms. The compression performance depends on the nature of compression: **lossy** or **lossless**. The space efficiency and time effectiveness would be more prominent in the event of a lossy compression algorithm when used to compress a specific source document and lower if there should arise an occurrence of the lossless compression algorithm.

Compression Ratio (CR)—it is the fraction of the size of the compressed file and the size of the source file.

$$\text{Compression Ratio} = \text{size after compression}/\text{size before compression} \quad (1)$$

where, size after compression = compressed file size, and size before compression = original file size.

Saving Percentage (SP)—it reckons the reduction of source file size in percentage.

$$\text{Saving Percentage} = (\text{size before compression} - \text{size after compression})/\text{size before compression} \quad (2)$$

where, size after compression = compressed file size, and size before compression = original file size.

Every one of the strategies characterized above guarantees the adequacy of compression calculations utilizing distinctive file sizes. There are different strategies to guarantee the functioning of compression calculations, for example, compression time, computational complexity, and probability distribution.

4.1 Comparison Between Huffman and Shannon–Fano

The following statistics display a comparison [18, 19] between the existing algorithms, i.e., Huffman Coding algorithm and Shannon-Fano algorithm (Fig. 2 and Fig. 3).

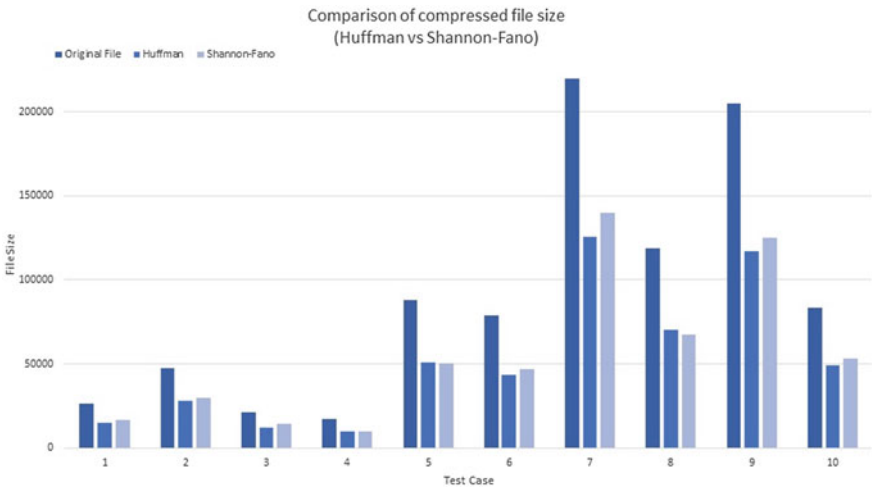


Fig. 2 Shows comparison between Huffman and Shannon–Fano where it depicts, Huffman provides slightly better compression than Shannon–Fano

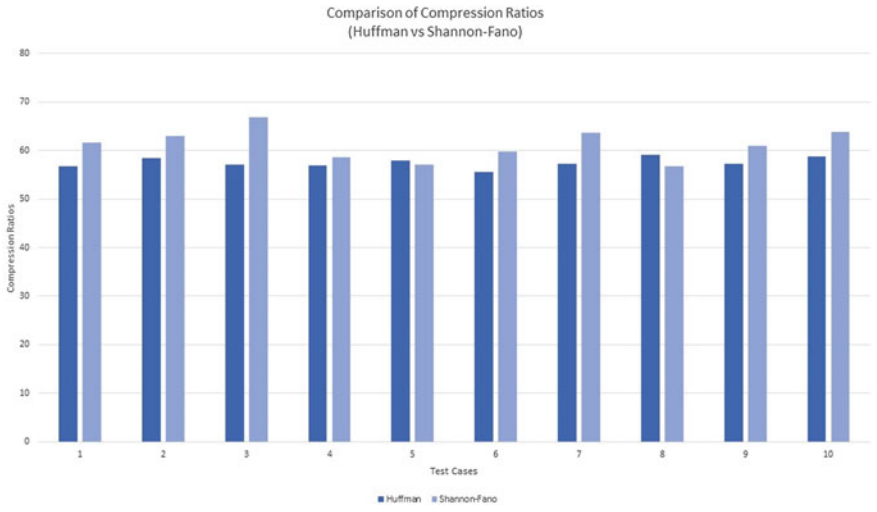


Fig. 3 Shows comparison between the compression ratios of Huffman and Shannon–Fano algorithm

In Table 1, Huffman coding provides us with,
In Table 2, Shannon-Fano provides us with,

Table 1 Huffman algorithm

File	Original file	Compressed file	CR	SP
1	26,574	15,068	56.70	43.29
2	47,654	27,801	58.34	41.66
3	21,076	12,030	57.08	42.92
4	16,908	9629	56.95	43.05
5	87,908	50,934	57.94	42.05
6	78,640	43,645	55.50	44.5
7	219,819	125,736	57.20	42.80
8	118,564	70,000	59.04	40.96
9	204,798	117,246	57.25	42.75
10	83,427	48,972	58.70	41.29

Average compression ratio = 57.47
Average saving percentage = 42.52

Table 2 Shannon–Fano

File	Original file	Compressed file	CR	SP
1	26,574	16,391	61.68	38.32
2	47,654	30,012	62.98	37.02
3	21,076	14,099	66.90	33.10
4	16,908	9894	58.52	41.48
5	87,908	50,160	57.06	42.94
6	78,640	47,058	59.84	40.16
7	219,819	139,848	63.62	36.38
8	118,564	67,285	56.75	43.25
9	204,798	124,967	61.02	38.98
10	83,427	53,243	63.82	36.18

Average compression ratio = 61.22
Average saving percentage = 38.78

4.2 *Comparison Between the Proposed Algorithm, Huffman, and Shannon-Fano*

The following statistics display a comparison [18, 19] between the proposed algorithm and the existing algorithms, i.e., Huffman Coding algorithm and Shannon–Fano algorithm.

In Table 3, Huffman coding provides us with,
In Table 4, Shannon–Fano coding provides us with,
In Table 5, the Proposed method provides us with,

Table 3 Huffman

File	Original file	Compressed file	CR	SP
1	24,567	13,902	56.59	43.40
2	45,756	26,858	58.70	41.29
3	12,076	7001	57.98	42.01
4	16,908	9620	56.90	43.10
5	77,890	44,506	57.14	42.86
6	40,786	22,693	55.64	44.36
7	121,989	70,216	57.56	42.44
8	181,456	106,950	58.94	41.06
9	240,789	136,647	56.75	43.25
10	73,428	43,506	59.25	40.75

Average compression ratio = 57.55
Average saving percentage = 42.45

Table 4 Shannon–Fano

File	Original file	Compressed file	CR	SP
1	24,567	15,374	62.58	37.42
2	45,756	28,222	61.68	38.32
3	12,076	8139	67.40	32.60
4	16,908	9857	58.30	41.70
5	77,890	45,222	58.06	41.94
6	40,786	24,039	58.94	41.06
7	121,989	76,401	62.63	37.37
8	181,456	104,791	57.75	42.24
9	240,789	146,688	60.92	39.08
10	73,428	46,406	63.20	36.80

Average compression ratio = 61.14
Average saving percentage = 38.85

Table 5 Proposed method

File	Original file	Compressed file	CR	SP
1	24,567	15,708	63.94	36.06
2	45,756	28,455	62.19	37.81
3	12,076	8211	68.00	31.99
4	16,908	9975	59.10	40.91
5	77,890	46,095	59.18	40.82
6	40,786	25,009	61.32	38.68
7	121,989	72,729	59.62	40.38
8	181,456	115,224	63.50	36.50
9	240,789	158,487	65.82	34.18
10	73,428	43,058	58.64	41.36

Average compression ratio = 62.13
Average saving percentage = 37.87

To ensure the performance of our proposed algorithm, it is exercised as complementary with the implementation of the Huffman Algorithm to compress a source text file. In Table 5, it can be seen that the proposed algorithm provides better compression than Huffman and Shannon–Fano (Fig. 4).

The proposed algorithm uses b–bits to compress the text file. The most favorable value of b depends on several aspects:

- The size and nature of the data file.
- The character's rate of recurrence within the data file.
- The conveyance of characters inside the file.

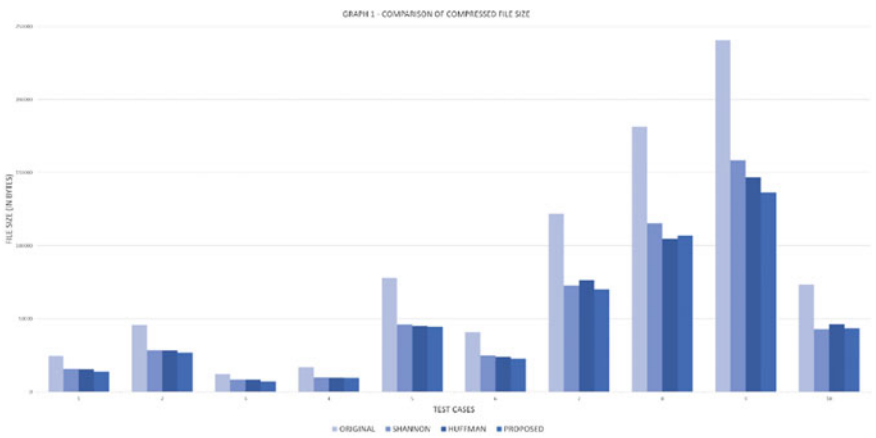


Fig. 4 Comparison of compressed file size

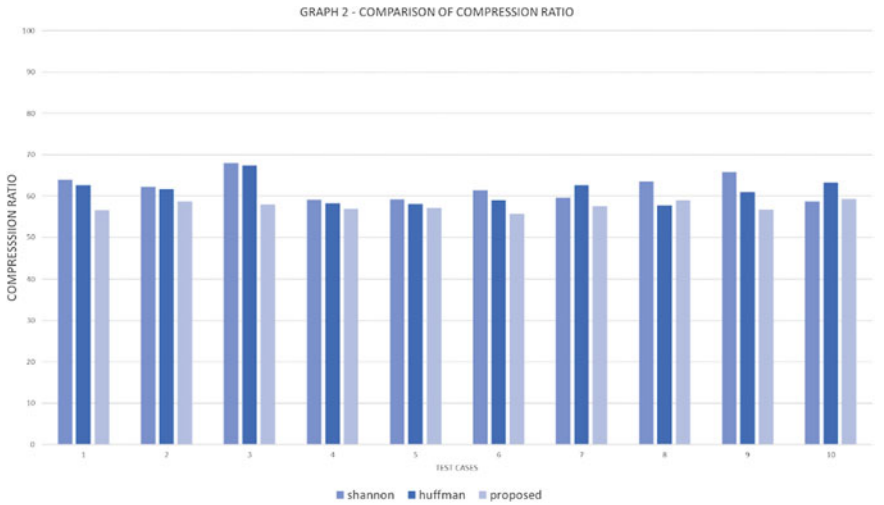


Fig. 5 Comparison of compression ratio

The compression ratio is calculated using Eq. (1), which shows that the fraction of the size of the compressed file and the size of the source file (Fig. 5). As the file size increases, a better compression ratio of our proposed algorithm is observed.

5 Conclusion and Future Scope

In this paper, the text data is converted into compressed data and the compressed data is converted back to its original form. Our proposed algorithm can be utilized as an integral component of any statistical lossless data compression algorithm, for example, Shannon-Fano coding, adaptive/static/dynamic Huffman coding, Arithmetic coding, a blend of these algorithms, or any revised form of them. The proposed algorithm has proved to have better compression than the Huffman algorithm and Shannon-Fano algorithm. In the future, more algorithms can be designed with a better compression ratio. The proposed algorithm can be used for reducing network congestion and conserving the storage capacity of various devices.

References

1. T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein, *Introduction to Algorithms, 3rd Rev* (Massachusetts Institute of Technology, London, 2009)
2. H. Al-Bahadili, S.M. Hussain, A bit-level text compression scheme based on the ACW algorithm. Int. J. Autom. Comput. **7**(1), 123–131 (2010). https://doi.org/10.1007/s11633-010-0123-6

3. H. Al-Bahadili, S.M. Hussain, An adaptive character word length algorithm for data compression. Comput. Math. Appl. **55**, 1250–1256 (2008)
4. S.R. Kodituwakku et al., Comparison of lossless data compression algorithms for text data. Indian J. Comput. Sci. Eng. **1**(4), 416–425
5. S. Tiwari, A. Lal, S. Agarwal, A. Kumar, A. Singh, N. Arora, Novel bit-level adaptive and asymmetric data compression technique. Int. J. Comput. Appl. **975**, 8887
6. A. Singh, S. Mahapatra, *Network-Based Applications of Multimedia Big Data Computing in IoT Environment* (2019), pp. 435–452. https://doi.org/10.1007/978-981-13-8759-3_17
7. R.M. Capocelli, R. Giancarlo, I.J. Taneja, Bounds on the redundancy of Huffman codes. IEEE Trans. Inf. Theory **32**(6), 854–857 (1986)
8. J.B. Connell, A Huffman-Shannon–Fano code. Proc. IEEE **61**(7), 1046–1047 (1973)
9. D. Cortesi, An effective text-compression algorithm. BYTE **7**(1), 397–403 (1982)
10. D.R. Mcintyre, M.A. Ano Pechura, Data compression using static Huffman code-decode tables. Commun. ACM **28**(6), 612- 616 (1985)
11. D.S. Parker, Conditions for the optimality of the Huffman algorithm. SIAM J. Comput. **9**(3), 470–489 (1980)
12. J.A. Storer, T.G. Szymanski, Data compression via textual substitution. J. ACM **29**(4), 928–951 (1982)
13. H. Tanaka, Data structure of Huffman codes and its application to efficient encoding and decoding. IEEE Trans. Inf. Theory **33**(1), 154–156 (1987)
14. H. Plantinga, An asymmetric, semi-adaptive text compression algorithm. in *Proceedings of IEEE Data Compression Conference* (1994)
15. S. Mahapatra, A. Singh, Application of IoT-based smart devices in health care using fog computing. in *Fog Data Analytics for IoT Applications*, ed. by S. Tanwar. Studies in Big Data, vol. 76 (Springer, Singapore, 2020)
16. D. Salomon, *Data Compression: The Complete Reference*. Springer Science & Business Media (2004)
17. R.K. Saini, S.L. Choudhary, A. Singh, A. Verma, Introduction to optimization algorithms–bio inspired. Inf. Secur. Optim. **189**, (2020)
18. L. Singh, A.K. Singh, P.K. Singh, Secure data hiding techniques: a survey. Multimed. Tools Appl. 15901–15921 (2020)
19. N. Agarwal, A.K. Singh, P.K. Singh, Survey of robust and imperceptible watermarking. Multimed. Tools Appl. 8603–8633 (2019)

Meta-Heuristic with Machine Learning-Based Smart e-Health System for Ambient Air Quality Monitoring

Pankaj Rahi, Sanjay P. Sood, and Rohit Bajaj

Abstract Health issues caused by air pollution are the primary concern nowadays. The air quality monitoring system aids in monitoring the level of air pollutants by measuring the concentration of a particular pollutant in the environment. A systematic study of the existing systems has been done to analyze the limitations in the current scenario of these systems. This paper proposes a smart e-health air quality monitoring system that utilizes a meta-heuristic Firefly algorithm as well as a Cat Swarm Optimization algorithm to efficiently optimize the selected features to give better results in the feature selection process. Further, these features are classified by using a Support Vector Machine which predicts the index level of the air quality and gives better precision and recall. The combination of all these algorithms has true positive and false positive rates as 99% and 88.4%, respectively while the recall rate is 89.6%. Also, the proposed work predicts a high air quality index based on the input dataset.

Keywords Air quality monitoring system · Firefly algorithm · Cat swarm optimization · Support vector machine · e-Health

1 Introduction

Air is the fundamental basis of a living being's existence as it provides fresh air to breathe. The atmosphere of Earth comprises numerous gases including dust particles.

P. Rahi (✉)
University Institute of Computing, Chandigarh University, Gharuan, Mohali, Punjab, India

S. P. Sood
Health Informatics Division, Centre for Development of Advanced Computing, Mohali, Punjab, India

R. Bajaj
Department of Computer Science and Engineering, Chandigarh University, Gharuan, Mohali, Punjab, India
e-mail: rohitbajaj.cse@cumail.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_38

Around 78% Nitrogen, 21% Oxygen, 0.9% Argon, 0.4% Carbon dioxide, and particulate matter together make up the Earth's atmosphere [1]. Air is contaminated by the induction of undesired particles or poisonous gases such as smoke, oxides of carbon, etc., that has a toxic effect. Air pollution results in various infections, allergies in humans, sometimes which are so serious that eventually lead to the death of individuals. It can cause respiratory problems to specific individuals if they are left untreated. Animals, as well as food crops, are adversely affected by this [2]. Air pollution is just not outside; the air inside the buildings can be contaminated as well. Some of the pollutants are discharged straight into the air while others are formed as a result of chemical reactions. The air pollution emissions are caused by various sources such as petroleum factories, dry cleaners, marine vessels, airplanes, highway vehicles, etc. In 2019, around 70 million tons of pollutants were emitted into the air in the US [3]. Environment and health are directly linked to each other. Good health can be defined as the strength to deal with numerous physical, mental, and, social difficulties [4, 5]. Human beings are constantly in touch with the environment in one way or the other. Therefore, bad air quality greatly influences human health. The continuously increasing air pollution has led to a large amount of skin and breathing problems. A large number of people across the world suffer from respiratory-related ailments such as asthma, cardiovascular diseases, and many others. So, the air quality needs to be at appropriate levels. This contamination can be controlled by measuring the Air Quality Index (AQI) that can help to analyze the current scenario of the atmosphere [6, 7]. Nowadays, air quality monitoring stations are being constructed in smart cities to perceive all the information associated with the quality of air [8]. These stations measure the number of toxic gases and pollutants present in the atmosphere at a specific time [9].

It is extremely significant for a nation to maintain its healthcare objectives if the air quality deteriorates. Therefore, an effective e-health system should be designed that can prevent aggravations of airborne infections. Such systems play a significant role in regulating airborne diseases and decreasing the rush in hospitals. There are certain limitations or gaps of the existing systems that need to be overcome to design an effective e-health system. A few of them are listed below:

1. Designing an e-health system that is capable of processing the information in real time is quite costly.
2. There is a need to apply machine learning and dimension reduction methods in combination to improve air quality monitoring [10].
3. The particular routines of Healthcare-APIs shall be derived for devising smart healthcare.

This paper proposes a Smart e-health system that employs a firefly algorithm and cat swarm hybrid optimization for optimization of the features plus, Support Vector Machines (SVM) to predict the quality of air. The deterministic system will have increased predictability and photo-realisticity of environmental properties. The organization of the paper is as follows: Introduction of the research topic is described in Sect. 1. The existing works in the particular research topic are explained in Sect. 2. The proposed work and methodology are described in Sect. 3. The graphical results

and outputs are elucidated in Sect. 4. Lastly, Sect. 5 consists of the conclusion and future work.

2 Related Work

In this section, some of the existing research works related to the topic are addressed. Most of the existing research works have proposed simple air quality monitoring systems using various techniques. The literature [2] proposed a system for observing air quality and pollution using an Arduino microcontroller unit. It consists of a sensor, a USB or power connection, an AC-to-DC converter, analog pins, and, a display. A value "V" is fetched from the sensor based on which the system predicts whether the air is fresh, polluted, or very polluted. This sensor-based system can be employed effectively. An air quality index modeling technique known as PANDA based on deep multitask learning (MTL) has been proposed for urban spaces [9]. It can solve spatially fine-grained AQI level evaluation plus forecasting tasks concurrently. The deep neural networks tend to learn the appropriate spatial and sequential patterns such as weather forecast, POI distribution, traffic, etc. It can also make a correlation between AQI and these representations. It is observed that PANDA outperforms the existing air quality modelling methods. Also, the literature [11] proposed a framework for monitoring air quality plus forecasting based on deep learning. It works in combination with domain-decomposition techniques so that the models can be deployed beyond the domain(s) on which it has been trained. Consistency constraints are proposed that train the surrogate models on small domains and then apply them to broader domains plus they allow including data external to the domain. In literature [7], a real-time monitoring system is depicted that utilizes low cost and power and comprises sensors for gas. Real-time Air Sense (RAS) consisting of AQI Collection Device (AQICD), an application used for aggregating the data, and cloud services such as Google Firebase that act as a real-time database is created. Finally, the air quality index value and corresponding air pollution status are displayed.

Another domain of the research topic deals with Indoor Air Quality (IAQ) systems. The literature [12] proposed a system to improve indoor air quality by introducing a context-aware non-intrusive scheme that can collect knowledge from its surroundings and adjust accordingly. It is responsible to reduce Sudden Infant Death Syndrome (SIDS) and is based on Big Data Context-aware Monitoring (BDCaM) to monitor medical data in great numbers [13]. It comprises an Ambient sensing system, Data collector and forwarder, Context aggregator, Context provider, Context management system, and, service provider. The system aids in reducing SIDS and developing safer surroundings for newborns. In the literature [14], an indoor Internet of Things (IoT) air quality monitoring system is presented that is low cost, battery-operated, and, portable. The battery life of the system is around 30 h. It can monitor the total amount of pollutants in the air as well as observe the temperature, humidity, and, illuminance in real time. A custom smartphone app named "Blynk" is also developed to work in low power modes and monitor the hourly average of the pollutants. Also, it makes

suggestions to increase ventilation and more. It works according to the standards of the Environmental Pollution Agency (EPA). The literature [15] focuses on IAQ to reduce the effects of poor air quality. The quality is measured using different sensors when the data is stored in the cloud. The system uses a Universal Asynchronous Receiver/Transmitter (UART) to control the interfacing of the device and computer. The MQ Telemetry Transport (MQTT) protocol is utilized for communication. The parameters like carbon dioxide, oxygen, dust particles, etc., are measured and an e-mail or SMS is sent to the users accordingly to notify them about the quality of air indoors.

The smart e-health system for air quality monitoring has a lot of applicability. The literature [16] focused on studying the effects of air pollutants based on air quality monitoring data and Baidu indices in Beijing. Concentrations of air pollutants, AQI values, mean level of wind, average temperature, and, average comparable humidity were measured every day.

AQI takes the maximum of six individual (SO_2, NO_2, PM_{10}, $PM_{2.5}$, O_3, and, CO) air quality index (IAQI) values as:

$$\text{IAQIi} = [\text{IAQIHi} - \text{IAQILi}/\text{BPHi} - \text{BPLi}]\,(\text{Ci} - \text{BPLi}) + \text{IAQILi}, \qquad (1)$$

where,

IAQIi: individual air quality index of i-th pollutant;

Ci: concentration of i-th pollutant;

BPHi: higher limit of concentration;

BPLi: Lower limit of concentration limits;

IAQIHi and IAQLi: individual air quality indices corresponding to BPHi & BPLi.

The relative risk's for respiratory, cardiovascular, and, cerebrovascular diseases are obtained which shows O_3 and NO_2 caused higher risks for these diseases comparative to others. Another research work [17] contributes to detecting abnormal measurements in the ozone layer using deep learning. It applies a combination of Deep Belief Networks (DBN) and a one-class SVM (OCSVM). OCSVM tends to separate normal from abnormal datasets, i.e., it is designed by using data only from one class. The system is tested on real datasets from Isere, France. The fault detection methodology includes two phases. Phase 1 is the Modelling phase that includes normal data collection, greedy layer-wise training of a DBN, fine-tune the DBN parameters, feature extraction from the last layer, training of OCSVM using output data, and, finding the maximal margin hyperplane while Phase 2 is the Anomaly detection phase comprising of acquiring and normalizing new data, generating data features using DBN model, evaluating feature vectors using OCSVM model, and finally declaring the anomaly. Results prove that DBN-OCSVM methodology functions reasonably well in detecting real ozone abnormalities. The research work [18] represents that the feature selection is carried out using Principle Component Analysis (PCA) and Support Vector Regression(SVR) with Radial Basis Function(RBF) kernel is deployed for accurately prediction of hourly density of particulates, concentration of gases/pollutants for further defining the Air Quality Index.The literature

[19] designed a system for the prediction and prevention of an airborne disease outbreak known as MERS-CoV. The system comprises data collection for combining information from the users utilizing sensors that are worn in the body, information granulation, BBN (Bayesian Belief Networks)-based initial analysis classifies the users into possibly infected or uninfected by analyzing the symptoms registered by users and sensors, assessment of risks based on GPS, and, communication as well as sharing of information. If the user is found to be infected after further tests at the hospital, an alert is sent to everyone associated with that person as well as the government agencies. Further, Google maps are updated consequently. Results show 80% accuracy in classifying accurately along with GPS rerouting. The research work [20] defined the expert system for monitoring the air quality and explained the optimization of feature selection process using Firefly algorithm with Support Vector Machine (SVM) for classification of pollutants information for further defining the AQI. Some of the existing works in the field of air quality monitoring as shown in Table 1.

3 Research Method

The background analysis and gaps in the existing works led to the development of a new smart e-health system. The objectives, basics of air quality monitoring, system design, and, various techniques used in the system are described in the following sections.

3.1 Objectives

The central aim of the work is to design an effective and smart e-health system. The proposed work completes by accomplishing the following objectives:

(1) To study and analyze the existing expert systems of air quality monitoring.
(2) To design an algorithm for monitoring the ambient air quality.
(3) To implement the predictive modelling for Smart e-Health system.

3.2 System Design

The proposed smart e-health system is designed in such a way so that it can detect the air quality accurately. Initially, the data consisting of environmental pollutants is fed to the system; the next step is to analyze what pollutants are present in the system. Further, an algorithm is designed to check the air quality by measuring the number of harmful air pollutants, and, finally, predicting whether the air is polluted or not. Figure 1 shows the general design of the proposed system.

Table 1 Existing air quality monitoring systems

Author	Approach used	Dataset	Outcomes
Dhingra et al. [21]	• Three phase pollution monitoring system: phase 1 detects concentration of air pollutants, phase 2 develops android app, and phase 3 predicts air quality level • Consists of Arduino IDE module, sensors for gas, and, a Wifi component • IoT-Mobair: an android application to access air quality data from cloud	Generated by gas sensors	• App has the following features: Indices of air quality for specific city with real-time computation, • Daily forecasts and map generation of air quality
Laskar et al. [22]	• Framework for air quality index constructed using Bayesian neural network • Hardware uses Atmel 2560 based platform while pollutants measured using solid state sensors	Kolkata Municipal Corporation study in winter season	• Provides real-time air quality data about a location through wireless communication • User gets information related to health and neighborhood due to pollutants in ambient air
Ma et al. [23]	• Methodology based on DL, transferred bi-directional long short-term memory (TL-BLSTM) model • Utilizes this methodology to grasp long-term reliability of pollutants and applies transfer learning	Case study in Guangdong, China	• Combines deep and transfer learning to predict quality of air at substantial temporal resolutions • Aids the government organizations to generate and examine correct predictions • Effectively decrease prediction error of BLSTM for $PM_{2.5}$ • These algorithms outperform other ML and DL models

(continued)

3.3 Methodology

The system consists of a GUI comprising several list boxes to input the uploaded data. There are pushbuttons on the GUI for uploading data of varying concentrations of gases and performing the classification. The proposed research work utilizes the

Table 1 (continued)

Author	Approach used	Dataset	Outcomes
Zhang et al. [24]	• Model of LightGBM is used to determine $PM_{2.5}$ concentration in next 24 h • Sliding window method used to improve training data amount • Resolves issue of processing large-scale data and inputs the forecasting data for prediction of air quality	Data collected at 35 monitoring stations in Beijing	• Deeply explores the high-dimensional and statistical features based on analysis done on big data • It surpasses the other models • Performance is increased by including forecasting data

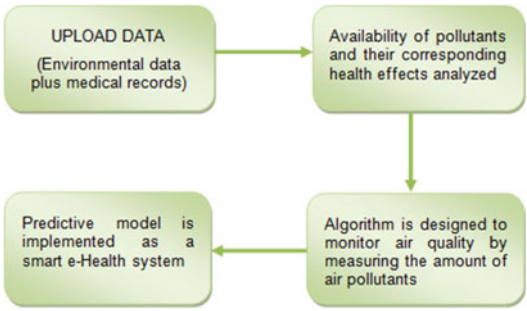


Fig. 1 General design of the system

Firefly and Cat Swarm Hybrid optimization techniques for optimizing the features or instance selection. Further, SVM is employed to classify the air quality of the surroundings. The air quality parameters and their health effects are clearly defined in the ambient air quality monitoring program known as National Air Quality Monitoring Programme (NAMP) implemented in India during 2014. Various agencies have set and established the rules for measuring the Air Quality Index for defining the pollution levels and their effects. Table 2 shows the number of various air pollutants such as PM_{10}, NO_2, $PM_{2.5}$, $S0_2$, CO, Pb_{24}, and, O_3 in the air and categorizes the air according to its quality. AQI is one of the significant means used for government organizations for reporting the overall location-wise quality of air and their trends based on already defined parameters of a specific standard. In India, the CPCB standard is used for calculating environmental pollution and air quality index.

The dataset and techniques used in the proposed work are explained in the sections below.

Table 2 AQI categories according to CPCB standard [25]

AQI category (range)↓	PM_{10} 24-h	$PM_{2.5}$ 24-h	NO_2 24-h	O_3 8-h	CO 8-h (mg/m3)	NH_3 24-h	Pb 24-h	Possible impact on health
Good (0–50)	0–50	0–30	0–40	0–50	0–1.0	0–200	0–0.5	Minimal effect
Satisfactory (51–100)	51–100	31–60	41–80	51–100	1.1–2.0	201–400	0.5 –1.0	Minor discomfort in breathing to sensitive people
Moderately polluted (101–200)	101–250	61–90	81–180	101–168	2.1–10	401–800	1.1–2.0	Discomfort in breathing amongst the people with asthma, lungs, and heart ailments
Poor (201–300)	251–350	91–120	181–280	169–208	10–17	801–1200	2.1–3.0	Breathing difficulty to people on prolonged exposure
Very poor (301–400)	351–430	121–250	281–400	209–748*	17–34	1200–1800	3.1–3.5	Respiratory illness on prolonged exposure
Severe (401–500)	430+	250 +	400+	748+	34+	1800+	3.5+	Affects healthy people and seriously affects those with existing diseases

3.3.1 Dataset

The dataset used in the proposed work is from air quality monitoring equipment that has multiple sensors as stated in [26]. A total of balanced 9358 instances in dataset responses from an array of 5 metal oxide chemical sensors were installed. The device was placed on the field in a notable contaminated area at the road level within an Italian city. The data was recorded from March 2004 to February 2005 that represents the longest readily obtainable records of field-deployed air quality chemical sensor device responses. The ground truth has hourly average concentrations for Carbon Dioxide, Benzene, Nitrogen Dioxide, Non-metallic Hydrocarbons, and Total Nitrogen Oxide that are given by the certified analyzer of ordered reference. The missing values are tagged with a −200 value.

3.3.2 Algorithms Used

The algorithms used in the proposed work are as explained below:

(A) Firefly Algorithm

This algorithm [27, 28] is motivated by the ability of the fireflies to transmit light for communication with other fireflies. The main motive of this kind of communication is either reproduction or search for food. This algorithm is used for optimization problems where the optimization function is associated with the intensity of the emitted light. FA has witnessed many applications in finding solutions for optimization problems. FA can be utilized in three forms, namely, standard FA, improved versions of FA, and hybridization of FA with other optimization techniques. The workflow and implementation [29, 30] of FA is shown in Fig. 2 and Table 3, respectively.

(B) Support Vector Machines

Support Vector Machines (SVMs) are supervised learning models that are used for classification problems owing to their ability to separate datasets by the most reliable hyperplane [31]. As the effectiveness of classification is independent of large feature space, it is quite popular as a classifier. SVM simulations imply instances as a point in space that is mapped by a method that differentiates one type's instances from others through a gap.

As described in [32], Let n be the dimensional inputs X_i, where, $i = 1, 2, 3,..., m$, and m is the number of individuals belonging to classes 1 and 2. The related labels are $Yi = 1$ and $Yi = -1$. This data is divided by a linear hyperplane that can be written as

$$F(x) = 0 \tag{2}$$

This can be defined by the following equation:

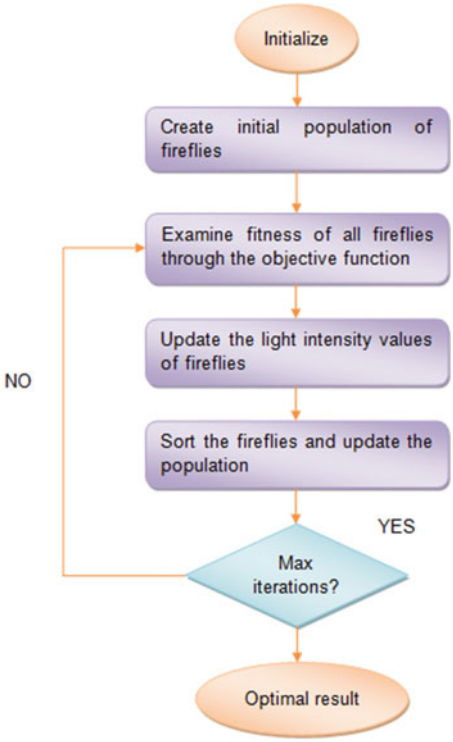


Fig.2 Workflow and implementation of the firefly algorithm [2]

Table 3 Firefly algorithm [28, 30]

Step 1	Generate the Objective function which is to be minimized: O(t) , t = (t1 , t2 , . . . , td)
Step 2	Initial population generation of the swarm particles such that F_p (p = 1 , 2 , … , s)
Step 3	Evalaution of the intensity of the particles which is assoaciated to O(t) i.e. Objective function
Step 4	Define absorption coefficient γ
	While (t < MaxGeneration)
	for i = 1 : n (all n fireflies)
	for j = 1 : n (n fireflies)
	if (I j > I i)
	Differ the attractiveness with distance r via $\exp^{(-\gamma r)}$
	firefly i is moved towards j;
	Assess new solutions and update light intensity;
	end if
	end for j
	end for i
	Sort the particles and evalaute the best fittest solution.
	end while
	end

$$f(x) = WTx + b = n\sum k = 1Wkxk + b \tag{3}$$

where, W is n dimensional vector and b is scalar. These variables decide the hyperplane position [31]. The constraints are $X_i \geq 1$ if $Y_i \geq 1$ and $f(X_i) \geq -1$ if $Y_i \geq$

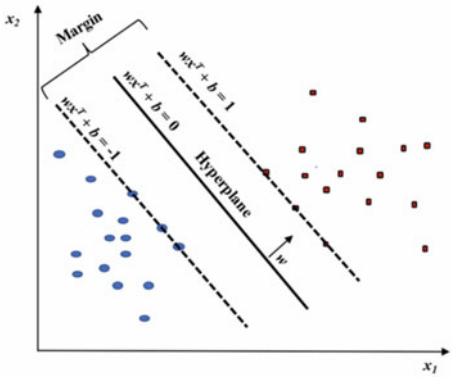


Fig. 3 Linear Support Vector Machine [33]

−1. They tend to analyze data as well as recognize patterns. A linear SVM is shown in Fig. 3.

(C) Cat Swarm Optimization

This algorithm is an intellectual algorithm that is based on the natural behavior of cats and is employed for optimizing the resources. It is designed based on cat behavior that comprises of two behaviors—the tracing mode and the seeking mode. In the CSO algorithm, originally, the number of cats that are to be used is selected then the cats are processed into the algorithm to resolve the issues [32, 34]. Every cat has its own position comprised of M dimensions, fitness value, and, a flag to identify whether the cat is in a seeking mode or tracing mode. The final solution is the best location for one of the cats as shown in Fig. 4. Through observation of cat behavior, it can be perceived that most of the time is spent by the cats resting while moving slowly from one position to other. So, to merge the two modes in CSO, a mixture ratio needs to be defined that must be a small value.

Modes of CSO

There are two modes of CSO described as follows:

- *The Seeking mode:*

It is used to replicate the cat situation that is resting, looking around, and, searching the following position. Four-dimensional factors are elucidated [36]:

(1) SMP-It defines the size of seeking memory for each cat. The cat may choose a point from the memory pool per the rules.
(2) SRD-Seeking a range of selected dimensions is used to announce the ratio of mutation for selected dimensions. As stated in [37], if a dimension is chosen to vary, the difference between the new and old values will be out of range that is defined by this parameter.

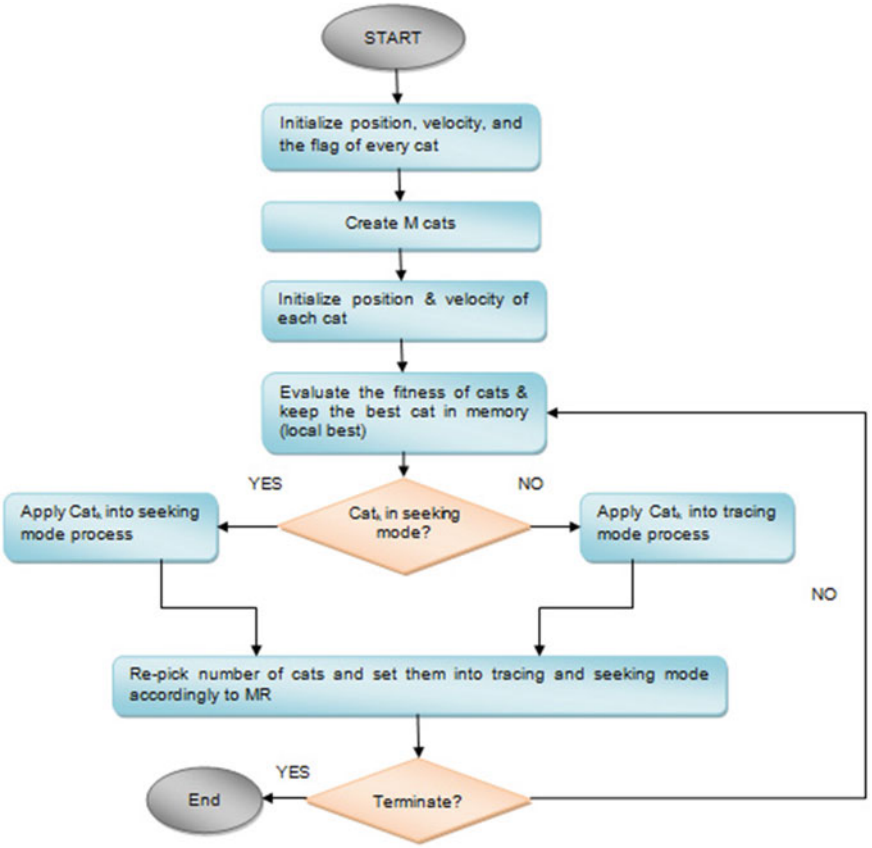


Fig. 4 CSO flowchart [3, 35]

(3) CDC-Counts of dimension to Change specifies that how many dimensions are varied.
(4) SPC-It determines whether the current location of the cat will be one of the candidates that can be moved to. It is a Boolean value.

It is shown in the literature [38] that the next position of the cat is based on the probability as shown:

$$Pi = |FSi - FSb| / FSmax - FSmin, \tag{4}$$

where, 0 < I < j and FS = Fitness value.

- *The Tracing mode:*

It is another way of modeling the state of a cat in search of prey. Once the cat switches into the tracing mode, it tends to move in every dimension as per its velocities.

4 Results and Analysis

4.1 Data Uploading

Figure 5 depicts the user panel that comprises list boxes as well as pushbuttons. When the push button "UPLOAD" is clicked, the list boxes are fed information consisting of concentrations of numerous air pollutants and other factors such as carbon monoxide, tin oxide, benzene chloride, particulate matter, temperature, etc., from the dataset.

Figure 6 displays the procedure for selecting the dataset that is fed to the user panel for further process. A folder is opened from which the dataset file is chosen that gets uploaded into the GUI. A dialog box appears with "Data loaded successfully."

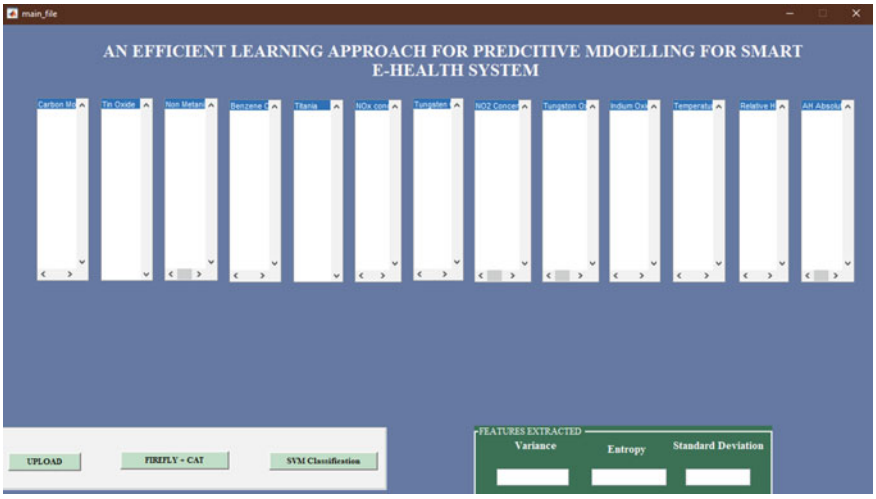


Fig. 5 Proposed panel

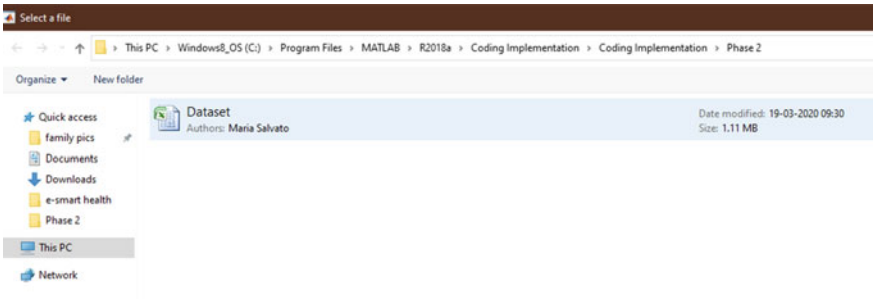


Fig. 6 Dataset selection

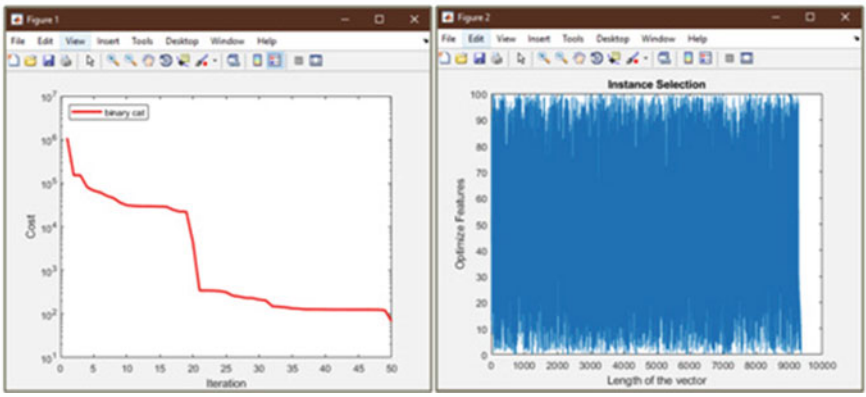


Fig. 7 CSO algorithm iterations and instance selection

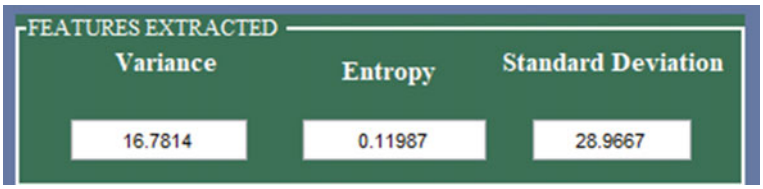


Fig. 8 Extracted features

4.2 Implementation of Firefly and CSO Algorithm

When the "FIREFLY + CAT" push button is pressed, 50 iterations are performed to get the best cost cat as shown in Fig. 7. These algorithms tend to optimize the solution as a dialog box with "Optimization Done Successfully" message appears.

4.3 Feature Extraction

As the optimization is done using Firefly and Cat Swarm Optimization algorithm, various feature values are extracted from the data that are variance, entropy, and, standard deviation as shown in Fig. 8.

4.4 Instance Selection

It can be explained as the selection of a subset from the population while keeping the underlying distribution intact so that the selected data signifies the overall population

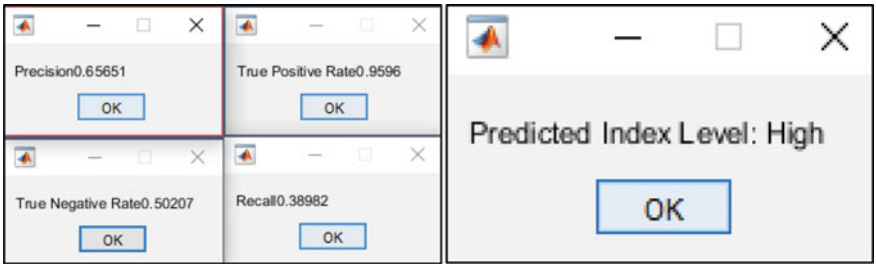


Fig. 9 Performance results using Firefly algorithm

features. The process of instance selection in the proposed system is executed as shown in Fig. 7. The training and testing are done based on this process of instance selection.

4.5 Result Metrics

Following terms are required to be understood and defined before observing the results as stated in literature [3]:

- *True Positive*—It is defined as the number of instances that are correctly predicted as required. It is usually denoted as TP.
- *True Negative*—The number of instances that are correctly predicted as not required are called True Negatives. It can be represented as TN.
- *False Positive*—The number of instances that are predicted as required but are incorrect. It is denoted as FP.
- *False Negative*—It can be defined as incorrectly predicted instances as not required. It can be represented as FN.

The result metrics used in the proposed system are described and screenshots are shown in Fig. 9.

4.5.1 Precision

It can be measured as the true positives divided by the total of true positives and false positives.

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{5}$$

4.5.2 Recall

It is defined as true positives divided by the summation of true positive plus false negative as shown below:

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{6}$$

4.5.3 Predicted Index Level

It is the main result metric as it defines whether the quality of air is good or poor. A High index level means the presence of harmful gases is higher in the atmosphere and the air quality is not fit while low index level means that the level of harmful gases present in the air is low and the quality of air is fit.

Figure 9 shows that the predictive index level results are classified as high which shows the presence of dangerous gases in the air which can be harmful for the humans and the important protective measures should be taken.

Figure 10 shows the performance evaluation in terms of high precision, high recall and high true positive rate, and high true negative rate using cat swarm optimization. From the above performance analysis, it can be seen that the results are coming better in terms of performance evaluations using firefly and cat swarm optimization hybridizations which is one of the efficient scenario to achieve high classification error rates.

Table 4 shows that using Firefly algorithm with Cat Swarm Optimization plus Support Vector Machine yields better results in terms of precision and recall as

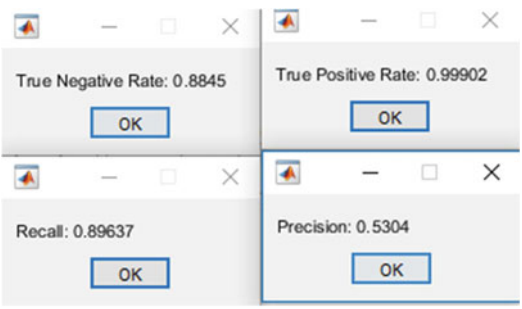


Fig. 10 Performance evaluation using Firefly + Cat Swarm hybridization

Table 4 Performance evaluation using different methods

Parameters	PCR SVR-RBF	(Firefly + CAT) + SVM
Precision	0.67	0.53
Recall	0.42	0.89
True positive rate	0.93	0.99
False positive rate	0.89	0.88

compared to using only Firefly algorithm with SVM as well as the results of algorithm of PCR with SVR-RBF algorithm. Hence, the proposed air quality monitoring system is more precise and accurate. This further triggers the areas wise categorizations and notifications/alerts to citizens/population residing in the close proximity, Controlling Bodies (Municipalities, Pollution Control boards, AirCleaning groups and even can alert Healthcare Providers/Rapid Response Teams) for initiating timely actions for reducing the exposure of bad air quality which further is helpfull in reducing the long or short-term health hazards. This is could also be utilized for reducing the burden of diseases by initiating timely geocentric-preventions, which further reduce the diseases burden and treatment cost at a large scale.

5 Conclusion

Air is the fundamental basis of a living being's existence. The atmosphere includes numerous gases along with several kinds of pollutants that contaminate the air resulting in severe health problems in humans. This paper largely contributes to designing a smart e-health system for monitoring the quality of the air. The current limitations have motivated us to work in this area. The study aims to analyze existing systems of air quality monitoring and then design an algorithm for observing the ambient air quality. The work proposes a Firefly algorithm and Cat Swarm optimization algorithm to perform feature selection. Further, classifying these features using a Support Vector Machine as explained in the paper. For evaluation of the proposed work, the results are evaluated in MATLAB software in terms of precision, recall, and, predicted index level. A high predicted index level index indicates that the quality of air is poor and is unsafe for humans as it can cause severe health as well as breathing issues as indicated in Table 2. The proposed work aid in improving the smart healthcare derived specific routines of healthcare-APIs.

The main aim for accumulating the data plus improving the air quality monitoring system is to access the concentration of pollutants in the air and inform the residents about the air quality. The data used in this research work help in analyzing the effects of air pollution on health that can be used by the governmental organization in remedying the diseases and Pollution Control Boards and National air quality department in controlling the concentration of pollutants in the air. Future work may include real-time monitoring apps that can instantly generate alerts based on basis of the air quality at any point of time. Also, define HeathyAir or e-CleanAir as cloud service of Smart e-Health System.

Acknowledgments The authors acknowledge the contributions of the University Institute of Computing, Chandigarh University, Mohali for undertaking this research.

Disclosure Statement
Not Applicable.

Funding The research is not funded by any organization.

References

1. Y. Su, Prediction of air quality based on gradient boosting machine method. in *2020 International Conference on Big Data and Informatization Education (ICBDIE)*, (2020). IEEE, pp. 395–397
2. M.A.A. Ahasan, S. Roy, A.H.M. Saim, R. Akter, M.Z. Hossain, Arduino-based real time air quality and pollution monitoring system. Int. J. Innov. Res. Comput. Sci. Technol. **6**(4), 81–86 (2018)
3. D. Chandirasekaran, T. Jayabarathi, Cat swarm algorithm in wireless sensor networks for optimized cluster head selection: a real time approach. Clust. Comput. **22**(S5), 11351–11361 (2017)
4. Y. Kumar, K. Sood, S. Kaul, R. Vasuja, Big Data Analytics and Its Benefits in Healthcare. in *Studies in Big Data Big Data Analytics in Healthcare*. IEEE, pp. 3–21
5. Y. Kumar, M. Mahajan, 5. Recent advancement of machine learning and deep learning in the field of healthcare system. in *Computational Intelligence for Machine Learning and Healthcare Informatics*, (2020). IEEE, pp. 77–98
6. M. Marć, M. Tobiszewski, B. Zabiegała, M.D.L. Guardia, J. Namieśnik, Current air quality analytics and monitoring: a review. Anal. Chim. Acta **853**, 116–126 (2015)
7. S.R. Enigella, H. Shahnasser, Real time air quality monitoring. in *2018 10th International Conference on Knowledge and Smart Technology (KST)*, (2018)
8. G. Marques, J. Saini, M. Dutta et al., Indoor air quality monitoring systems for enhanced living environments: a review toward sustainable smart cities. Sustainability **12**, 4024 (2020)
9. L. Chen, Y. Ding, D. Lyu, X. Liu, H. Long, Deep multi-task learning based urban air quality index modelling. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **3**(1), 1–17 (2019)
10. S. Tanwar, Q. Bhatia, P. Patel, A. Kumari, P.K. Singh, W. Hong, Machine learning adoption in blockchain-based smart applications: the challenges, and a way forward. IEEE Access **8**, 474–488 (2020)
11. P. Hähnel, J. Mareček, J. Monteil, F. Odonncha, Using deep learning to extend the range of air pollution monitoring and forecasting. J. Comput. Phys. **408**, 1–13 (2020)
12. D.D.L. Iglesia, J.D. Paz, G.V. González, A. Barriuso, J. Bajo, A context-aware indoor air quality system for sudden infant death syndrome prevention. Sensors **18**(3), 2–22 (2018)
13. Y. Kumar, K. Kaur, G. Singh, Machine Learning Aspects and its Applications Towards Different Research Areas. in *2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM)*, (2020). IEEE, pp. 150–156
14. S. Esfahani, P. Rollins, J.P. Specht, M. Cole, J.W. Gardner, Smart city battery operated IoT based indoor air quality monitoring system. in *2020 IEEE Sensors*, (2020). IEEE, pp. 1–4
15. V. Ladekar, R. Daruwala, Indoor air quality monitoring on AWS using MQTT protocol. in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, (2019). IEEE, pp. 1–6
16. S. Zhong, Z. Yu, W. Zhu, Study of the effects of air pollutants on human health based on baidu indices of disease symptoms and air quality monitoring data in Beijing, China. Int. J. Environ. Res. Public Health **16**(6), 1–19 (2019)
17. F. Harrou, A. Dairi, Y. Sun, F. Kadri, Detecting abnormal ozone measurements with a deep learning-based strategy. IEEE Sens. J. **18**(17), 7222–7232 (2018)
18. M. Castelli, F.M. Clemente, A. Popovič, S. Silva, L. Vanneschi, A machine learning approach to predict air quality in California. Complexity **2020**, 1–23 (2020). https://doi.org/10.1155/2020/8049504
19. R. Sandhu, S.K. Sood, G. Kaur, An intelligent system for predicting and preventing MERS-CoV infection outbreak. J. Supercomput. **72**(8), 3033–3056 (2015)
20. P. Rahi, S.P. Sood, R. Bajaj, Y. Kumar, Air quality monitoring for Smart eHealth system using firefly optimization and support vector machine. Int. J. Inf. Technol. **13**(5), 1847–1859 (2021). https://doi.org/10.1007/s41870-021-00778-9
21. S. Dhingra, R.B. Madda, A.H. Gandomi, R. Patan, M. Daneshmand, Internet of Things mobile-air pollution monitoring system (IoT-mobair). IEEE Internet Things J. **6**(3), 5577–5584 (2019)

22. M.R. Laskar, P.K. Sen, S.K.D. Mandal, An IoT-based e-health system integrated with wireless sensor network and air pollution index. in *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, (2019). IEEE, pp. 1–5
23. J. Ma, J.C. Cheng, C. Lin, Y. Tan, J. Zhang, Improving air quality prediction accuracy at larger temporal resolutions using deep learning and transfer learning techniques. Atmos. Environ. **214**, 1–10 (2019)
24. Y. Zhang, Y. Wang, M. Gao, Q. Ma, J. Zhao, R. Zhang, Q. Wang, L. Huang, A predictive data feature exploration-based air quality prediction approach. IEEE Access **7**, 30732–30743 (2019)
25. EPA, air quality—national summary, (2020). https://www.epa.gov/air-trends/air-quality-national-summary. [Online]
26. D. Dua, C. Graff, *UCI Machine Learning Repository* (2019). https://archive.ics.uci.edu/ml. [Online]
27. F. Wahid, R. Ghazali, L.H. Ismail, Improved Firefly algorithm based on genetic algorithm operators for energy efficiency in smart buildings. Arab. J. Sci. Eng. **44**(4), 4027–4047 (2019)
28. K. Kaur, Y. Kumar, Swarm Intelligence and its applications towards Various Computing: A Systematic Review. in *2020 International Conference on Intelligent Engineering and Management (ICIEM)*, (2020). IEEE, pp. 57–62
29. R. Nand, P. Sharma, Iteration split with Firefly algorithm and genetic algorithm to solve multi-dimensional knapsack problems. in *2019 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, (2019)
30. S. Banerjee, D. Mandal, Array pattern optimization for a steerable circular isotropic antenna array using the firefly algorithm. J. Comput. Electron. **16**(3), 952–976 (2017)
31. A. Saxena, S. Shekhawat, Ambient air quality classification by Grey Wolf optimizer based support vector machine. J. Environ. Public Health **2017**, 1–12 (2017)
32. S. Goyal, S. Bhushan, Y. Kumar, A.U.H.S. Rana, M.R. Bhutta, M.F. Ijaz, Y. Son, An optimized framework for energy-resource allocation in a cloud environment based on the whale optimization algorithm. *Sensors* **21**(5), 2021
33. S. Huang, N. Cai, P.P. Pacheco, S. Narrandes, Y. Wang, W. Xu, Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genomics Proteomics **15**(1), 41–51 (2018)
34. S. Kaul, Y. Kumar, Nature-inspired metaheuristic algorithms for constraint handling: challenges, issues, and research perspective. in *Constraint Handling in Metaheuristics and Applications* (2021), pp. 55–80
35. B. Crawford, R. Soto, N. Berríos, F. Johnson, F. Paredes, C. Castro, E. Norero, A binary cat swarm optimization algorithm for the non-unicost set covering problem. Math. Probl. Eng. **2015**, 1–8 (2015)
36. M. Chen, Y. Hao, K. Hwang, L. Wang, L. Wang, Disease prediction by machine learning over big data from healthcare communities. IEEE Access **5**, 8869–8879 (2017)
37. Q. Yang, G. Webb, PRICAI 2006: trends in artificial intelligencelspringer link (2006). https://link.springer.com/book/. https://doi.org/10.1007/978-3-540-36668-3. [Online]
38. A.M. Ahmed, T.A. Rashid, S.A. Saeed, *Cat Swarm Optimization Algorithm—A Survey and Performance Evaluation* (2020)

Smart eHealth System for Pervasive Healthcare

Pankaj Rahi, Sanjay P. Sood, and Sanjay K. Sharma

Abstract The recent technological developments in Information and Communication Technology (ICT) is offering multiple tools and techniques for upholding the large set of data pertaining to the pollutants availability in the air, which could further be utilized for devising the mechanism for empowering the pervasive healthcare for reducing the burden of diseases in the larger set of population. Cloud computing, Internet of Medical Things (IoEMT), and wireless networks have revitalized the new era and methodologies which are effectively utilized for handling the challenges of healthcare in terms of Quality of Healthcare Service (QoHS). It is also helpful in improving the information-sharing mechanism for preventive healthcare, integration and management of Decision Support Systems. The aim of this research is to define the framework for collating the real-time data of air quality parameters and analyzing the same for improving preventive health analytics as a cloud computing service. This platform of technology further derives the healthy-mapping or re-routing assisted prevention as an early aspect of cure from the airborne diseases. The substantial increase in the diseases caused by the air pollutants in the air and with increasing population the vehicles density on road has also increases, giving us the reason for proposing the mechanism to address the concern with technological capabilities. The proposed architecture in this research may also be helpful in improving user-centric, low-cost modeling of Smart eHealth System which is beneficial in reducing the morbidity and mortalities occurring with the exposure of pollutants available in the local environment.

Keywords Smart eHealth · Cloud computing · Pervasive healthcare · Green computing · Predictive healthcare · Internet of medical things (IoMT)

P. Rahi (✉)
University Institute of Computing, Chandigarh University, Gharuan, Mohali, Punjab, India

S. P. Sood
Health Informatics Division, Centre for Development of Advanced Computing, Mohali, Punjab, India

S. K. Sharma
University of Hydrabad, Hydrabad, Tengana, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_39

1 Introduction

Environment is considered to play an extensive role in the spread and control of many disorders such as asthma. In 2014, it was estimated that 92% of the world's population lived in areas where the air quality standards defined by WHO were not met. This further led to the occurrence of approximately million premature deaths in both rural and urban areas of the globe in 2012. Out of which 88 percent of the premature deaths occurred in the underdevelopment countries specifically among the individuals of low or middle-income groups. Large numbers of deaths belong to Western Pacific and South-East Asia regions. The "disease" is also defined as the abnormal condition and response caused by human body-like causes pain, dysfunctioning of organ, mental stress, death of person, or may cause similar problems for those living beings who are in contact with the affected person. Air quality monitoring is the core thrust area for preventing the occurrences of many diseases in the community and urban households in the developing countries. Evidences of the tightly coupled connections between the ecosystem and health are strongly established and thoroughly elaborated in the medical and environmental health sciences. The World Health Organization (WHO) has also determined the environmental factors as triggers which are accountable for almost 24% of the world's burden of diseases and also 23% of all deaths [1] occurred.

Seventeen percent of deaths are in the developed regions which have defined the environmental components as the major contributor and accountable for happenings. Also, the environment plays a significant role in the occurrence and cure of chronic diseases such as lung cancer (30%). The adverse environmental health impacts on living beings have been well acknowledged by Canadian authorities. The environmental authority of Canada also defines that the "asthma, lung disease, cardiovascular syndrome, allergies and many other human health issues have been linked to poor quality of air [2]."

Environmental factors and air quality (outdoor or indoor) are major components plays vital role in spread and cure of respiratory or communicable diseases. Allergy and Asthma are two major diseases largely linked with environmental factors and air quality. Approximately, 20–30% of total population of India have allergic diseases, out of which 15% are having asthma or asthma-like symtoms (WHO-Research 2009). The biomedical and health sector are the major contributors of data.

Since the healthcare Information Systems are linked with human-centric care with further outcomes as diseases cure or mortality, hence mHealth, eHealthcare, Digital Healthcare technologies, Electronic Health Records (EHRs), electronic Personal Health Records (ePHR) are advantageous for tackling the emergencies or enforcing the timely responses for saving the lives. In healthcare systems, the patients connected information is used for continuous real-time monitoring of disease or illnesses especially among the children, old aged people, and patients of chronic diseases even during home by using the expert Smart-Healthcare architecture of mHealth System. This help to evaluate the symptomatic assertions with clinical trials are sufficient, provide the medical advice or generates the alerts to the patients, family physician, government health provides for timely inteventions for controlling of the disease or

prompt patient for immediate expert-medical consultations or spreading of Communicable disease and strengthen the Public Health Surveillance System. The combination of Internet of Medical Things in consonance with Hospital Information System and Environmental oriented ambient air analytics will increase the productivity of this architecture. In current scenario, ICT is considering the promising technology utilized for improving and streamlining the Indian Healthcare Delivery System. Timely availability and delivery of healthcare service to the patients in India is still the challenge due to shortfall of human and physical/infrastructural resources. Despite recent developments, the maintenance of Central Repository for Patient—illness or other clinical trial or less availability of Electronic Medical Records for policy decisions still requires more efforts and computing integrations. Since the diagnosis of any illness is linked with the past history of diseases and care of the patient and even with other social determinants as well environmental factors. Hence the integration, virtualization of these factors using technology along with its linking with electronic medical records is essentially needed for delivering or controlling the quality of Healthcare services or diseases vulnerability in India. Hence the linking of Climate change, Weather factors, air quality, socioeconomic, and health datasets using Cloud Computing and Big Data-enabled platforms will provide new insights to human healthcare and wellness, and pushes the development of decision support/alert tools which are quite beneficial for providing healthy living environment for all living beings. There are multiple facts which are demonstrating the occurrences of airborne allergies in consonance with environmental phenomena as defined in Fig. 1.

There are significant spatiotemporal lags between change in air quality and health outcomes like occurrence of allergy and asthma diseases. The real-time air quality monitoring is a critical public health concern and essential element for improving the quality of life as human passes 90% of their total time in indoor activities at home or workplace. The concentration of CO_2 and Particulate Matter (PM_{10} and $PM_{2.5}$) is relatively higher in closed areas which could generate prolonged serious health issues [3].

The aim of this research is to define the algorithm to combine multiple evidence sources using Cloud Computing and Big Data Analytics for increased sensitivity at manageable alert rates to achieve effective Health Informatics and Bio-surveillance The Innovative Healthcare framework [4, 5] suggested the innovation having thrust on three crore areas, i.e., how the patient is seen, how the patient is heard, how the patient's needs are met during the emergencies or disaster or biomedical threats which are still needing lots of ICT integrations for intelligent electronic Personal Health Records (ePHR), electronic Community Health Records (eCHR), and EMR or effective Health Informatics.

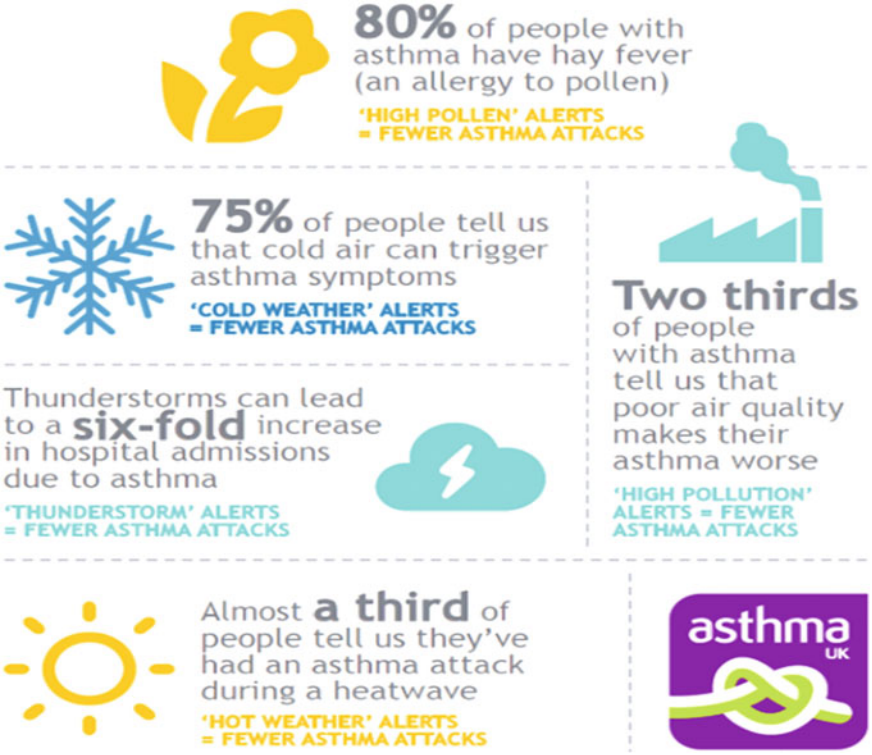


Fig. 1 Environmental factor make difference in case of Asthma (*Source* Asthma UK18 Mansell Street London https://www.asthma.org.uk)

2 Motivation and Objectives

The local environment (indoor and outdoor) have the remarkable manifestation in the existence as well as treatment of various chronical diseases like lung cancer (30%). The adverse environmental health impacts on living beings have been well acknowledged by Canadian authorities. The environmental authority of Canada has also defines that "asthma, lung diseases, cardiovascular syndrome, allergies as human health issues linked to the poor quality of air [6]." The pollutants availability in the air are the most possible risk factor for causing asthma and people having comorbidities like heart diseases, lung diseases are at higher risk group, may be affected by pollution even on moderate or low pollution levels. Moreover, the children and young adults already having asthma may have faster breathing rates and the lung development of such cases may be continuous, can take longer time than the common period.

The eHealth system of real-time air quality monitoring helps in preventing the exacerbations of airborne diseases and also helps in addressing the community health

vulnerabilities. The main aim of this research paper is to define the technology linked architecture for meeting the following objectives:

- To design the conceptual architecture of intelligent air quality monitoring system for leveraging the smart predictive health analytics of air as Cloud service.
- To define the intelligent system for allergic and asthmatic patients for Pervasive Healthcare.
- To define the Smart eHealth System for Airborne diseases vulnerabilities with advantages and limitations of ICT.
- To ensure geolocational, real-time indicative parameters of prevention of health hazards for the society at large.

3 Related Work

According to the reports of World Air Quality Report 2019, India was the fifth most polluted country in 2019 and having two-thirds of the world's most polluted cities based on the paramater of expousers level of $PM_{2.5}$.The core motivation of this research is to quantify the contribution of air pollutants expousers and also define the ICT-enabled architecture for devising the preventive alerts for limiting the expouser of pollutants for limiting the health hazards. Since the human and other well being are under the expouser of ambient air which is the necessity required for their survival on the earth. Further, the proposed system of air quality monitoring will help in preventing and controlling airborne allergies and reducing the burden of disease and the cost of treatments. Epidemiological studies conducted earlier elaborated that the air particles availability in the ambient air with high or low concentration can even cause serious health hazards having short and long-term health effects (Brunekreef et al. 2005). The air pollutants expouser have diverged impact on human being ranges from nausea, difficulty in breathing (COPD), cardiovascular disease to causing cancer (Hochadel et al. 2006). The presence of Particulate Matters ($PM_{2.5}$ and PM_{10}) in the air of urban areas has a high association and contribution in rising the emergency visits or hospital admissions because of respiratory illnesses (WHO 2003). Children less than 15 years of age and peoples above than 60 years at higher risk and expouser of polluted air environmental may cause allergies or asthma. The adverse environmental health impacts on living beings have been well acknowledged by Canadian authorities. The environmental authority of Canada has also defines that the "asthma, lung diseases, cardiovascular syndrome, allergies as human health issues linked to the poor quality of air [6]." The pollutants availability in the air are the most possible risk factor for causing asthma and people having comorbidities like heart diseases, lung diseases are at higher risk group, may be affected by pollution even on moderate or low pollution levels. Moreover, the children and young adults already having asthma may have faster breathing rates and the lung development of such cases may be continuous, can take longer time than the common period. Fig. 2 of World Resources Institute represented in their monthly update on August 2008 as Air Pollution's Causes, Consequences, and Solutions, represents the potential long

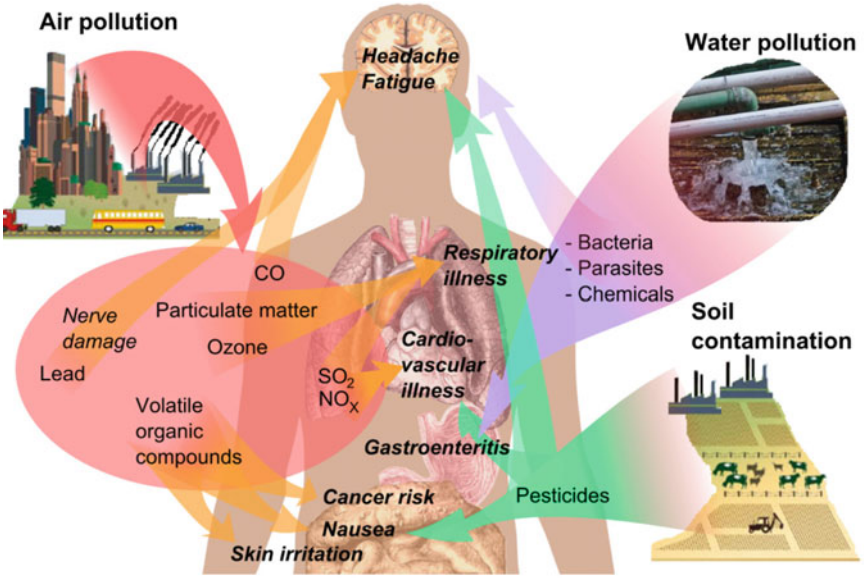


Fig. 2 Health effects of pollution (World Resources Institute 2018)

and short-term health effects of air pollutants [7]. The research work [8] defined the new algorithm for calculating the air quality index using the optimized feature selection process with the Firefly algorithm and Support Vector Machine for further calculation of air quality index.

4 Methodology and Architecture

Since the proposed eHealth System is based on the input of core component, i.e., quality of the air in the local environment. The Air Quality Monitoring System (AQMS) is required for defining the facility for measuring the wind speed, humidity, other weather parameters, concentration of air pollutants, and particulate matters for specific period or time frame. Hence the pollutants availabilty, expouser of bad air quality required to be measured for specific duration for deriving the predictions of preventive healthcare.

4.1 Air Quality Index

The color-coding scheme utilized for representation of numeric values for defining the level of pollutants availability in the ambient air for better understanding and

remembrance is known as air quality index. It is defined in multiple ways by the various countries using different methodologies of calculations for measuring the quality of air, defining the availability of harmful gasses. It is also known as Pollution Standard Index (PSI) during 1974 [9]. Various countries across the globe have defined the air or pollutions monitoring index like Green Index (GI), Fenstock Air Quality Index (FAQI) Canadian Environmental Quality Index (CEQI), Environmental Protection Agency-Pollution Standard Index (EPA-PSI), National Ambient Air Quality Standards (NAAQS), National Ambient Air Quality Standards And Significant Harm Level (NAAQS-SHL), Air Pollution Index (API), Oak Ridge Air Quality Index (ORAQI), Greater Vancouver Air Quality Index (GVAQI), Most Undesirable Respirable Contaminants Index (MURC), National Air Quality Index, Indian-National Air Quality Index (IND-NAQI), Air Quality Depreciation Index (AQDI) and Air Quality Risk Index (AQRI). Table 1 defines the comparison of various air quality indexes derived by the different countries for defining the targeted pollutants of air [9].

Table 1 defined the comprehensive comparison of various air quality indexes defined internationally along with the category of pollutants analysis for predicting the air quality. These air quality indexes defined variable ranges or color-coded bifurcations for generalizing the country or region-specific definitions of air quality index.

4.2 Indian-National Air Quality Index (NAQI) and Color-Coding Definition

The color-coding scheme for representing the status of the air quality and its effects on living beings in context to India the IND-AQI is shown in the form of color pattern in Fig. 2. IND-AQI is defined in six levels for eight major pollutants like PM_{10}, $PM_{2.5}$, NO_2, SO_2, CO, O_3, NH_3, Pb (Fig. 3).

4.3 Architecture and Pervasive Healthcare Prospectives

In order to obtain the objectives of Smart eHealth System of pervasive healthcare, the IoT-based cloud computing architecture has been defined for predicting the air polluting and corresponding analytics for representing the HealthMap for individual references. The architecture also generates preventive health alerts for initiating the air cleaning process specially in case of indoor users. And also help them to auto-initiate the disinfection process through UVC-lamp for specific duration.

Table 1 Air quality indexes and targeted air pollutants

Air pollutants/AQI	GI	FAQI	CEQI	EPA-PSI	NAAQS	NAAQS-SHL	API	ORAQI	GVAQI	MURC	IND-NAQI	AQDI	AQRI
O_3					X	X	X		X		X		
PM_{10}						X	X		X		X		X
$PM_{2.5}$					X	X	X	X	X		X		X
CO		X			X		X	X	X		X		X
SO_2	X	X			X	X	X	X	X		X	X	X
NO_2					X		X		X		X	X	X
COH	X									X			
TSP		X										X	
Photochemical oxidants								X					
Suspended PM												X	
Other pollutants or indices			X	X		X					X	X	X

Numerical Value of Air Quality Index	Color Code	Health Concens / Levels	Possible Health concernes and Impact
0 – 50	Green	Good	Quality of Air is Satisfactory. Poses no risk or minimal impact of pollutants on health.
51 – 100	Light Green	Satisfactory	Air contamination is minimum and cause minimal health concerns to the sensitive individuals.
101 – 200	Yellow	Moderate	Air quality is unhealthy for sensitive individuals, health may be effected , having breething discomforts to the people having asthama and lungs diseases.
201 – 300	Orange	Poor	Large number of people may experience health concerns, problem in breathing , heart illnesses with long expouser of air quality. Long terms health impact may also occurs.
301 – 400	Red	Very Poor	Serious Health impacts on general public. All sensitive individuals will have serious illnesses and others have short longtems health impacts.
>400	Dark Red	Sever	Air quality is bad, entire population is at risk,sever health concerns. Emergency conditon and health warning is suggested as entire population will have serious health issues with the expouser such air.

Fig. 3 Representation of color coding scheme of NAQI

4.4 System Design

The core modeling of the system having the Sensor Array Unit of gaseous sensors and PM-monitoring sensors fixed along with GPS and Bluetooth modules arranged along with local storage-memory module compatible with SD card and configured using Microcontroller as Processing unit (Fig. 4).

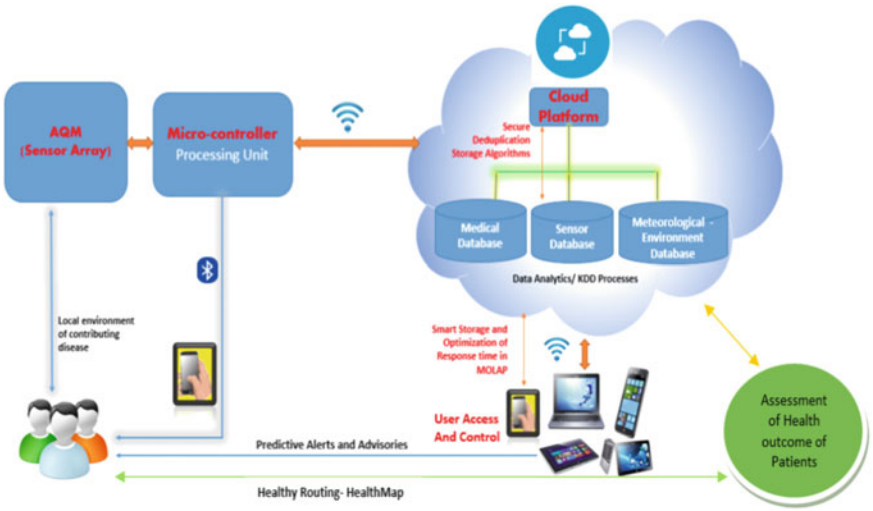


Fig. 4 Conceptual design of smart eHealth system of ambient air quality monitoring

This data acquisition system comprising of micro-controller and sensors array programmed for capturing the parameters for generating the prospective Local-area Air Quality Index (LAQI). The system modeling technique will be able to produce the LQAI (both for indoor and outdoor environments) on mobile devices using bluetooth or internet connectivity.

4.5 Core System Architecture

The core architecture of the system comprises of the Hardware Unit, Communication technology, Cloud platform, and predective modelling techniques being untilized for initiating the predictive health alerts along with process of initiating the disinfection processes or control measure for air cleaning. Figure 5 is representing the core system architecture of Smart eHealth System of ambient air quality monitoring.

The Cloud platform is utilized for integrating multiple platforms and databases like Medical Database, Real-time Sensor Database, and Meteorological-Environment Database. Further this plateform will support the multiple data analytical processes, knowledge discovery techniques for generating the Knowledge Discovery Database by linking the parameters of health, environment, and meteorological departments. Based on the parametric comparison of sensor outcomes (LAQI), Knowledge Discovery Database, and Electonic Health Record, the predictive health analytics will be generated with healthy routing plans-HealthMap. This deterministic

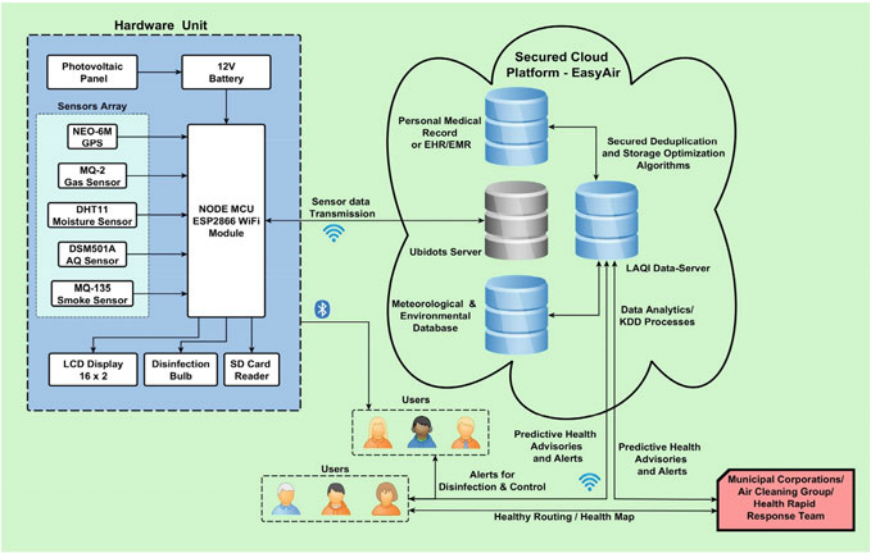


Fig. 5 Core system architecture of ambient air quality monitoring in smart eHealth system

system is having increased predictability by devising LAQI, Healthrouting, initiating disinfection alerts and control measures.

This health map and predictive alerts and advisories will represent the places or routes which may be avoided for restricting the airborne allergies and asthma exacerbation. Moreover, the Fog Computing platforms are also beneficial in this architecture as multiple data aggregations, analytical processes in context to EHR/PHR, Meteorological, Environmental Data integration will be undertaken in this system for generating the knowledge discovery for releasing the preventive health alerts.

Predictive Health alert will also trigger information-sharing mechanism for timely interventions as:

- Timely alert to the patients for start air clearing process (SOPs to be designed and followed).
- Trigger the room disinfection process in case of poor indoor air quality either by advising ventilation processes and also auto switching of UVC disinfection lamp for predefined interval. This mechanism will help to reduce the density of virus, pathogens, bacterias, COVID-19, etc. The protection alerts of UV light exposures of human body as safety measure.
- Since n-numbers of nodes are contributing the parameters for defining the air quality parameters of areas in case of outdoor air so alerts to Municipal Corporation or AirCleaning groups be initiated for timely actions like:
  - Initiate alerts for restricting the burning of Solid waste, closure of emission industries, encourage the carpooling or use of cycle, electric vehicles for travelling. Channelized the traffic on signal points to reduce the stoppage time.
  - Switch-on alerts for the PM purifiers, Switch-on alerts for Artificial Raining guns in highly polluted areas (Deployment of AirCleaning Brigades).
  - Alerts for restricting public gatherings and social events.
  - Sensitize the public based on relative air quality index (calculated on the air quality parameters of LAQI) and also represent the color coding of areas as per predefined range of color codes of IND-NAQI.

The other aspects which are helpful in reducing the carbon emission with the deployment of Cloud Computing are represented in Fig. 6 [9, 10] as mentioned.

5 Advantages and Limitations of Proposed System

Although this system yields multiple benefits for technology-driven Health ecosystem like advanced preventive health alerts to the individual based on their local environmental conditions, which are further beneficial in curtailing the number of hospitals visits and improve cost-effectiveness as well as the burden of the diseases. Since multiple sources and databases are integrated for defining the healthcare outcomes and predictive alerts hence the accuracy, completeness of information populated in the databases specially the electronic health records, and deployment

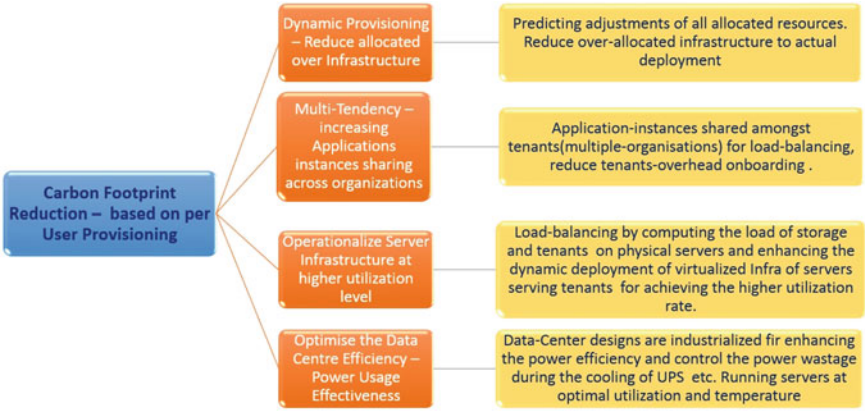


Fig. 6 Benefits of cloud computing platform in context to carbon emission [9, 10]

of chemical Sensors in sensor array unit is recommended otherwise the false results may be expected from this system. The special calculation factors are required to be defined for representing the LAQI of areas like national highways, population gathering locations, and Industrial areas where there is possibility of higher level of outdoor air pollution. The limitation of higher rate of breathing may also be taken in to account for defining the exposure of pollutants using such systems. Also, the real-time air quality monitoring (AQM) is confined to the area hence, it may require more numbers of data acquisition nodes to be put in place for covering the large geographical areas, which may increase the cost, complexity in analytics and also increase the response time of the system.

6 Challenges

Since the data collections, transmission, and storage activities of Wireless Sensor Network (WSN) is on real-time basis, hence violation of logical consistency during the transaction processes in Real-time Relation Database Management System (RT-RDBMS) are required to be monitored and addressed with prioritized methods. The clear-cut methods and authentic correlation mechanism are needed to be defined for sensor data transmission during the routing of queries through physical network among sensors and their neighborhood sensors. The quality of services are also required in systems equipped with wireless sensor network so the common challenges of big data in wireless sensor networks are also defined using Fig. 5 [7].

Continuous power-management, dynamic storage management and refreshing methods, and multiple query handing processes are the challenges as the real-time data of Smart eHealth System are dealing with the human life aspects and

cure aspects. The secured dynamic storage without-missing data values transmission, missing-value estimation to near-real, and then its storage in WSN requires enforcement of adequate security protocols and data-analytical algorithms (Fig. 7).

Real-time medical datasets with current state of medical data, history, and treatment details are challenges due to lack of deployment of uniform hospital information system across many countries. Hence, real-time preventive health alerts as per current state of patient's health and location/areas of proposed system may create confusion or wrong alerts, therefore medical database of allergic and asthmatic patients, etc., need to be updated in real-time basis. The other issues with WSN are Denial of service requests (DoS) by disrupting the availability and services of linked medical devices over the network leading to the overflowing of system with user requests [11]. This overflowing request creates massive problem like over-congestion over network when operating on distributed systems architechture. Hence, secured framework is required

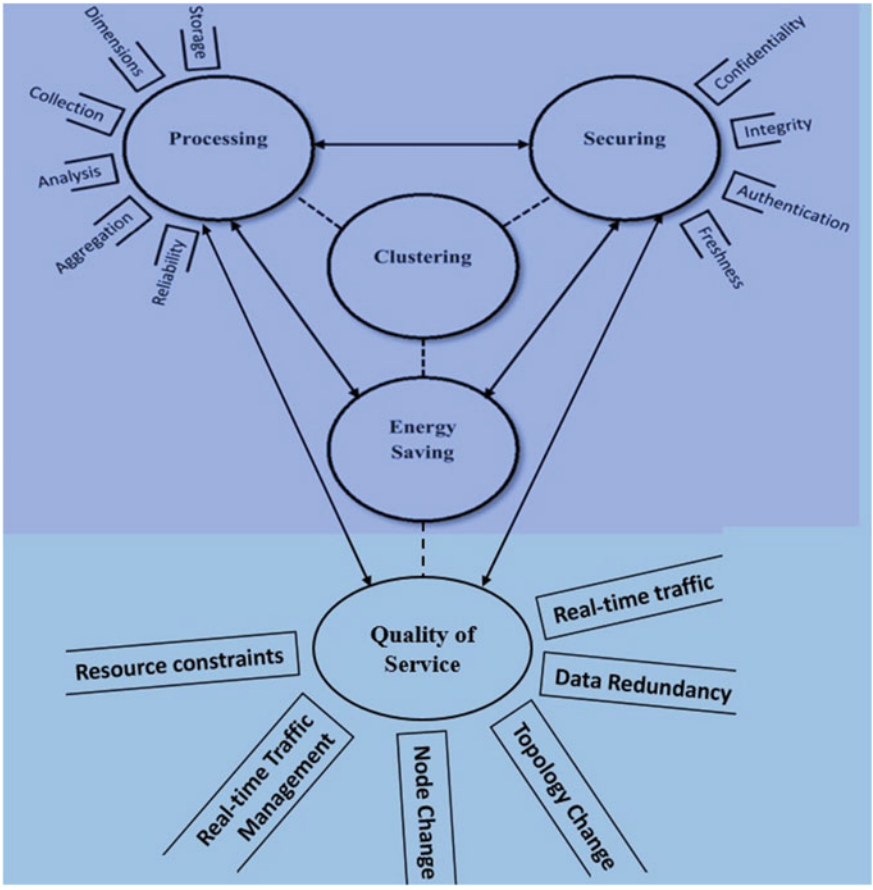


Fig. 7 Challenges of WSN linked large-scale data aggregation system

to be deployed in context of IoMT for addressing Denial or Distributed Denial of services, so that authentic and correct preventive health analytics be provided to the users distributed geographically.

7 Future Work

Ethics and guidelines for the usages of IoT devices and IoMT services for health-centric system needs to be defined. Standardization of usages of IoT-based WSN systems of Healthcare models and IoMT is suggested as such devices are having linkages with human lives and curative behaviors of diagnostics and treatments. This will help in defining the new feature as Prevasive Healthcare as a Service (PhaaS) of Cloud Computing. Also, standardization and modelling of data catalogs/metadata with semantic indexes for uniformity and interoperable medical data exchange are also needed for QoS in Smart eHealth System. Framework and guidelines for addressing the privacy issues while dealing with Medical Datasets are desired as friendliness of Smart eHealth Systems and IoT systems are needed every time for service provisioning.

Acknowledgments The authors acknowledge the contributions of their friend Mr Rajinder Maurya, B.Tech (Electronics & Communication Engineering) in realizing the functioning as well as tune-in the optimization of the Hardware Unit for this research.

Funding Information The research is not funded by any organization.

References

1. WHO, preventing disease through healthy environments (2002)
2. A. Prüss-Üstün, C. Corvalán, *Preventing Disease through Healthy Environments* (2006), pp. 5–13
3. G. Marques, J. Saini, M. Dutta, P.K. Singh, W.-C. Hong, Indoor air quality monitoring systems for enhanced living environments: a review toward sustainable smart cities. Sustainability. **12**, 4024 (2020)
4. T. Truong, A. Dinh, K. Wahid, An IoT environmental data collection system for fungal detection in crop fields. Can. Conf. Electr. Comput. Eng., 0–3 (2017)
5. M. Firdhous, B. Sudantha, P. Karunaratne, IoT enabled proactive indoor air quality monitoring system for sustainable health management. Comput. Commun. Technol. (ICCCT), 2017 2nd Int. Conf., 216–221 (2017)
6. WHO, prevention of allergy and allergies asthma, pp. 1–14 (2003)
7. World resources institute. August 2008 monthly update. Air pollution's causes, consequences and solutions. https://www.wri.org/resources
8. P. Rahi, S.P. Sood, R. Bajaj, et al, Air quality monitoring for Smart eHealth system us-ing firefly optimization and support vector machine. Int. J. Inf. Tecnol. **13**, 1847–1859 (2021). https://doi.org/10.1007/s41870-021-00778-9

9. *Cloud Computing and Sustainability: The Environmental Benefits of Moving to the Cloud*. Accenture & WPS (2010)
10. P. Rahi, S.P. Sood, R. Bajaj, Smart platforms of air quality monitoring: a logical literature exploration. in *Futuristic Trends in Networks and Computing Technologies. FTNCT 2019*, ed. by P. Singh, S. Sood, Y. Kumar, M. Paprzycki, A. Pljonkin, W.C. Hong. Communications in Computer and Information Science, vol. 1206 (2020). https://doi.org/10.1007/978-981-15-4451-4_5
11. H. Vahdat-Nejad, S. Eilaki, S. Izadpanah, Towards a better understanding of ubiquitous cloud computing. Int. J. Cloud Appl. Comput. **8**(1), 1–20 (2018). https://doi.org/10.5555/3212646.3212647
12. G. Muhammad, S.M.M. Rahman, A. Alelaiwi, A. Alamri, Smart health solution integrating IoT and cloud: a case study of voice pathology monitoring. IEEE Commun. Mag. **55**(1), 69–73 (2017)
13. P.E. Stevenson, H. Arafa, S. Ozev, H.M. Ross, J.B. Christen, Toward wearable, crowd-sourced air quality monitoring for respiratory disease (2017), pp. 140–143
14. C. Poon, Y. Lussier, G.-Z. Yang, Big data for health. IEEE J. Biomed. Health Inform. **19**(4), (2015)
15. B. Coats, S. Acharya, Bridging electronic health record access to the cloud. in *2014 47th Hawaii International Conference on System Science*
16. H.J. Lee, S.H. Lee, K.S. Ha, H.C. Jang, W.Y. Chung, J.Y. Kim, Y.S. Chang, D.H. Yoo, Ubiquitos healthcare service using Zigbee and mobile phone for elderly patients. Int. J. Med. Inform. **78**, 193–198 (2009)
17. D.M. Kendall, R.A. Kaplan, C.F. Paulson, J.L. Parkes, A.M. Tideman, Accuracy and utility of a 10-test disk blood glucose meter. Diabetes Res. Clin. Pract. **67**, 29–35 (2004)
18. P. Lecomte, I. Romon, S. Fosse, D. Simon, A. Fagot-Campagna, Self- monitoring of blood glucose in people with type 1 and type 2 diabetes living in France: the entry study 2001. Diabetes Metab. **34**, 219–226 (2008)
19. H.J. Hermens, M.M.R. Vollenbroek-Hutten, Towards remote monitoring and remotely supervised training. J. Electromyogr. Kinesiol. **18**, 908–919 (2008)
20. V.K. Omachonu, N.G. Einspruch, Innovation in healthcare delivery systems: a conceptual framework. Public Sect. Innov. J. **15**(1), (2010)
21. H.-K. Kim, Y. Hyun, Design for U-health care hybrid control systems. Int. J. Softw. Eng. Its Appl. **8**(2), 375–384 (2014)
22. A. Tahmasbi, S. Adabi, A. Rezaee, Behavioral reference model for pervasive healthcare systems. J. Med. Syst. **40**(12), 1–23 (2016). https://doi.org/10.1007/s10916-016-0632-0. Last accessed 1 Dec 2016. [Online]
23. A.P. Mukherjee, S. Tirthapura, Enumerating maximal bicliques from a large graph using MapReduce. Proc.—2014 IEEE Int. Congr. Big Data Congr. **1374**(2004), 707–714 (2014)
24. Asthma UK, https://www.asthma.org.uk/
25. V. Omachonu, N. Einspruch, *Innovation in Healthcare Delivery Systems: A Conceptual Framework* (The Public Sector Innovation Journal, The Innovation Journal, 2010)
26. A. Alnahdi, S.-H. Liu, Mobile Internet of Things (MIoT) and Its Applications for Smart Environments: A Positional Overview. in *2017 IEEE International Congress on Internet of Things (ICIOT) 2017 Jun 25*. IEEE, pp. 151–154 (2017)
27. J. Kaur, K.S. Mann, AI based healthcare platform for real time, predictive and prescriptive analytics. Commun. Comput. Inf. Sci. **805**, 138–149 (2018). https://doi.org/10.1007/978-981-13-0755-3_11
28. G. Marques, R. Pitarma, Indoor air quality monitoring for enhanced healthy buildings. Indoor Environ. Qual. 1–17 (2019). https://doi.org/10.5772/intechopen.81478
29. A. Sangeetha, A. Thangavel, Pervasive healthcare system based on environmental monitoring. Intell. Pervasive Comput. Syst. Smarter Healthc. 159–178 (2019). https://doi.org/10.1002/9781119439004.ch7
30. L.A. Tawalbeh, R. Mehmood, E. Benkhlifa, H. Song, Mobile cloud computing model and big data analysis for healthcare applications. IEEE Access **4**(c), 6171–6180 (2016). https://doi.org/10.1109/ACCESS.2016.2613278

31. B. Arnrich, O. Mayora, J. Bardram, G. Tröster, Pervasive healthcare. Methods Inf. Med. **49**(01), 67–73 (2010)
32. U. Varshney, Pervasive healthcare. Computer **36**(12), 138–140 (2003). https://doi.org/10.1109/MC.2003.1250897
33. S. Tanwar, Q. Bhatia, P. Patel, A. Kumari, P.K. Singh, W. Hong, Machine learning adoption in blockchain-based smart applications: the challenges, and a way forward. IEEE Access **8**, 474–488 (2020). https://doi.org/10.1109/ACCESS.2019.2961372
34. F. Alsubaei, A. Abuhussein, S. Shiva, Security and privacy in the internet of medical things: taxonomy and risk assessment. in *Proceedings—2017 IEEE 42nd Conference on Local Computer Networks Workshops, LCN Workshops 2017, September 2018*. IEEE, pp. 112–120 (2017). https://doi.org/10.1109/LCN.Workshops.2017.72
35. K. Kanchan, A. Gorai, P, Goyal, A review on air quality indexing system. Asian J. Atmos Environ. **9**, 101–113 (2015). https://doi.org/10.5572/ajae.2015.9.2.101

Identification of Missing Person Using Fusion of KNN and SVM Approach

Sandeep Rathor, Afreen Hasan, and Ankur Omar

Abstract Each day an enormous number of people get missing from kids to senior citizens because of some mental illness or Alzheimer's or Dementia. Out of all, most of them are trackless. Therefore, an artificial intelligence-based technique is proposed in this paper. It helps to find missing persons and can be used by the government officials like police or the common people. If a person goes missing, relatives of that person can visit the nearest police station and provide the details of the missing person which includes a photo (picture). When any person got such a missing person then, they can capture a picture of that person and upload it in our proposed model. Now, our proposed model tries to find that match in our database using the face recognition encoding technique. If a match is found, the proposed model will inform police officers along with the current location of the missing person.

Keywords Use of artificial intelligence · Rescue of missing person · Missing person identification · Face encoding

1 Introduction

The incident of a "missing person" is very painful for any family. It can be caused due to medical illness or by any person intentionally. Data and statistics of missing persons show that there is a drastic increase in the number of lost person in the past few years and is assumed to be increased further. Many authorities pursue to identify the victims so they can be returned to their respective families, for investigative purposes and/or for legal reasons [1]. There are numerous categories of missing persons of different ages and different reasons. Most of the cases are juveniles who runaways due

S. Rathor (✉) · A. Hasan · A. Omar
Department of Computer Engineering and Applications, GLA University, Mathura, India
e-mail: Sandeep.rathor@gla.ac.in

A. Hasan
e-mail: Afreen.hasan_cs17@gla.ac.in

A. Omar
e-mail: Ankur.omar_cs17@gla.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_40

to family restrictions or different family issues. Children are the future of the nation so children missing has been a worrying issue. The government never earmarked any budget for tracking the missing persons. The child and women trafficking rackets are big reasons behind more than half of missing cases. Other than that, most of the lost people are subjected to many severe crimes such as kidnapping, rape, or murder, etc. This era is the era of computers and the internet. So we require a system that can work effectively to search for a missing person in real time [2, 3]. Definitely, such type of system will help the family members of missed persons.

The photo and the details given by the relatives of the missing person will be uploaded in our system using the Desktop Application and will be stored in our database. Where the details will be stored and the image will be gone through the method of facial extraction using the dlib facial landmark generator and it will lastly be downloaded for finding a match using a machine learning algorithm. The public is powered to upload a picture of any person in undetermined circumstances. While uploading the detail they need to give their name, mobile no. and their current location along with the image of the person found. Henceforth, the proposed system will be matching the uploaded images through our database, and if a match s found it will be notified to the police officers and in case no match is found it will be stored in our database.

2 Literature Survey

In this section, we discussed various techniques proposed by the researchers in the same context. Formerly, many models have been proposed to find a missing person and have been proved helpful and faster than the manual investigation system. However, they have their pros and cons.

A detailed survey on face recognition techniques is proposed in 2020 [3]. It uses face recognition and a linear SVM classifier. The cons of the system are that if the missing person's age is in between 0 and 10 years, the efficiency and accuracy of the system. This proposed system is not suitable for the real-time environment. Chandran et al., came up with the same idea as presented with model the using deep learning-based face recognition following the match with SVM [4]. The limitation of the system is that there is no such dependency on the common people who can prove to be helpful. In the same context, another research is proposed by [5]. In this paper, the authors used principal component analysis to reduce the spatial features and designed a face recognition system. As a result, they came up with two disadvantages one dealt with the complexity and the other was that the system only read similar facial expressions. A research using LBPH technology to search the missing person is proposed by Muyambo [6]. The accuracy of the proposed system for face recognition was only 67.5% and gets comparatively difficult with accessories occlusion. The dataset used in this research was limited. In the same context, Birari Hetal [7] proposed an android-based model by using SWF-SIFT technique. This technique is used for comparing the face of two images. The demerit they faced was that the system became

slow and complex with cost-effective. It required more time to execute the process. A new technique for face recognition is proposed by Patel et al. [8]. In this paper, the author used the Line Edge Method for face recognition and implemented it to search for a missing person. The system worked well however, it has an efficiency of 85%. Another mobile-based model was proposed by Thomas and Kahonge [9]. In this model, no advance technique is used. It was just the modification of the manual and the old system. In this model, the trusted person or any family member of the missing person can upload the details of the person and when the missing person is found by someone they can contact the parents and relative accordingly to detail. The new research in the same context was proposed that focuses on finding missing persons using RFID technology [10, 11]. The main limitation of this system was that the person has to wear physically all the time RFID tag which is unusual. There have been many other face recognition methods to be developed for many other purposes. Eigenface algorithm [12, 13] is one such algorithm which is simple and effective but it is pose-effective which is the biggest drawback. Another system in this context, used deep learning approach and CNN Algorithm to build the model for the purpose of face recognition but ultimately resulted in low accuracy [14, 15].

Several models have come up with many exceptional ideas, irrespective of that there are lots of drawbacks which are listed in Table 1. In this paper, we have made tremendous effort to overcome the limitation faced by different authors.

Table 1 shows the work proposed by various researchers in a similar context. All the previous proposed work have either less accuracy or limited datasets.

The objective of the proposed research is to identify the missing person accurately with low computation time. The proposed research used a set of existing machine learning techniques to obtain acceptable accuracy. The proposed research can be useful for society at a low cost.

3 Proposed Methodology

In this section, an effective method is used for searching for a missed person. The proposed system mainly focuses on two parameters, i.e., searching for a missing person works easier and faster. If the proposed model finds the missed person successfully, then only we can rescue him/her safely.

The proposed framework for identifying the missing person for desktop application and android application is proposed as.

Figure 1 shows two different applications, i.e., Desktop application and Android Application. The processes start with uploading the data in our desktop application from which the facial key will extract and stored in firebase. Android application goes hand in hand by registering a missing person in the application which leads to the submission of all cases in the same database, i.e., firebase. Firebase database is used to store because it is simple to use and not complex as rest databases. The comparison is made between the facial key extracted from both the application which is stored in database using KNN Algorithm. If a match is found, it is notified. If the

Table 1 Comparison of different state of art techniques

S. no.	Title	Technique used	Limitations	Accuracy (%)
1	Efficient Face Recognition System for Identifying Lost People	SVM Classifier	1 In case of age between 0 and 10 accuracy decreases 2 Testing has been done on very limited dataset	97.6
2	Missing Child Identification System Using Deep Learning and Multiclass SVM	Multiclass SVM Classifier	The algorithm is complex which leads to slowing down the system	80.4
3	Missing Child Identification Using Face Recognition System	PCA Algorithm	Similar facial expression detection only	96
4	An Investigation on the Use of LBPH Algorithm for Face Recognition to Find Missing People in Zimbabwe	LBPH	Less interactive with accessory occlusion	67.5
5	Robust Face Recognition System for E-Crime Alert	Line Edge Mapping	Only for legitimate customer	85
6	Using a Mobile-Based Web Service to Search for Missing People—A Case Study of Kenya	Web Application	No modern technologies used just simple complaint registration system	–

match is not found, that means the image with which we are matching is not yet uploaded in our database and will ask to register in the android as well as in the desktop applications.

The proposed system consists of a GUI Application designed using python that can be utilized by the police stations to maintain a record of the missing person. In the backend, once the image is submitted it will be processed and facial key points are captured and will be stored in the database. Different police stations will be having access to this application and ultimately all the data that is uploaded will be stored in a central database.

The second important module of the system is our Android Application which can be accessed by any common person. In this application, if any person finds anyone suspicious or in any uncertain situation can upload the image and the location where the missing person was last seen, anonymously. The GUI Application uses Machine Learning Algorithm for matching the pictures that have been uploaded by the police and the public. If there is a match, the location can also be recognized for the same.

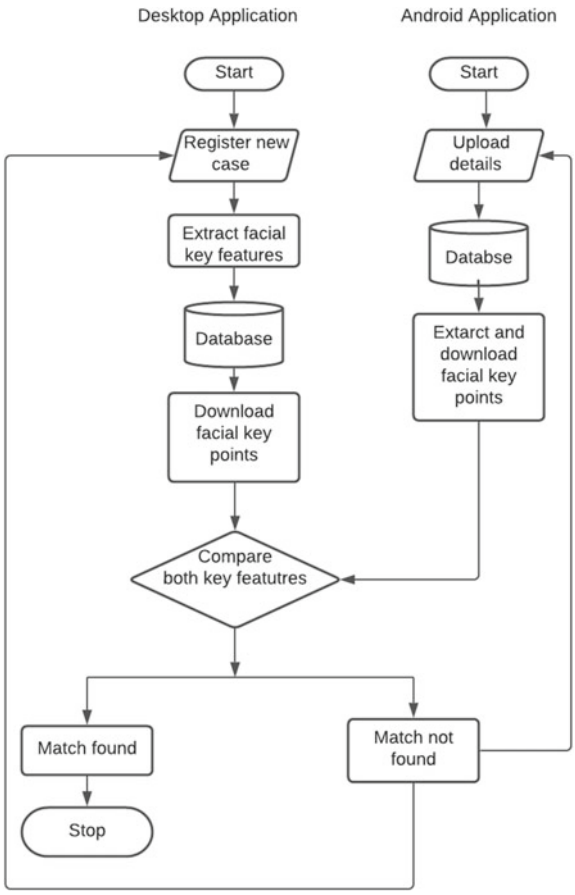


Fig. 1 Proposed framework for identifying of missing person

4 Results and Discussion

In this section, we discuss about the dataset used in the proposed system along with the obtained results.

4.1 Dataset Used

The proposed model is implemented on LFW face dataset. It comprises of approx. 13,000 images. We have tested all the state of art classification algorithms. However, our model works efficiently on KNN Algorithm. The achieved accuracy on different classifiers is shown in Fig. 2. It shows the accuracy of different algorithms with the dataset. We also used 4-cross fold validation technique in our proposed model and

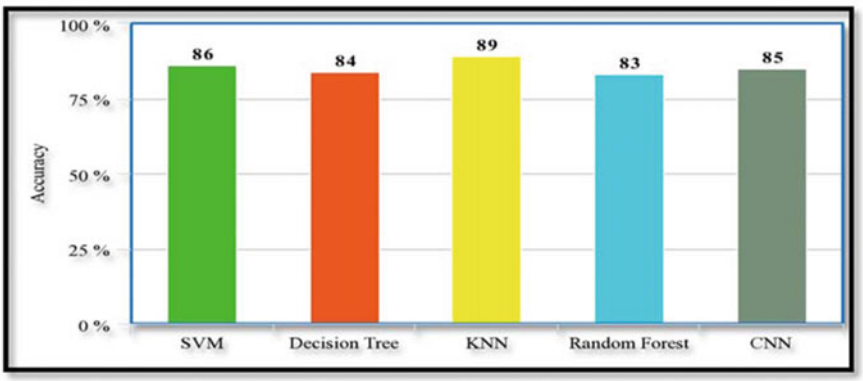


Fig. 2 Individual accuracy on different state of art classifiers

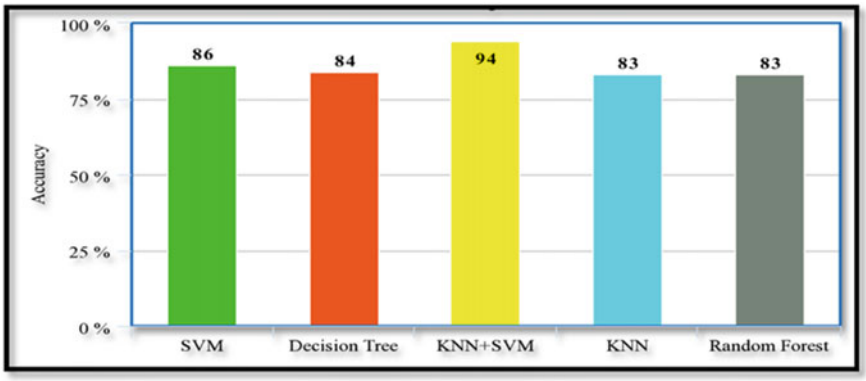


Fig. 3 Accuracy of the proposed model

found that KNN and SVM work effectively. Therefore, the proposed system used the fusion of KNN and SVM both so that, it works effectively and it is shown in Fig. 3.

Figure 3 shows that individually KNN and SVM have the accuracy of 83% and 86%, respectively. However, in our proposed model its accuracy is 94%, which can be suitable for real-time implementation.

4.2 Working of Proposed System on Desktop Application

The GUI application is which is designed using python and is authenticated to police and other government officials to register a new case. The information uploaded in the application is the name, father's name, mobile number, age, and image of the missing person. The uploaded data is stored in our database. Now, the 68 facial

Fig. 4 Main page of desktop application

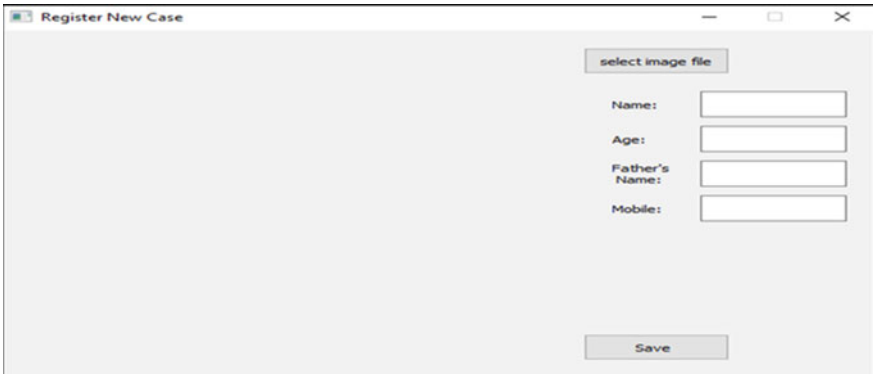


Fig. 5 Working of proposed system on desktop application

key points are extracted from every person's face by using the dlib facial landmark generator, and these facial key points are stored in the firebase.

Figures 4 and 5 show the main GUI interface and registration process used in the proposed system for desktop applications.

4.3 Working of Proposed System on Android Application

Common people like us can use this application to register a missing person as being anonymous. The information that is needed to be uploaded by the person is just the location, image, and mobile number. Here, the image is encoded in base64 form and stored in firebase to minimize the load on firebase. Therefore, 68 facial key points are again generated in the encoded image by using the dlib facial landmark generator

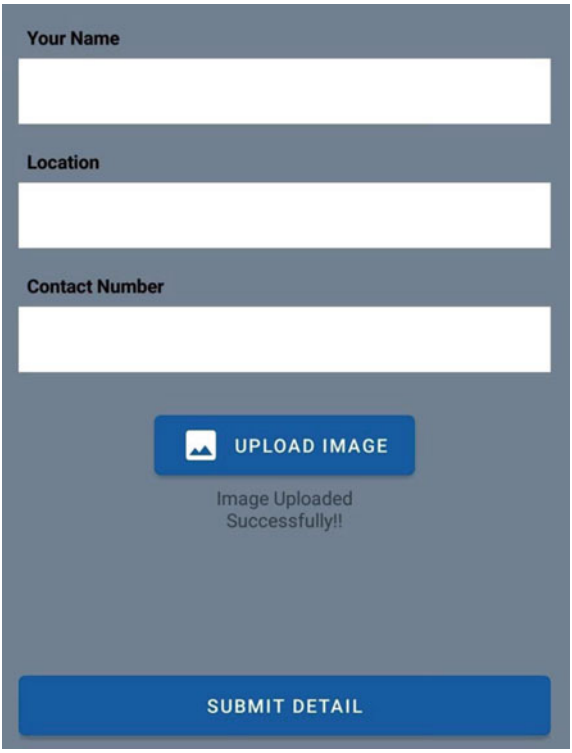


Fig. 6 Working of proposed system on Android Application

and stored in the database. The screenshot of the working process on the android application is shown in Fig. 6.

Finally, the matching step is processed in our GUI Application which uses machine learning algorithm to match the pictures uploaded by police and users. The facial landmarks are downloaded from firebase and the KNN classifier is trained using these points. To find a match all the points registered by the user are downloaded and the value is predicted using the trained KNN classifier. If the value is above 60%, it is predicted to be a match.

The proposed system has an accuracy of approx. 94%. The testing is performed on 1800 images, out of which 1400 images were used as training dataset and 600 images were used for testing purposes. After the execution, 564 images were matched correctly and the rest 36 images were misclassified.

5 Conclusion and Future Scope

The proposed model of searching a missed person is fastened, underprivileged. It is also affordable for poor people due to its low-cost features. The manual way of searching is very slow and cost-effective. Tracking a missing person cannot only be

limited to android applications and it can work on desktop applications also. The target ultimately is to build a system that can work efficiently to search the person. As soon as a picture is uploaded, the model starts working on it. Therefore, the search is promoted and integrated into the security cameras everywhere to make the search easier and faster which will be happening in a real-time environment.

From a future point of view, the proposed model can be implemented to rescue the missing person. The process of rescue is possible only when we find the person. Therefore, we will focus on rescue techniques also in near future.

References

1. M.D. Vigeland, F.L. Marsico, M.H. Piñero, T. Egeland, Prioritising family members for genotyping in missing person cases: a general approach combining the statistical power of exclusion and inclusion. Forensic Sci. Int. Genet. 49, 102376 (2020)
2. Z. Domozi, D. Stojcsics, A. Benhamida, M. Kozlovszky, A. Molnar, Real time object detection for aerial search and rescue missions for missing persons, in *2020 IEEE 15th International Conference of System of Systems Engineering (SoSE)* (IEEE, June 2020), pp. 000519–000524
3. S. Balaji, R. Saravanan, P. Pavadharani, I. Deepalakshmi, S. Suvetha, A survey on facial recognition techniques and stipulated algorithms. J. Interdiscip. Cycle Res., 1120–1126 (2020)
4. P.S. Chandran, N.B. Byju, R.U. Deepak, K.N. Nishakumari, P. Devanand, P.M. Sasi, Missing child identification system using deep learning and multiclass SVM, in *2018 IEEE Recent Advances in Intelligent Computational Systems (RAICS)* (IEEE , December 2018), pp. 113–116
5. G.C. Iannacone, R.C. Parra, Genetic structure and kinship analysis from the Peruvian Andean area: limitations and recommendation for DNA identification of missing persons, in *Forensic Science and Humanitarian Action: Interacting with the Dead and the Living* (2020), pp. 473–489
6. P. Muyambo, An investigation on the use of LBPH algorithm for face recognition to find missing people in Zimbabwe. Int. J. Eng. Adv. Technol. (IJEAT) (2018)
7. B. Hetal, Android based application-missing person finder, in Iconic Res. Eng. J. 1(12) (June 2018)
8. S. Patel, D. Maurya, V. Mhatre, P. Yadav, Robust face recognition system for e-crime alert. Int. J. Res. Eng. Appl. Manag. (IJREAM) (March 2016). ISSN: 2494-9150, Special Issue-01
9. T.M. Omweri, Using a mobile based web service to search for missing people–a case study of Kenya. Int. J. Comput. Appl. Technol. Res. **4**(7), 507–511 (2015)
10. Díaz, J. V., & Urueña, C. V. (2020). The Colombian experience in forensic human identification. Forensic Science and Humanitarian Action: Interacting with the Dead and the Living, 693–702.
11. S.B. Arniker, K.S.R. Rao, G. Kalyani, D. Meena, M. Lalitha, K. Shirisha, RFID based missing person identification system, in *2014 International Conference on Informatics, Electronics & Vision (ICIEV)* (IEEE, 2014), pp. 1–4
12. S. Sharma, V. Kumar, Low-level features based 2D face recognition using machine learning. Int. J. Intell. Eng. Inf. **8**(4), 305–330 (2020)
13. F. Tabassum, M.I. Islam, R.T. Khan, M.R. Amin, Human face recognition with combination of DWT and machine learning. J. King Saud Univ. Comput. Inf. Sci. (2020)
14. F. Zhao, J. Li, L. Zhang, Z. Li, S.G. Na, Multi-view face recognition using deep neural networks. Futur. Gener. Comput. Syst. **111**, 375–380 (2020)
15. S. Rathor, S. Agrawal, A robust model for domain recognition of acoustic communication using Bidirectional LSTM and deep neural network. Neural Comput. Appl., 1–10. (2021)

Current Trends and Future Prospects: Detection of Breast Cancer Using Machine Learning Techniques

Ruqsar Zaitoon, Ashwani Kumar, and Syed Saba Raoof

Abstract The root cause of death among the global population is considered Cancer. Cancer is originated from the compulsive accumulation of cells forming a tumor. One among these cancers is breast cancer. The 2020 cancer report states that 19.3 breast cancer cases were estimated and 10 million cancer deaths. In the past few years, a significant increase in breast cancer among women was perceived and it stands at fifth position among all the cancers, where 11.7% of female breast cancer cases were recorded according to the 2020 cancer report and 6.9% of deaths are caused. Many of the researchers noticed that the implementation of various machine learning techniques like feature extraction, feature selection, and classification can ease the task of physicians in breast cancer detection and diagnosis. Timely detection of breast cancer can save many lives and improve the treatment process. Thermograph is the most suitable screening technique for all age groups and is affordable compared to mammogram, Magnetic Resonance Imaging (MRI), and ultrasound. Precise tumor classification alleviates patients from pain and assists physicians with an accurate diagnosis. According to the survey done the main aim of many researchers is to implement automated and apparent breast cancer identification and classification system. Our paper presents the review of the most updated machine learning techniques and methods implemented by various researchers' for detecting, and diagnosing breast cancer. And future directions are outlined regarding various machine learning algorithms like SVM, ANN, and K-NN utilized for detecting breast cancer.

Keywords Breast cancer · Machine learning · Classification · Detection · Diagnosis

R. Zaitoon
Genpact, Gachibowli, Hyderabad, India

A. Kumar (✉)
Sreyas Institute of Engineering and Technology, Hyderabad, India

S. S. Raoof
SCOPE, Vellore, VIT, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_41

1 Introduction

Cancer is defined as uncontrollable expansion of tumor cells at any specific body part. Cancerous cells effect the healthy cell growth leading it into a peculiar way and even changes the shape of the cells [1]. This uncontrolled growth of cells may lead to death if not controlled at initial phase. Expansion of cells at breast forming cancerous tumor is known as "Breast cancer" [2]. Breast cancer is one of the major cancers caused among feminine group [3]. If diagnosed at initial phase, mortality rate can be controlled and better treatment can be provided which extends the life span of patient. Screening mammography is the most widely used and accepted method for breast cancer diagnosing that is breast X-ray. Symptomatic patients are recommended to undergo mammography screening so that cancer can be discovered at an early stage. As the mortality rate of woman due to breast cancer is high amidst all other cancers, there is a need for early cancer discovery. Medical imaging is the best technique for breast cancer diagnosing. There are different medical imaging techniques for breast cancer diagnoses like Digital Mammogram (DM), Infrared Thermography (IRT), Magnetic Resonance Imaging (MRI), Breast-Ultrasound (US), Computed Tomography (CT), and Biopsy. Among all these screening techniques, the most employed is mammogram and it is considered an extremely useful and vigorous technique to detect breast cancer and to lower fatality rate [4]. Usually due to its high radiation, it's not preferred for the age group below 40 years, above this, it's mostly recommended as it gives exact precise tumor screening images [5]. Due to this drawback, other screening methods were introduced like IRT, CT, and MRI. The following Table 1 represents various breast cancer screening techniques and their consequences.

Table 1 Various breast screening procedures and their respective merits and demerits

Screening method	System process	Merits	Demerits
Digital mammogram (DM)	X-ray analysis to detect variations of breast	Most effective and accurate method	High radiation due to which not recommended for age group below 40
Breast-ultrasound (US)	Ultrasound waves are employed to analyze the inner parts of the body	Simple and trouble-free method Less radiation	Inadequate resolution Couldn't screen Complete breast area
Breast MRI	Magnetic and radio waves are used	Serious conditions can be analyzed accurately	Physical examinations are advised
Thermography	Infrared scans are used to map the breast variations	Completes the screening process in less time and it's a noninvasive method	Results depend on body temperature due to which leads to errors

Based on the cancer stage and patient condition the screening method is recommended by physician as the breast region is the most sensitive part of body. DM is the cheap and safe screening method at initial phase, but in case of youths, its ineffective breast-US can be used in this scenario [6].

By the opinion of many experts, surgeons are able to diagnose any cancer with 80% precision, whereas Computed Automated Diagnostic systems (CAD) are able to diagnose with 91–94% accuracy. With the advancement of technologies to handle enormous data, the cancer detection can be improved. Various Machine Learning (ML) classification techniques like Support Vector Machine (SVM), Artificial Neural Network (ANN), K-Nearest Neighbor (KNN), and Decision Tree (DT) can be utilized to classify the cancerous and non-cancerous images and benign or malignant.

This paper represents a comprehensive literature of various techniques and approaches used for detecting cancer and mainly breast cancer detection and diagnosing techniques. Paper is organized in the following way. Sect. 2 explains summary of literature done, Sect. 3 presents various breast cancer database, Sect. 4 demonstrates ML and its techniques employed for breast cancer detection, and Sect. 5 represents conclusion.

2 Literature Review

Distinct approaches had been proposed for classifying mammogram scans. Therefore to figure out the best cancer classification technique, Table 2 represents some of the classification techniques with high accuracies and this table presents their corresponding techniques, features used for classification, datasets, and performance measures. It had been noticed that Mammographic Image Analysis Society dataset (MIAS) is mostly used. The main aim of breast cancer detection problem is to classify the input data into cancerous or non-cancerous images and whether it's benign or malignant tumor. As medical images like mammogram scans are complex images due to variance accompanied along with it, it becomes a difficult task for classification [7].

Several researchers had worked on breast cancer classification, detection, and diagnose tasks by employing various machine learning algorithms and some of those research works are discussed below. Barrett et al. [8] proposed a system to identify breast cancer by utilizing microwave radiometry which is based on human body temperature. Authors' combined data of two screening methods, i.e., microwave and thermography to provide high efficient system and recorded 96% accuracy rate.

Ibrahim et al. [9] proposed a CAD model to diagnose breast cancer by utilizing Radial Basis Function (RBF) and Multiple-Layer Perceptron (MLP) as a classifier and recorded 79.2% and 54.2% accuracy, respectively. Bushra et al. [10] developed a classification network by employing Back-Propagation NN (BPNN). Two datasets were used for experimental analysis, i.e., MIAS and Digital Database for Screening Mammograms (DDSM). The proposed system accurately classified the breast lesion at initial phase with an accuracy of 99%. Sharma et al. [11] proposed a cancer detection system by utilizing three different ML techniques viz. KNN, RF, and Navie

Table 2 Summary of literature review

Author name	Machine learning method	Screening method	Dataset used	Accuracy
Khan et al. [17]	SVM	DM	MIAS	68%
Diz et al. [18]	SVM	DM	INbreast BCDR	89.3% 83.1%
Phadke et al. [19]	SVM	DM	MIAS	93%
Cai et al. [20]	SVM	US	Private	86.9%
Choi et al. [21]	ANN + SVM	DM	DDSM	93.2% + 92.5%
Silva et al. [22]	SVM	DM	DDSM	83.53%
Liu et al. [23]	SVM	DM	INbreast	86.76%
Prabu et al. [24]	SVM	US	Private	95.85%
Sharma et al. [25]	SVM	DM	DDSM IRMA	97% + 99%
Wu et al. [26]	SVM	US	Private	96.67%
Azar et al. [27]	SVM	DM	WBC	97.14%
Behesthi et al. [28]	SVM	DM	MIAS	99%
Cai et al. [29]	SVM	MRI	Private	82.8%
Tan et al. [30]	SVM	DM	DDSM	80.5%
Garcia et al. [31]	SVM	DM	DDSM	93%
Hoffmann et al. [32]	SVM	MRI	Private	Not given
Retter et al. [33]	SVM	MRI	Private	87%
Sidropolous et al. [34]	SVM	IRT	Private	88.10%
Ding et al. [35]	SVM	US	Private	91%
Jian et al. [36]	SVM	DM	Private	100%
Ramos et al. [37]	SVM	DM	Private	96.9%
Zakeri et al. [38]	SVM	US	Private	95%
Gardezi et al. [39]	KNN	DM	MIAS IRMA	97.13% + 92.81%
Shibusawa et al. [40]	KNN	US	Private	93%
Wagh et al. [41]	KNN	MRI	Private	74.7%
R Hupse et al. [42]	KNN	DM	Private	Not Given
Singh et al.[43]	ANN	DM	DDSM MIAS	87.27% + 89.38%
Azar et al. [44]	ANN	DM	WBC	97.66%
Masmoudi et al. [45]	ANN	DM	Private	79%

(continued)

Table 2 (continued)

Author name	Machine learning method	Screening method	Dataset used	Accuracy
Dheeba et al. [46]	Swarm optimization ANN	DM	MIAS Private	97% + 91%
Dheeba et al. [47]	Differential evolution wavelet ANN	DM	MIAS	97%
Dong et al. [48]	RF	DM	DDSM	97.73%
Abdel et al. [49]	RF	US	Private	99%
Bruno et al. [50]	RF	DM	BCDR	100%
Gubern et al. [51]	RF	DCE + MRI	Private	95%

Bayes. Wisconsin Breast Cancer Data Repository is used for experimental analysis and results of these algorithms were compared with others which had demonstrated that KNN algorithm had achieved highest accuracy of 94% compared to other algorithms. Mohamed et al. [12] employed ANN algorithm with one fully connected and one hidden layer. Wisconsin Breast Cancer Data Repository was utilized to train and test the proposed model. Ahmed et al. [13] developed an automated system to detect breast cancer using SVM and recorded 87.12% accuracy. Kaur et al. [14] proposed a model based on neural networks technology. UCI breast cancer dataset was employed to train the model. The proposed model was compared with other machine learning algorithms like KNN and Naive Bayes and the comparison demonstrated that the proposed model had achieved high accuracy of around 91% then other algorithms. Whereas Sathy et al. [15] developed a model to detect breast cancer based on Decision Tree (DT) algorithm. Breast cancer classification system was proposed by Sun et al. [16] deep CNN algorithm was employed for this purpose.

3 Breast Cancer Benchmark Datasets

3.1 Digital Database for Screening Mammography (DDSM)

This dataset is created by the University of South Florida at Computer Science and Engineering Department with the help of Massachusetts General Hospital and Sandia National Laboratories. It comprises of 2500 sets and each set includes two breast images of particular patient and the CSV file includes the features such as age, abnormality information (ACR), sensitivity value of abnormality, density of breast, spatial information, and scanner details. It also provides ground truth details regarding tumor spot.

3.2 Wisconsin Breast Cancer Database (WBCD)

It consists of 699 files of breast acquired by Fine Needle Aspirates (FNA), where 444 are benign and 239 are malignant images. All the files include 9 attribute features of breast scans.

3.3 Full-Field Digital Mammography (FFDM)

It consists of 739 images where 327 are benign and 412 are malignant images. This dataset was captured on12 bit quantification and size of pixel is 100 × 100 m.

3.4 Mammographic Image Analysis Society (MIAS)

It is created by the research team of UK (UK National Breast Screening Programme). It consists of 322 breast digitized scans of 1024 × 1024 pixels. Ground truth data is also provided at the abnormality spot, i.e., at lesions location. This dataset is available at and can be downloaded from the University of Essex through PEIPA (Pilot European Image Processing Archive).

4 Machine Learning Techniques

ML is a method where systems are pre-trained and then input data is fed to make optimum decision, respectively, to the input data. It is the sub-branch Artificial Intelligence (AI) subjected that the system may learn and analyze the result from data given. The key concept of ML is it enables the systems/machines to learn and analyze by its own. It's broadly categorized into three types based on learning methods and data provided, i.e., Supervised Learning, Unsupervised learning and Reinforcement Learning. In supervised learning data, it is structured and labeled into groups based on this it decides the output. Whereas, Un-supervised learning data is not labeled, the system itself finds out the relation from input. The following Fig. 1 explains all the machine learning algorithms. Most widely applied ML algorithms are Support Vector Machine (SVM), Artificial Neural Network (ANN), k-Nearest Neighbor (KNN), Decision Tree (DT), and XGBoost.

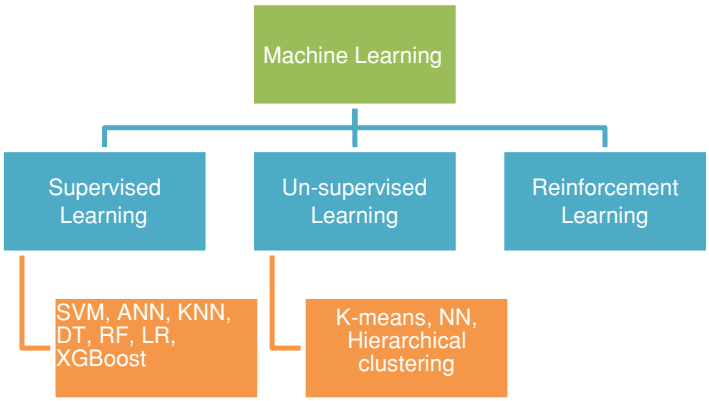


Fig. 1 Types of machine learning algorithms

4.1 Challenges

The occurrence of cancer among patients differs from one another. After treatment in some patient cancer may occur again after half-year while it won't occur again in some patients. Hence, designing the cancer detection system based on this occurrence is necessary as there is no such method designed so far. Through the research work, it is observed that algorithms comprising of multiple layers with neurons, activation functions, and memory units can be implemented to address this challenge. Some examples of such algorithms are ANN, LSTM, and RNN.

4.2 Artificial Neural Network (ANN)

ANN is inspired by working of neurons in brain where large number of neurons/nodes are connected with each other and work together simultaneously depending on each other. ANN network is created by connecting various layers, i.e., input, output, and hidden layers as shown in Fig. 2 [52] and further each layer are created by connecting numerous neurons and activation functions. Information is passed to input layer and then it is passed to preceding layers by performing necessary operations and finally passed to output layer which classifies the output.

4.3 Support Vector Machine (SVM)

SVM in supervised machine learning algorithms SVM is the most flexible learning algorithm. It can be employed to solve both regression and classification problems

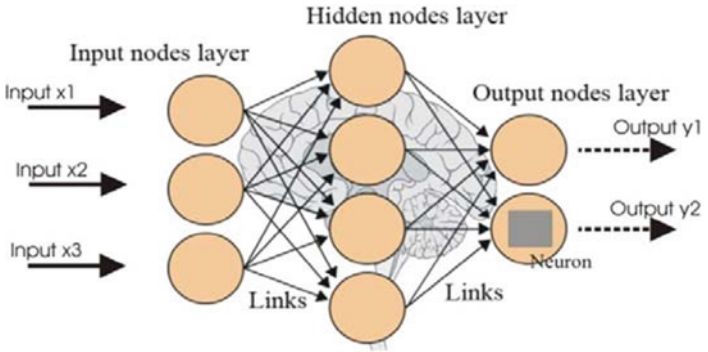


Fig. 2 Architecture of ANN [52]

precisely. Hyperplane the boundary is the line which separates outs the data into groups, i.e., classes. Based on the support vectors the hyperplane is created, support vectors are the values plotted on the graph. It can be utilized for classification problems, recognition, and detection tasks. It generally works well for binary classification tasks. It is noted by many researchers that it best suits for prognosis and diagnosis medical imaging, i.e., classifying diseases. The following Fig. 3 shows the SVM classification of data using hyperplane.

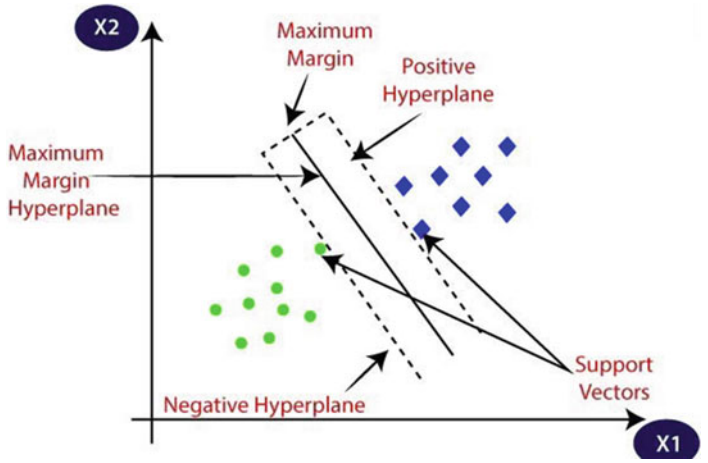


Fig. 3 Classification of data through SVM [52]

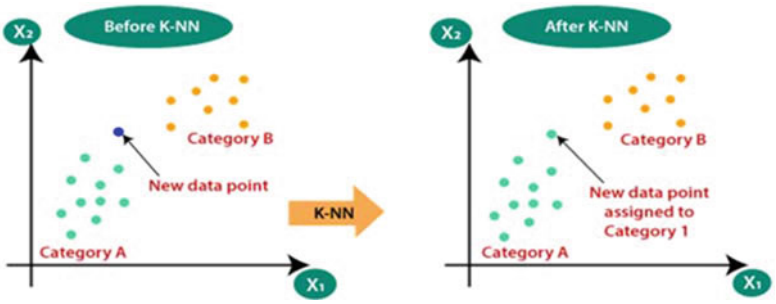


Fig. 4 Representation of KNN classifier [52]

4.4 K-Nearest Neighbors (K-NN)

KNN is one classification algorithm of supervised learning. It classifies the object based on the neighboring objects in other words classification is done based on the similarity of neighboring objects and to do so the distance between the two nearest objects is calculated by using various mathematical formulas like Euclidean distance, Manhattan, minkowski, etc., among which Euclidean is mostly preferred. The following Fig. 4 displays the KNN architecture.

4.5 Future Directions

New advancements and various future directions are addressed in the section in order to enhance and improve the efficiency of existing methods. The summary table of literature review, i.e., Tables 2 and 3 demonstrate the various databases that can be used for breast cancer analysis and performance measures with respect to algorithm. Some of these techniques are [23, 33, 41] accuracy of the system can be improved

Table 3 Breast cancer dataset and their details

Dataset	Screening type	No. of images	No. of patients
MIAS	Mammogram	322	161
WBCD	Mammogram	1398	699
FFDM	Mammogram	739	360
DDSM	Mammogram	Not given	2620
INbreast	Mammogram	419	115
QIN Breast	MRI, PET, CT	100,835	67
Breast MRI	MRI	99,058	64
DCE-MRI	CT	76,328	10
ISPYI	MR	386,528	237

further by using feature extraction techniques. Other segmentation techniques can be applied in [45] to procure quantitative measures aimed to highlight the lesion precisely. Fuzzy clustering techniques, K-means can be utilized in [22, 23, 29, 30, 34] to enhance segmentation process. To obtain higher results than 91% in [35] different clustering methods and distinct features can be extracted. The optimistic classifier employed in [46, 47] can be compared to other classifiers like SVM, ANN, and K-NN. Kumar et al. [52–56] proposed an object detection method for blind people to locate objects from a scene. They have used machine learning-based methods along with single SSMD detector algorithm to develop the model.

5 Conclusion

In this survey paper, we presented a brief summary of breast cancer, index rate of mortality due to breast cancer globally, and various breast cancer datasets available to analyze and detect breast cancer. Machine learning techniques and mostly used ML algorithms like ANN, SVM, and K-NN to detect and classify objects are explained. According to the survey, it is noted that majority of researchers had adopted supervised machine learning techniques to solve classification problems as they produce accurate outcomes. Through our research work, it had been observed that SVM algorithm is widely applied to classify and diagnose cancer.

References

1. W. Schulz, *Molecular Biology of Human Cancers* (Springer, New York, USA, 2007)
2. American cancer society: Breast cancer facts & figures 2017–2018. https://www.cancer.org/research/cancer-facts-statistics/breast-cancer-facts-figures.html. Accessed 3 November 2017
3. J. Eric, L.M. Wun, C.C. Boring, W. Flanders, J. Timmel, T. Tong, The lifetime risk of developing breast cancer. JNCI J. Nat. Cancer Inst. **85**(11), 892–897 (1993). https://doi.org/10.1093/jnci/85.11.892
4. A.F.M. Agarap, On breast cancer detection: an application of machine learning algorithms on the Wisconsin diagnostic dataset (2019 February). arXiv:1711.07831v4 [cs.LG] 7
5. A.A. Bataineh, A comparative analysis of nonlinear machine learning algorithms for breast cancer detection. Int. J. Mach. Learn. Comput. 9(3) (2019, June). 20:734–42 (2010)
6. A. Alkhateeb, Breast cancer computer-aided diagnosis system from digital mammograms, JAMMR 30(5), 1–15 (2019)
7. A.H. Barrett, P.C. Myers, N.L. Sadowsky, Detection of breast cancer by microwave radiometry. Radio Sci. **12**(6S), 167–171 (1977)
8. A.O. Ibrahim, Classification of mammogram images using radial basis function neural network, in: *Emerging Trends in Intelligent Computing and Informatics. IRICT 2019. Advances in Intelligent Systems and Computing*, edited by F. Saeed, F. Mohammed, N. Gazem (Springer, Cham, 2020)
9. B. Mughal, *Early Detection and Classification of Breast Tumor from Mammography* (COMSATS Institute of Information Technology, Islamabad, 2019)
10. Cancer Detection Using Machine Learning Algorithms, International Conference on Computational Techniques, Electronics and Mechanical System (2018)

11. M. Nemissi, H. Salah, H. Seridi, *Breast Cancer Diagnosis Using an Enhanced Extreme Learning Machine based-Neural Network* (IEEE, 2019)
12. A.M. Abdel-Zaher, A.M. Eldeib, *Breast Cancer Classification Using Deep Belief Networks, Expert Systems with Application* (Elsevier, 2016)
13. A. Kaur, P. Kaur, Breast cancer detection and classification using analysis and gene-back proportional neural network algorithm. Int. J. Innov. Technol. Explor. Eng. (2019)
14. P .Sathy Anaranan, in *Proceedings of International Conference on Systems Computation Automation and Networking* (IEEE, 2019)
15. W. Sun, T.L.B. Tseng, J. Zhang, W. Qian, Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. ComputMed Imaging Graph **57**, 4–9 (2017)
16. S. Khan, M. Hussain, H. Aboalsamh, G. Bebis, A comparison of different Gabor feature extraction approaches for mass classification in mammography. Multimed. Tools Appl. **76**(1), 33–57 (2017)
17. J. Diz, G. Marreiros, A. Freitas, Applying data mining techniques to improve breast cancer diagnosis. J. Med. Syst. **40**(9), 203 (2016)
18. A.C. Phadke, P.P. Rege, Fusion of local and global features for classification of abnormality in mammograms. Sādhanā 41(4), 385–395 (2016)
19. L. Cai, X. Wang, Y. Wang, Y. Guo, J. Yu, Y. Wang, Robust phase-based texture descriptor for classification of breast ultrasound images. Biomed. Eng. Online 14(1), 26 (2015)
20. J.Y. Choi, A generalized multiple classifier system for improving computer-aided classification of breast masses in mammography. Biomed. Eng. Lett. **5**(4), 251–262 (2015)
21. J. de Nazaré Silva, A.O. de Carvalho Filho, A.C. Silva, A.C. de Paiva, M. Gattass, Automatic detection of masses in mammograms using quality threshold clustering, correlogram function, and SVM. J. Digital Imaging 28(3), 323–337 (2015)
22. X. Liu, M. Mei, J. Liu, W. Hu, Microcalcification detection in full-field digital mammograms with PFCM clustering and weighted SVM-based method. EURASIP J. Adv. Signal Process. 2015(1), 73 (2015)
23. K.M. Prabusankarlal, P. Thirumoorthy, R. Manavalan, Assessment of combined textural and morphological features for diagnosis of breast masses in ultrasound. Human-Centric Comput. Inf. Sci. **5**(1), 12 (2015)
24. S. Sharma, P. Khanna, Computer-aided diagnosis of malignant mammograms using Zernike moments and SVM. J. Digital Imaging **28**(1), 77–90 (2015)
25. W.-J. Wu, S.-W. Lin, W.K. Moon, An artificial immune system-based support vector machine approach for classifying ultrasound breast tumor images. J. Digital Imaging **28**(5), 576–585 (2015)
26. A.T. Azar, S.A. El-Said, Performance analysis of support vector machines classifiers in breast cancer mammography recognition. Neural Comput. Appl. **24**(5), 1163–1177 (2014)
27. S.M.A. Beheshti, H. AhmadiNoubari, E. Fatemizadeh, M. Khalili, An efficient fractal method for detection and diagnosis of breast masses in mammograms. J. Digital Imaging **27**(5), 661–669 (2014)
28. H. Cai, L. Liu, Y. Peng, Wu. Yaopan, Li. Li, Diagnostic assessment by dynamic contrast-enhanced and diffusion- weighted magnetic resonance in differentiation of breast lesions under different imaging protocols. BMC Cancer **14**(1), 366 (2014)
29. M. Tan, J. Pu, B. Zheng, Optimization of breast mass classification using sequential forward floating selection (SFFS) and a support vector machine (SVM) model. Int. J. Comput. Assisted Radiol. Surg. **9**(6), 1005–1020 (2014)
30. A. García-Manso, C.J. García-Orellana, H.G. Velasco, R. Gallardo-Caballero, M.M. Macías, Consistent performance measurement of a system to detect masses in mammograms based on blind feature extraction, Biomed. Eng. Online 12(1), 2 (2013)
31. S. Hoffmann, J.D. Shutler, M. Lobbes, B. Burgeth, A. Meyer-Bäse, Automated analysis of non-mass-enhancing lesions in breast MRI based on morphological, kinetic, and spatial-temporal moments and joint segmentation-motion compensation technique. EURASIP J. Adv. Signal Process. 2013(1), 172 (2013)

32. F. Retter, C. Plant, B. Burgeth, G. Botella, T. Schlossbauer, A. Meyer-Bäse, Computer-aided diagnosis for diagnostically challenging breast lesions in DCE-MRI based on image registration and integration of morphologic and dynamic characteristics. EURASIP J. Adv. Signal Process. **2013**(1), 157 (2013)
33. U. Rajendra Acharya, E.Y.K. Ng, J.-H. Tan, S. Vinitha Sree, Thermography based breast cancer detection using texture features and support vector machine, J. Med. Syst. 36(3), 1503–1510 (2012)
34. J. Ding, H.D. Cheng, J. Huang, J. Liu, Y. Zhang, Breast ultrasound image classification based on multiple-instance learning, J. Digital Imaging 25(5), 620–627 (2012)
35. W. Jian, X. Sun, S. Luo, Computer-aided diagnosis of breast micro calcifications based on dual-tree complex wavelet transform. Biomed. Eng. Online **11**(1), 96 (2012)
36. R. Ramos-Pollán, M.A. Guevara-López, C. Suárez-Ortega, G. Díaz-Herrero, J.M. Franco-Valiente, M. Rubio-del-Solar, N. González-de-Posada, M.A.P. Vaz, J. Loureiro, I. Ramos, Discovering mammography-based machine learning classifiers for breast cancer diagnosis, J. Med. Syst. 36(4), 2259–2269 (2012)
37. F.S. Zakeri, H. Behnam, N. Ahmadinejad, Classification of benign and malignant breast masses based on shape and texture features in sonography images. J. Med. Syst. **36**(3), 1621–1627 (2012)
38. S.J.S. Gardezi, I. Faye, J.M.S. Bornot, K. Nidal, M. Hussain, Mammogram classification using dynamic time warping, Multimed. Tools Appl. 76(282), 1–22 (2017)
39. M. Shibusawa, R. Nakayama, Y. Okanami, Y. Kashikura, N. Imai, T. Nakamura, H. Kimura, M. Yamashita, N. Hanamura, T. Ogawa, The usefulness of a computer-aided diagnosis scheme for improving the performance of clinicians to diagnose non-mass lesions on breast ultrasonographic images. J. Med. Ultrason. **43**(3), 387–394 (2016)
40. S.A. Waugh, C.A. Purdie, L.B. Jordan, S. Vinnicombe, R.A. Lerski, P. Martin, A.M. Thompson, Magnetic resonance imaging texture analysis classification of primary breast cancer. Eur. Radiol. **26**(2), 322–330 (2016)
41. R. Hupse, M. Samulski, M. Lobbes, A. den Heeten, M.W. Imhof-Tas, D. Beijerinck, R. Pijnappel, C. Boetes, N. Karssemeijer, Standalone computer-aided detection compared to radiologists' performance for the detection of mammographic masses. Eur. Radiol. **23**(1), 93–100 (2013)
42. S.P. Singh, S. Urooj, An improved CAD system for breast cancer diagnosis based on generalized pseudo-zernike moment and Ada-DEWNN classifier. J. Med. Syst. **40**(4), 105 (2016)
43. A.T. Azar, S.A. El-Said, Probabilistic neural network for breast cancer classification. Neural Comput. Appl. **23**(6), 1737–1751 (2013)
44. A.D. Masmoudi et al., LBPV descriptors-based automatic ACR/BIRADS classification approach. EURASIP J. Image Video Process. **2013**(1), 19 (2013)
45. J. Dheeba, S.T. Selvi, A swarm optimized neural network system for classification of xzcvmicro calcification in mammograms. J. Med. Syst. **36**(5), 3051–3061 (2012)
46. J. Dheeba, S. Tamil Selvi, An improved decision support system for detection of lesions in mammograms using differential evolution optimized wavelet neural network. J. Med. Syst. 36(5), 3223–3232 (2012)
47. M. Dong, X. Lu, Y. Ma, Y. Guo, Y. Ma, K. Wang, An efficient approach for automated mass segmentation and classification in mammograms. J. Digital Imaging 28(5), 613–625 (2015)
48. M. Abdel-Nasser, J. Melendezb, A. Morenoa, O.A. Omerc, D. Puiga, Breast tumor classification in ultrasound images using texture analysis and super-resolution methods. Eng. Appl. Artif. Intell. 59, 84–92 (2017)
49. D.O. Tambasco Brunoa, M.Z. do Nascimentoa, R.P. Ramos, V.R. Batistaa, L.A. Neves, A.S. Martins, LBP operators on corselet coefficients as an algorithm to describe texture in breast cancer tissues. Expert Syst. Appl. 55, 329–340 (2016)
50. A. Gubern-Mérida, R. Martí, J. Melendez, J.L. Hauth, R.M. Mann, N. Karssemeijer, B. Platel, Automated localization of breast cancer in DCE-MRI. Med. Image Anal. **20**(1), 265–274 (2015)
51. Deep neural network https://www.analyticsvidhya.com/blog/2016/08/evolution-core-concepts-deep-learning-neural-networks/ Accessed 10 May 2021

52. A. Kumar, S.S.S.S. Reddy, V. Kulkarni, An object detection technique for blind people in real-time using deep neural network, in *2019 Fifth International Conference on Image Information Processing (ICIIP), Shimla, India* (2019), pp. 292–297. https://doi.org/10.1109/ICIIP47207.2019.8985965
53. A. Kumar, A review on implementation of digital image watermarking techniques using LSB and DWT, in *The Third International Conference on Information and Communication Technology for Sustainable Development (ICT4SD 2018), held during August 30–31, 2018 at Hotel Vivanta by Taj*, Goa, India (2018)
54. S.S.S. Reddy, A. Kumar, Edge detection and enhancement of color images based on bilateral filtering method using K-means clustering algorithm, in *ICT Systems and Sustainability. Advances in Intelligent Systems and Computing* edited by M. Tuba, S. Akashe, A. Joshi, vol. 1077 (Springer, Singapore, 2020)
55. A. Kumar, S.S.S. Reddy (eds.), Advancements in security and privacy initiatives for multimedia images. IGI Glob. (2021). https://doi.org/10.4018/978-1-7998-2795-5
56. A. Kumar, Design of secure image fusion technique using cloud for privacy-preserving and copyright protection. Int. J. Cloud Appl. Comput. (IJCAC) **9**(3), 22–36 (2019)

E-Learning Cloud and Big Data

Analysis of Blockchain Secure Models and Approaches Based on Various Services in Multi-tenant Environment

Pooja Dhiman and Santosh Kumar Henge

Abstract This research paper is focused on analyzing past proposed papers on using blockchain with cloud computing to enhance the present security system of the cloud. Homomorphic encryption technique is the latest and most secure way of protecting cloud data on the server. It allows computational operations on encrypted data as there is no need to decrypt the data for processing it. Still, we cannot guarantee privacy and data integrity. Here comes the concept of blockchain, it is a decentralized approach as compared to the cloud's centralized approach. Decentralization means data on the blockchain is stored at different locations or servers which is owned by different organizations. Hence, it is not dependent on any single third party for its execution or processing. The blockchain can be combined with fully or partially homomorphic encryption schemes to build a strong security system in a multi-tenant environment.

Keywords Blockchain (BC) · Cloud computing (CC) · Multi-tenancy (MT) · Homomorphic encryption (HE) · Hyperledger (HL) · Merkle tree (MT) · Fully homomorphic encryption (FHE)

1 Introduction

Cloud computing is a centralized approach. It stores the data on different servers located globally. Since the data is stored on remote servers, security is always the major concern. The fully Homomorphic Encryption technique (FHE) is considered to be the most secured technique as of now. It permits the processing of encrypted data, maintaining the privacy of data. But for some cases, there may be a need to share the public key with the cloud provider to perform some operations. If the provider is not a trusted party, it can manipulate the user's data. In this case, we can use Blockchain along with the FHE schemes to build a more secured network.

P. Dhiman (✉) · S. K. Henge
School of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India

S. K. Henge
e-mail: santosh.24372@lpu.co.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_42

Blockchain once created cannot be altered or modified. Cloud data stored on the blockchain is considered to be a safer technique. Data is stored in a decentralized form in different remote locations. If an attacker anyhow gets access to the cloud's data which is stored using the blockchain technique, he may get access to only a small part of the information or data, the rest of the data is stored on different blocks. And the accessed data is encrypted using the FHE scheme which is hard to crack.

In the case of Multi-tenant cloud architecture, blockchain allows to record transactions and communications between the tenants and the cloud provider using the smart contract technique.

To achieve decentralization in cloud storage, the Ethereum blockchain and CP-ABE (ciphertext policy attribute-based encryption) algorithm can be used. The communication between the cloud service provider and the user is done using Ethereum smart contract technology [8, 12, 14].

Wenlei Qu et al. concluded the use of blockchain in electronic voting using homomorphic encryption schemes. The Hyperledger method was used in blockchain. It uses a homomorphic signcryption algorithm to encrypt digital signatures, and smart contract technology is used for maintaining integrity in the votes. Blockchain makes the process public and transparent which cannot be altered or modified.

2 Paper Objective

This paper focuses on analyzing past proposed approaches on cloud security using blockchain to ensure better security and maintaining integrity. It discusses the Merkle tree implementation using Ethereum in a Multi-tenant system.

3 Blockchain Architecture for Multi-tenant Cloud Environment

In a Multi-tenant environment, each tenant can maintain their information using an individual permissioned blockchain. Data is divided into smaller parts or blocks in the blockchain database and encrypted using the hash technique. Each block has a previous hash value connecting to other blockchains. The consensus algorithm is used to verify the valid transactions or data. Merkle tree maintains the transactions of all the nodes using hashing or anchoring the nodes. The data is transmitted to each node by encrypted transactions and data using a hash function. Homomorphic encryption algorithms can be used for encryption purpose to ensure better security.

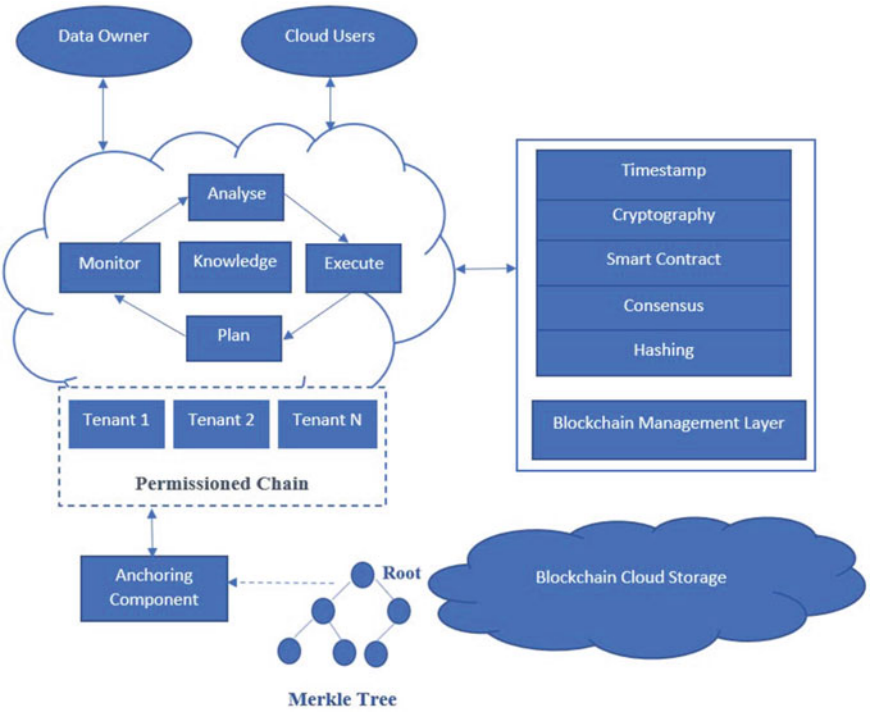


Fig. 1 Blockchain-based multitenant cloud architecture

When a new data block is inserted in the blockchain, block header and nonce (4-byte field) are used to measure cryptographic hash function SHA-256. Then it will check in the Merkle tree the anchored node, if it validates only then the new node can be included in the blockchain. Also, all the other nodes in the blockchain network must validate the new node. The new entry will be recorded in the Merkle tree. To maintain data integrity, smart contract technology is used. A smart contract is a self-execution program based on some condition that helps in building trust and data integrity.

In the architecture, Ethereum's Merkle-Patricia library is used for implementing the Merkle tree. All the tenants can be dependent on the cloud owner for the anchoring process [4]. A trigger is used to communicate with the anchor nodes and writing the data to the blockchain (see Fig. 1).

4 Analysis of Cloud Security Using Blockchain Technology

Author name and indexing	Title	Methodology	Problem taken	Techniques and security parameters used	Achievements	Lagging issues/future scope
Reantongcome et al. [8]	Securing and Trustworthy Blockchain-based Multi-Tenant Cloud Computing	Truffle framework is used with Ethereum to implement Blockchain	Co-resident attack in multi-tenancy caused due to data leakage by malicious tenant	Ethereum is used to implement Blockchain via Smart contract used in Truffle framework. which is coded in Solidity Parameter Implementation; Strong integrity of transactions logs, Integration of Smart Contract, Resource Allocation and Policy checking	Blockchain will help in recording transactions between cloud owner and tenants	For future scope, proposed model can be improved for better confidentiality and integrity
Park and Park [6]	Blockchain Security in Cloud Computing: Use Cases, Challenges, and Solutions	Surveys blockchain technology and finds solution for securing bitcoin and cloud	Ensuring user anonymity	Secure bitcoin protocol is used to install electronic wallet in the cloud and deletes the user details after using the service. It uses a public key for encryption purpose Parameter implementation; Electronic wallet, Two-factor authentication, Total currency, Longest chain wins	Verifies complete removal of the electronic wallet so as to ensure user anonymity	Bitcoin can be further studied for future risks
Wang et al. [12]	A Secure Cloud Storage Framework with Access Control based on Blockchain	This paper proposes better access control in cloud storage using Ethereum blockchain and CP-ABE (ciphertext-policy attribute-based encryption)	To improve access policy in cloud	Two operating systems are required for implementation purpose, Linux/Unix is used to deploy Ethereum blockchain and Windows 10 for the encryption process Parameter implementation; Smart contract, Valid access periods	Cost required in accessing the files is less. Decryption process can be done in valid access period which leads to better cloud security. Tracing of function has become easier	The centralized Cloud storage platforms can be replaced by decentralised technologies in future as it lacks in data integrity

(continued)

(continued)

Author name and indexing	Title	Methodology	Problem taken	Techniques and security parameters used	Achievements	Lagging issues/future scope
Weber et al. [14]	A Platform Architecture for Multi-Tenant Blockchain-Based Systems	To design a scalable multi-tenant architecture by evaluating quantitative and qualitative analysis and identifying the requirements for fulfilment	Data integrity and performance isolation among tenants	A proof-of-concept prototype is used for implementation purpose in Ethereum with Laava industry partner. Individual blockchain is allocated to each tenant and tenant chains root is used for anchoring Parameter implementation; Individual permissioned blockchains	Proposed architecture is able to achieve low cost, data integrity, isolation in performance of tenants	Anchored chains can be evaluated for better flexibility
Sukhodolskiy and Zapechnikov [10]	A Blockchain-Based Access Control System for Cloud Storage	Ensures cryptographic operations privacy using protocols without the participation of cloud owner	Access control on cloud	The proposed prototype is tested on Ethereum and implemented using smart contracts. CP-ABE (Ciphertext-policy attribute-based encryption scheme) is the used in access control mechanism. Parameter implementation; Hashing, Access policy assignment, Smart contract	Only hash code ciphertexts are transferred on blockchain ledger, hence ensures privacy	
Murthy et al. [2]	Blockchain Based Cloud Computing: Architecture and Research Challenges	To develop an integrated architecture by investigating survey on blockchain with a scalable cloud environment	Challenges faced by cloud in security, integrity, user data management	Discussed various issues in cloud and suggested solution by using blockchain. Compares different blockchain platforms and explains consensus algorithms	Some of the advantages using blockchain with cloud platform can be better security, maintaining trust, better data management, scalability and usability	Blockchain can further be studied and applied in practical with cloud platform

(continued)

(continued)

Author name and indexing	Title	Methodology	Problem taken	Techniques and security parameters used	Achievements	Lagging issues/future scope
Shah et al. [9]	Decentralized Cloud Storage Using Blockchain	Focuses on decentralized data storage, utilization of data, security and privacy using IPFS (InterPlanetary File System) protocol	Decentralized data, Data security, utilization of resources, privacy	IPFS is used to store the AES encrypted file on the number of peers in the network. Details about file is stored in the blockchain using Smart contracts. Cryptocurrency is transferred to peer's wallet from user's wallet Parameter implementation; IPFS protocol, Hashing	Centralized approach in cloud is not secure as compared to the decentralized approach in blockchain	
Qu et al. [7]	A electronic voting protocol based on blockchain and homomorphic signcryption	Proposes a technology combining blockchain and homomorphic signcryption for electronic voting	Security and privacy issue in electronic voting	To maintain privacy of votes in the ballot and to count the votes, Homomorphic signcryption technique is used. Hyperledger blockchain type is used. Smart contract is used to aggregate the voting results. It also verifies the valid and invalid votes	Proposed model ensures safe and reliable voting, improving the efficiency of voting and reducing the total time required in the process. Transparency of data is achieved by blockchain. Third party is not required in this process	Research can be done on the proposed model to implement it practically with improved security
Yaji et al. [16]	Privacy Preserving in Blockchain based on Partial Homomorphic Encryption System for AI Applications	Proposed a technique to preserve the privacy in the blockchain using partial homomorphic encryption schemes	Attacks on wallet, collision attack reimage attack in the blockchain	The PHE schemes Goldwasser-Micali and Paillier are used for encryption process in the blockchain to make it stronger and more secure. Processing time required is very less and the proposed model can defeat most of the attacks	Proposed scheme is more secured and takes less time	Different attacks can be experimented on the blockchain and improved by using modified and improved encryption schemes in the blockchain technology

(continued)

(continued)

Author name and indexing	Title	Methodology	Problem taken	Techniques and security parameters used	Achievements	Lagging issues/future scope
Yahaya et al. [15]	A Blockchain based Privacy-Preserving System for Electric Vehicles through Local Communication	Proposed a privacy preserving algorithm for finding a potential supplier for demander in Electric vehicles	Energy trading between EVs demander and supplier	P2P communication is used for searching the suitable supplier while partial homomorphic encryption is used for implementation. Verification process for energy transmission is done by Blockchain Parameter implementation; Private Blockchain	The proposed model is faster than the Bichromatic Mutual Nearest Neighbor (BMNN) algorithm	Implementation of the proposed model can be done practically on hardware to optimize its performance
Kumar et al. [5]	BMIAE: blockchain-based multi-instance Iris authentication using additive ElGamal homomorphic encryption	Proposed a privacy algorithm based on blockchain multi-instance iris authentication using ElGamal Homomorphic Encryption (BMIAE)	Focussed mainly on malicious attacks on untrusted server	In BMIAE, smart contract is used to perform the distance compute on the encrypted templates. Elgamal security is based on the hardness of solving the discrete problems Parameter implementation; Smart contract	The proposed model guarantees confidentiality and integrity. Computational cost and execution time is improved	

5 Past Proposed dependency parameters in Existing Secure Systems

Existing parameters are not sufficient to support present security system in Multi-tenant environment so we have concluded some advanced parameters suitable for advanced systems or future technology systems. The following dependable parameters are proposed and implicated in existing secure system.

- **Strong integrity of transactions logs**: Blockchain is an integrated secure system with proper auditing and well-formed transactions. Consensus is used to verify the transactions for inclusion in a Blockchain.
- **Integration of smart contract**: It helps in building the trust between the two parties by making a contract in the form of a computer code developed in Ethereum's programming language, Solidity. It is stored in a decentralized unchangeable form in a public Blockchain.
- **Policy checking**: For using a resource, tenants request the request listener which classifies the request based on the business type requirement and forwards the

request to the policy checking smart contract. Its duty is to verify and validate the request [8, 17].

- **Resource allocation**: A robust blockchain based decentralized framework for managing resources can minimize the energy consumed by using smart contract technique. After policy checking process, the smart contract validates the request and send it to allocation service. For allocating a resource, server then shares the available virtual machines to allocation contract [8].

6 Integration of Blockchain Based Advanced Secure Parameters in Multi-tenant Environment

The advanced secure parameters such as Token based Salting, Key based salting, Session based salting, token-based resource allocation, and Token based policy checking can help in generating unique and customized keys for the required resource and services.

7 Conclusion

The blockchain technology can be combined with a partial or fully homomorphic encryption scheme by adding some advanced parameters which are mostly based on tokens, to ensure better privacy and security. The Ethereum or Hyperledger methods in blockchain are used to store the data on the cloud which is encrypted using homomorphic encryption schemes. Smart contract technology is used for data integrity and trustworthiness. Existing parameters are not sufficient to support present security system in Multi-tenant environment so we have concluded some advanced parameters suitable for advanced systems or future technology systems. Using tokens with salting, policy checking and session-based salting are some of secured parameters which can help in generating unique and customized keys for the required resource and services. Further research can be done to implement this technique in practical.

References

1. J. Bethencourt, A. Sahai, B. Waters, Ciphertext-policy attribute-based encryption, in *Proceedings of the Symposium on Security and Privacy* (2007), pp. 321–334. https://doi.org/10.1109/SP.2007.11
2. C.H.V.N.U.B. Murthy, M.L. Shri, S. Kadry, S. Lim, Blockchain based cloud computing: Architecture and research challenges. IEEE Access **8**, 205190–205205 (2020). https://doi.org/10.1109/ACCESS.2020.3036812
3. M. Dameron, Beigepaper: an ethereum technical specification, pp. 1–25 (2018) https://github.com/chronaeon/beigepaper/blob/master/beigepaper.pdf

4. N. Helil, K. Rahman, CP-ABE access control scheme for sensitive data set constraint with hidden access policy and constraint policy. Secur. Commun. Netw. **2017** (2017). https://doi.org/10.1155/2017/2713595
5. M.M. Kumar, M.V.N.K. Prasad, U.S.N. Raju, BMIAE: blockchain-based multi-instance Iris authentication using additive ElGamal homomorphic encryption. IET Biometrics **9**(4), 165–177 (2020). https://doi.org/10.1049/iet-bmt.2019.0169
6. J.H. Park, J.H. Park, Blockchain security in cloud computing: Use cases, challenges, and solutions. Symmetry (Basel) **9**(8), 1–13 (2017). https://doi.org/10.3390/sym9080164
7. W. Qu, L. Wu, W. Wang, Z. Liu, H. Wang, A electronic voting protocol based on blockchain and homomorphic signcryption. Concurr. Comput. (2019), 1–17 (2020). https://doi.org/10.1002/cpe.5817
8. V. Reantongcome, V. Visoottiviseth, W. Sawangphol, A. Khurat, S. Kashihara, D. Fall, Securing and trustworthy blockchain-based multi-tenant cloud computing, in *ISCAIE 2020 IEEE 10th Symposium on Computer Applications & Industrial Electronics* (2020), pp. 256–261. https://doi.org/10.1109/ISCAIE47305.2020.9108796
9. M. Shah, M. Shaikh, V. Mishra, G. Tuscano, Decentralized cloud storage using blockchain, in *Proceedings of the 4th International Conference on Trends in Electronics and Informatics, ICOEI 2020*, no. Icoei (2020), pp. 384–389, https://doi.org/10.1109/ICOEI48184.2020.9143004
10. I. Sukhodolskiy, S. Zapechnikov, A blockchain-based access control system for cloud storage, in *Proceedings of the 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering ElConRus 2018*, vol. 2018-January (2018), pp. 1575–1578. https://doi.org/10.1109/EIConRus.2018.8317400
11. S. Tanwar, Q. Bhatia, P. Patel, A. Kumari, P.K. Singh, W.C. Hong, Machine learning adoption in blockchain-based smart applications: the challenges, and a way forward. IEEE Access **8**, 474–448 (2020). https://doi.org/10.1109/ACCESS.2019.2961372
12. S. Wang, X. Wang, Y. Zhang, A secure cloud storage framework with access control based on blockchain. IEEE Access **7**, 112713–112725 (2019). https://doi.org/10.1109/ACCESS.2019.2929205
13. B. Waters, Ciphertext-policy attribute-based encryption: an expressive, efficient, and provably secure realization. Lecture Notes Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 6571 LNCS, no. subaward 641 (2011), pp. 53–70. https://doi.org/10.1007/978-3-642-19379-8_4
14. I. Weber, Q. Lu, A. B. Tran, A. Deshmukh, M. Gorski, M. Strazds, A platform architecture for multi-tenant blockchain-based systems, in *Proceedings of the 2019 IEEE International Conference on Software Architecture ICSA 2019*, no. January 2021 (2019), pp. 101–110. https://doi.org/10.1109/ICSA.2019.00019
15. A.S. Yahaya, N. Javaid, R. Khalid, M. Imran, N. Naseer, A blockchain based privacy-preserving system for electric vehicles through local communication, in *IEEE International Conference on Communications*, vol. 2020-June (2020). https://doi.org/10.1109/ICC40277.2020.9149129
16. S. Yaji, K. Bangera, B. Neelima, Privacy preserving in blockchain based on partial homomorphic encryption system for Ai applications, in *Proceedings of the 25th IEEE International Conference on High Performance Computing Work . HiPCW 2018* (2019), pp. 81–85. https://doi.org/10.1109/HiPCW.2018.8634280
17. D. Unal, M. Hammoudeh, M.S. Kiraz, Policy specification and verification for blockchain and smart contracts in 5G networks. ICT Express **6**(1), 43–47 (2020). https://doi.org/10.1016/j.icte.2019.07.002

Harmonic Minimization in Multilevel Inverters Using Ant Lion Optimization Algorithm

Tushar Tyagi, Amit Kumar Singh, Himanshu Sharma, and Rintu Khanna

Abstract Total Harmonic Distortion minimization in multilevel inverters is a strenuous optimization problem as it includes the solution of non-linear transcendent equations having multiple minima's. With the increase in the number of levels in multilevel inverters, the optimization task becomes more tedious. Several nature-inspired algorithms have been used in past to address the solution of the Total Harmonic Distortion optimization problem. The main of objective of this paper is to test the newly developed ant lion optimization algorithm for minimization of harmonics in multi-level inverter. In this paper ant lion optimization algorithm, which is inspired by the hunting mechanism of ant lion is used to optimize the harmonic content in multilevel inverters output voltage. In the proposed work 5, 7, 9, 11, 13 and 15 level inverters are considered for optimization. The comparative analysis has been performed with the Whale Optimization algorithm and an updated variant of grey wolf optimization for the eleven-level inverter. From the comparative analysis, it has been evaluated that the Ant-Lion optimization algorithm is best suitable for the Total Harmonic Distortion minimization problems in multi-level inverters.

Keywords Total harmonic distortion · Multi-level inverter · Random Walk · Ant lion · Optimization

T. Tyagi (✉) · A. K. Singh · H. Sharma
Lovely Professional University, Phagwara, Punjab, India
e-mail: tushar.20586@lpu.co.in

A. K. Singh
e-mail: amit.20267@lpu.co.in

H. Sharma
e-mail: himanshu.23441@lpu.co.in

R. Khanna
Punjab Engineering College (Deemed to be University), Chandigarh, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_43

1 Introduction

In the recent years multi-level inverter has gained lot of interest in the field of power electronics. In terms of inverter topologies, diode-clamped, capacitor-clamped and cascaded H-bridge inverters are the most common type of Multi-level inverter (MLI) found in the literature. Based on the modulation MLI was divided into two categories from the standpoint of strategies. Fundamental switching frequency and high switching frequency are the two major classes. The basic switching frequency group includes Selective Harmonic Elimination (SHE) and space vector control, while the second group includes high switching frequencies, pulsed width modulation strategies (PWM), space vector, and multilevel sinusoidal. The cascaded H-bridge multilevel topology is used in this research, as well as the selective harmonic elimination (SHE) control technique [1, 2]. In a cascaded H-Bridge multilevel topology, various H-bridge inverter systems with different dc sources are associated in cascade or sequence. SHE aims to zero out low-order harmonics while keeping the fundamental variable at the idealistically [3–7].

In SHE to eliminate the lower order harmonics, a series of non-liner equations need to be solved for switching angles of the power electronics devices of MLI. A variety of techniques have been used to solve this series of equations, stretching from evolutionary algorithms such as Newton–Raphson [8] to simulation models such as Genetic Algorithms (GA) [9–11] and Particle Swarm Optimization (PSO) [12, 13]. Every one of those methods, however, has disadvantages that limit their application in selective harmonic removal. The Newton–Raphson approach is not suggested for solving problems with a large range of angles, such as in multilevel inverters, and GA has a slow convergence rate when compared to other approaches. PSO [14], on the other hand, is a swarm-based meta-heuristic with a high convergence rate but stagnation while aiming for local minima. Some of these problems have been solved using meta-heuristics including the bee algorithm [15], ant colony [16], modified version of the fish algorithm [17], and firefly algorithms [18], with better results than PSO. The SHE method has recently been updated with the whale optimization algorithm (WOA) [19], an updated version of grey wolf optimization (MGWO) [20], and the Black Widow Optimization Algorithm (BWOA) [21]. Recently a modified version of WOA has been used for SHE problem in 5 and 7 level inverter [22].

In this paper ant lion optimization [23] algorithm developed by S. Mirjalili is used for the elimination of harmonics in multilevel inverters. The SHE technique is applied to five, seven, nine, eleven, thirteen and fifteen level inverter. The switching angles are determined by the solution of the equations using ant lion algorithm for minimum total harmonic distortion, to validate the results the switching angles are fed to the simulation model of MLI in MATLAB Simulink and THD is calculated by taking Fourier transform of the output waveform. The obtained results are compared with Black Widow Optimization Algorithm (BWOA) [21] Whale optimization algorithm (WOA) [21] and an updated variant of grey wolf optimization (MGWO) [21] for eleven level inverter.

2 Ant Lion Optimization Algorithm

The ant lion optimisation algorithm is based on the larval hunting process of ant lions. They dig conical pits by moving along a rounded path scattering out sand. After digging the trap, it conceal itself at the bottom staking out for the prey (ants) which falls easily as the edge of the conical pit is sharp. If an ant falls into the trap, ant lion catch the prey and if it try to escape, it intelligently pushes the sand towards the edge preventing its prey to leave the pit. After eating the prey, it throws the leftovers out and modify the pit for next hunt.

2.1 Operators of ALO Algorithm

ALO algorithm models the antlions method of catching ants. Since ants move in stochastic manner, a random walk is defined for each ant by the given equation:

$$X(t) = [0,\ \text{csum}(2r(j_1) - 1),\ \text{csum}(2r(j_2) - 1), \ldots\ldots,\ \text{csum}(2r(j_n) - 1)] \quad (1)$$

where csum indicates the collective sum, n indicates number of iterations, j indicates the step of the random walk and $r(j)$ is a random function defined as follows:

$$\overrightarrow{r(t) = \left\{ \begin{matrix} 1,\ if\ rand > 0.5 \\ 0,\ if\ rand \leq 0.5 \end{matrix} \right\}}, \quad (2)$$

Figure 1 shows example of 3 random walks over 500 iterations.

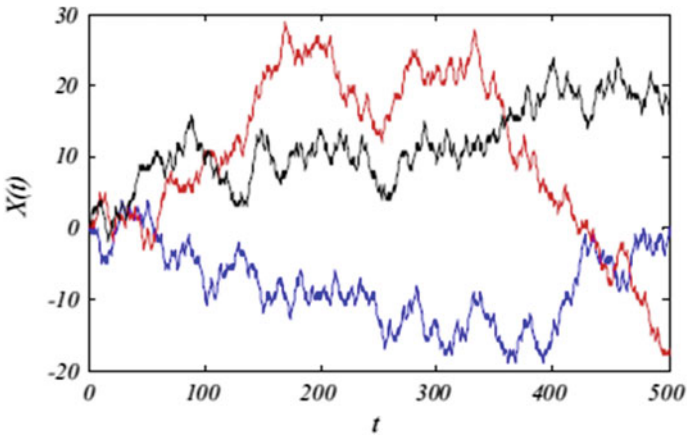


Fig. 1 Random walks [22]

The location of ants and ant lions is stored in a matrix as shown and is used for optimisation process later.

$$N_{\text{Ant}} = \begin{bmatrix} B_{1,1} & B_{1,2} & \ldots\ldots & B_{1,d} \\ B_{2,1} & B_{2,2} & \ldots\ldots & B_{2,d} \\ : & : & : & : & : \\ : & : & : & : & : \\ B_{n,1} & B_{n,2} & \ldots\ldots & B_{n,d} \end{bmatrix} \tag{3}$$

Here N_{Ant} indicates the location of each ant, $B_{x,y}$ shows the value of the y-th variable (dimension) of x-th ant, n is the number of ants, and d is the number of dimensions.

$$N_{\text{Antlion}} = \begin{bmatrix} BL_{1,1} & BL_{1,2} & \ldots\ldots & BL_{1,d} \\ BL_{2,1} & BL_{2,2} & \ldots\ldots & BL_{2,d} \\ : & : & : & : & : \\ : & : & : & : & : \\ BL_{n,1} & BL_{n,2} & \ldots\ldots & BL_{n,d} \end{bmatrix}. \tag{4}$$

Here N_{antlion} is the antlion position, $B_{x,y}$ shows the y-th dimension's value of x-th ant lion, n is the number of antlions, and d is the number of dimensions.

For calculating the fitness of each ant lion and ant, an objective function is used and correspondingly the values are stored in a matrix.

$$N_{\text{OB}} = \begin{bmatrix} f([B_{1,1,}B_{1,2,}\ldots, B_{1,d}]) \\ f([B_{2,1,}B_{2,2,}\ldots, B_{2,d}]) \\ : \\ : \\ f([B_{n,1,}B_{n,2,}\ldots, B_{n,d}]) \end{bmatrix}, \tag{5}$$

where N_{OB} is matrix that contains the fitness value evaluated for each ant with the objective function, $BL_{i,j}$ indicates j-th dimension of i-th ant.

$$N_{\text{OBL}} = \begin{bmatrix} f([BL_{1,1,}BL_{1,2,}\ldots, BL_{1,d}]) \\ f([BL_{2,1,}BL_{2,2,}\ldots, BL_{2,d}]) \\ : \\ : \\ f([BL_{n,1,}BL_{n,2,}\ldots, BL_{n,d}]) \end{bmatrix}, \tag{6}$$

where N_{OBL} is the matrix containing the fitness values of all antlions, $BL_{i,j}$ indicates j-th dimension's value of i-th antlion, f is the objective function and n is the total number of iterations.

2.2 Random Walks of Ants

Ants change their positions with random walk at every stage of optimization. As each searching area has a limit, random walks inside the search space are normalized using the following equation:

$$Y_m^k = \frac{(Y_m^k - b_m) \times (e_m - d_m^k)}{(e_m^k - b_m)} + d_m, \tag{7}$$

where b_m is the least value of m-th variable, e_i is the highest value in m-th variable, $Y^k{}_m$ is the least value of m-th variable at k-th iteration, and $d^k{}_m$ indicates the highest value of m-th variable at k-th iteration.

2.3 Trapping in Ant Lion Pits

As ants move randomly, their stochastic movement is affected by the trap of the antlion and this assumption is modelled by the following equations:

$$R_i^k = al_j^k + r^k \tag{8}$$

$$S_i^k = al_j^k + s^k \tag{9}$$

where r^t is least variable at k-th iteration, s^t indicates the vector including highest variable at k-th iteration, ${R_I}^t$ is the least variable for k-th ant, ${s_i}^t$ is highest variable for ith ant, and ${al_j}^t$ indicates location of j-th antlion selected randomly at k-th iteration.

2.4 Building Trap

Ant lions hunting ability is determined by roulette wheel operator selection. Since ants are supposed to get into the trap of one selected ant lion as shown in Fig. 2, the ant lion which is considered as fittest from the objective function is selected by using this operator.

2.5 Sliding Ants Towards Ant Lion

As ant falls into the trap, antlions slides it down to prevent its escape by pushing sand towards to the edge. This can be modelled by gradually reducing the radius of the random walk of that particular ant by the following equations:

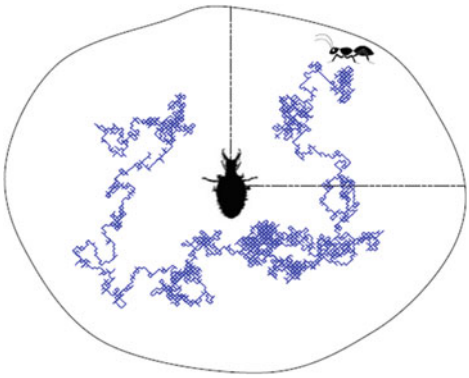


Fig. 2 Random walk of ant inside an ant lion trap [22]

$$a_t = \frac{a_t}{M} \tag{10}$$

$$b_t = \frac{b_t}{M} \tag{11}$$

where a_t is the local minima at k-th iteration, and b_t indicates the local maxima at k-th iteration and M is a ratio defined as follows:

$$M = 10^R \frac{k}{K} \tag{12}$$

where k is the current iteration, K is the maximum number of iterations, and R is a constant that determines the exploitation capabilities of the ALO algorithm and is equal to values that depends on the current and maximum iterations which is shown as follows:

($R = 2$ when $k > 0.1$ K, $R = 3$ when $k > 0.5$ K, $R = 4$ when $k > 0.75$ K, $R = 5$ when $k > 0.9$ K, and $R = 6$ when $k > 0.95$ K).

2.6 Trapping and Catching Ant and Re-Building the Pit

Finally, when an ant is caught by an antlion, it is eaten, and this process is imitated by the presumption that prey is caught when ants become more fit than their corresponding ant lion. The ant lion then shifts to the position of the captured ant for its next hunt, knowing that it will have a good chance of catching its next prey.

So,

$$\text{antlion}_m^k = \text{ant}_n^t \; if \; f\left(\text{ant}_n^k\right) > f\left(\text{antlion}_j^k\right)$$

where k shows the current iteration, $\text{antlion}_m{}^k$ shows the location of the antlion m selected randomly at k-th iteration, and $\text{ant}_n{}^k$ shows the location of the ant n selected at k-th iteration.

2.7 Finding an Elite

Finding an elite is an important trait of these algorithms that makes them to retain the best solution obtained in each level during the optimization. The antlion which gives best solution in each stage is an elite that affects the random walks of ants in each iteration. So, the random walks of ant is influenced by both the position of the antlion which was selected by the roulette wheel operator and as well as the elite which is modelled by the following equation:

$$\text{Ant}_i^t = \frac{R_A^t + R_E^t}{2} \tag{13}$$

where $R_E{}^t$ is ant's random walk around an elite antlion at k-th iteration, $R_A{}^t$ is ant's random walk around an antlion which is randomly selected by the roulette wheel at t-th iteration, and $\text{ant}_i{}^t$ is the position of i-th ant at k-th iteration.

3 Problem Formulation

3.1 Total Harmonic Distortion Equation

Initially, THD equation is considered as the Primary Factor and the goal here is to minimize this THD to least possible value. This is done by the optimization algorithm as it explores and exploits the Search Space for the better Switching Angles which in turn corresponds to Minimal Levels of THD.

THD is given by

$$\text{THD} = \frac{\sqrt{\sum_{i=2}^{\infty} V_{i_rms}^2}}{V_{1_rms}} \tag{14}$$

The modulation index is given by

$$M = \frac{4nV_{DC}}{\pi m}, \tag{15}$$

where n represents the number of H-Bridges, m is the modulation index and V_{DC} is the DC voltage level.

3.2 Voltage Equation

Secondly, the Voltage THD, also whose minimization forms the Secondary Focus, as the lower levels of Voltage THD corresponds to an improved Power Quality. This is linked to the Power Quality attribute that Voltage Quality is directly equal to Power Quality. A better explanation to this would be that in any Power System, Voltage is the only electrical property that can be precisely controlled irrespective of the load and on the other hand Control of Current is practically not plausible as Current is a dependent quantity and varies as per the Load requirements and is subject to change every now and then.

The voltages are considered from Fundamental (V_1) to a frequency of 49 (V_{49}) and is given by

$$\begin{bmatrix} V_1 \\ V_3 \\ . \\ . \\ V_{49} \end{bmatrix} = \begin{bmatrix} \frac{4}{\pi V_{DC}\left(\sum_{x=1}^{15}\cos(n(x)+\ldots\cos(49(x))\right)} \\ \frac{4}{3\pi V_{DC}\left(\sum_{x=1}^{15}\cos(n(x)+\ldots\cos(49(x))\right)} \\ . \\ . \\ \frac{4}{49\pi V_{DC}\left(\sum_{x=1}^{15}\cos(n(x)+\ldots\cos(49(x))\right)} \end{bmatrix}, \tag{16}$$

where n represents the number of H-Bridges, and V_{DC} is the DC voltage level.

The Voltage THD equation is as follows:

$$V_{_\mathrm{THD}} = \left| M - |V_1| + \left| \sum_{x=3}^{49} V_X \right| \right|, \tag{17}$$

where M = Modulation Index.

3.3 Harmonic Frequencies

The Third aspect is the calculation of all the harmonic frequencies present in the Voltage Waveform. This enables us to calculate the THD and gives us an insight to the contribution of each Harmonic Frequency to the total THD. Here, the calculations are performed up to 49th order of the Fundamental frequency, i.e., 50 Hz. A valid explanation for this is that, as the order of harmonics escalates, their contribution to

the overall THD drops down exponentially. This means lower order harmonics are more predominant than their higher order counterparts.

All the equations representing the Frequency Spectrum is given by

$$\begin{bmatrix} f_1 \\ f_3 \\ . \\ . \\ f_{49} \end{bmatrix} = \begin{bmatrix} \left(\sum_{x=1}^{15} \cos(n(x) + \ldots \cos(n(x))\right) \\ \left(\sum_{x=1}^{15} \cos(3n(x) + \ldots \cos(3 \times n(x))\right) \\ . \\ . \\ \left(\sum_{x=1}^{15} \cos(n(x) + \ldots \cos(49 \times n(x))\right) \end{bmatrix}. \tag{18}$$

The THD due to the Harmonic Frequencies is given by

$$\text{THD}_f = \left[\frac{\sum_{i=3}^{49} \sqrt{f_n}}{\sqrt{i} \times f_1}\right]^2, \tag{19}$$

where f_1 is Fundamental Frequency.

3.4 Fitness Function

Finally, linking all the above three aspects is the Final Fitness Function which is a combination of all the three above mentioned equations.

$$F = V_{_\text{THD}} + 100 \times \text{THD}_f + K \times \text{Penalty}, \tag{20}$$

where

$$\text{Penalty} = |C_1| + |C_2|, \tag{21}$$

and

$$C_1 = -n(1), \tag{22}$$

$$C_2 = M - |V_1|, \tag{23}$$

where

C_1 and C_2 are Type Constraint,

$K = 100$: Penalty on each constraint violation.

4 Results and Discussion

The ant lion optimization algorithm is implemented to find the optimal values of switching angles for 5, 7, 9, 11, 13 and 15 level inverter. The modulation index is changed from 0.1000 to 0.9999 with a step increment of 0.0100 to find out the minimum THD for a given level. At each modulation index the algorithm ran 50 times with population size of 50 and 500 iterations. The best modulation index for minimum THD and the corresponding switching angles fed to the Simulink model of inverter in MATLAB. To validate the result obtained by the optimization the THD of the output voltage waveform obtained after the simulation is evaluated using the Fast Fourier analysis tool available in MATLAB Simulink. The result obtained for different levels are presented here. The computational environment is as follows.

- Coding Platform: MATLAB 2016a
- Runs: 50 and Iterations: 500
- Simulation Platform: Mat Lab Simulink 2016a
- Type of Inverter Bridge used: Universal H Bridge
- Harmonic analysis tool: FFT analysis tool in POWER GUI.

Similarly as mentioned from Tables 1, 2 and 3 the optimal switching angles and THD are calculated for 11, 13, and 15 level. The overall best result is shown in Table 4 with comparative analysis.

It is evident from Table 4 that ALO outperformed the BWOA, WOA and MGWOA for eleven level inverter in terms of minimum THD values and the optimal range of the modulation index is from 0.63 to 0.85 for different level inverter.

Table 1 Five level inverter optimal switching angles and THD

Modulation index	THD anltlion (%)	THD by simulation (%)	Switching angles in degrees
0.13	27.0975	28.48	27.3, 89.9
0.29	27.102	28.48	27.3, 89.9
0.36	27.102	28.48	27.3, 89.9
0.46	26.2816	27.59	24.1, 89.5
0.51	28.4807	30.86	18.1, 85.9
0.7	23.9456	23.98	16.4, 63.8
0.8	16.0916	17.42	16.2, 50.2
0.86	**14.6115**	**16.44**	**12.378. 41.94**
0.91	16.3701	17.88	10.7, 33.1

Table 2 Seven level inverter optimal switching angles and THD

Modulation index	THD antlion (%)	THD by simulation (%)	Switching angles in degrees
0.18	20.7428	22.78	14.3869, 49.876, 89.4788
0.27	15.0576	18.64	14.4843, 41.6082, 90
0.39	15.0288	18.34	8.5313, 33.7132, 72.9833
0.49	9.8366	10.63	8.961, 30.51, 59.805
0.57	9.8093	10.12	8.004, 27.095, 47.486
0.63	**9.5942**	**11.57**	**8.319, 28.519, 48.834**
0.78	11.5658	13.12	9.7171, 23.422, 45.447
0.83	11.3148	13.97	7.1562, 20.5348, 46.8794

Table 3 Nine level inverter optimal switching angles and THD

Modulation index	THD antlion (%)	THD by simulation (%)	Switching angles in degrees
0.18	9.5739	12.62	18.3976, 36.5146, 52.557, 82.8783
0.24	10.0877	13.16	8.5313, 25.8633, 42.4848, 78.8676
0.39	7.1564	10.28	7.9182, 20.615, 35.775, 57.1754
0.48	**6.7415**	**9.4**	**7.305, 22.616, 39.345, 59.558**
0.55	7.6314	10.78	6.835, 19.933, 41.516, 59.782
0.62	7.2750	10.36	6.371, 18.4034, 35.758, 55.748
0.78	7.3465	10.48	6.39, 20.414, 1.516, 59.782
0.84	7.6318	10.71	4.824, 17.681, 35.609, 52.173

Table 4 Minimum THD values of 5, 7, 9, 11, 13 and 15 level inverter

Level	Modulation index	THD ant lion	THD simulation	Comparison		
5	0.86	14.6115	16.44	NA		
7	0.63	9.5942	11.57	NA		
9	0.48	6.7415	9.4	NA		
11	0.85	4.9245	7.65	5.01 (BWOA) [21]	5.56 (WOA) [21]	5.71 (MGWOA) [21]
13	0.76	5.6910	6.85	NA		
15	0.65	3.6397	5.47	NA		

5 Conclusion and Future Scope

The presented work introduces the ant lion optimization algorithm for the SHE problem of multi-level inverters. The algorithm is tested on different level inverters like 5, 7, 9, 11, 13 and 15, and it is concluded that ALO outperformed the existing algorithms mentioned in the literature for the minimization of the THD and have reduced it to 5.01% which is 9.8% reduction as compared to WOA. The ant lion algorithm is computationally efficient, consistent and providing near optimal solutions even at higher number of levels. The results obtained are validated by simulation of the inverters in MATLAB using the switching angles generated by ALO for minimum THD, and they are in close agreement with the THD calculated and THD by simulation. However the minuscule difference is due to consideration of harmonic order in case of THD provided by the FFT analysis tool as it considering high order harmonics as well, but in case of THD calculated, the maximum harmonic is considered till 49th only. In future ant lion optimization algorithm can be applied in various fields like economic emission dispatch and real world optimization problems The ant lion algorithm can also be applied for the optimization of Z-source inverters for the better tuning of shoot through duty ratio and THD.

References

1. H.S. Patel, R.G. Hoft, Generalized techniques of harmonic elimination and voltage control in thyristor inverters: part I—harmonic elimination. IEEE Trans. Ind. Appl. **3**, 310–317 (1973)
2. H.S. Patel, R.G. Hoft, Generalized techniques of har-monic elimination and voltage control in thyristor inverters: part II—voltage control techniques. IEEE Trans. Ind. Appl. **10**(5), 666–673 (1974)
3. J. Rodriguez, L.G. Franquelo, S. Kouro, J.I. Leon, R.C. Portillo, M.A.M. Prats, M.A. Perez, Multilevel converters: an enabling technology for high-power applications, in *Proceedings of IEEE*, vol. 97, no. 11 (2009), pp. 1786–1817
4. Y. Liu, H. Hong, A.Q. Huang, Real-time calculation of switching angles minimizing THD for multilevel inverters with step modulation. IEEE Trans. Industr. Electron. **56**(2), 285–293 (2009)
5. J. Kumar, B. Das, P. Agarwal, Selective harmonic elimination technique for a multilevel inverter, in *Proceedings of the Fifteenth National Power Systems Conference (NPSC), Mumbai, India* (2008 December), pp. 608–613
6. H.R. Massrur, T. Niknam, M. Mardaneh, A.H. Rajaei, Harmonic elimination in multilevel inverters under unbalanced voltages and switching deviation using a new stochastic strategy. IEEE Trans. Industr. Inf. **12**(2), 716–725 (2016)
7. A. Ajami, M.R.J. Oskuee, A.O. Mokhberdoran, Implementation of novel technique for selective harmonic elimination in multilevel inverters based on ICA. Adv. Power Electron. **20**, 1310 (2013)
8. J. Sun, H. Grotstollen: Solving nonlinear equations for selective harmonic eliminated PWM using predicted initial values, in *Proceedings of the 1992 International Conference on Industrial Electronics, Control, Instrumentation, and Automation, San Diego, CA, USA* (1992), pp. 259–264
9. B. Ozpineci, L.M. Tolbert, J.N. Chiasson, Harmonic optimization of multilevel converters using genetic algorithms. IEEE Power Electron. Lett. **3**(3), 92–95 (2005)

10. S.S. Lee, B. Chu, N.R.N. Idris, H.H. Goh, Y.E. Heng, Switched-battery boost-multilevel inverter with GA optimized SHEPWM for standalone application. IEEE Trans. Industr. Electron. **63**(4), 2133–2142 (2016)
11. K. El-Naggar, T.H. Abdelhamid, Selective harmonic elimination of new family of multilevel inverters using genetic algorithms. Energy Convers. Manage. **49**(1), 89–95 (2008)
12. W. Razia Sultana, S.K. Sahoo, S. Prabhakar Karthikeyan, I.J. Raglend, P.H. Vardhan Reddy, G.T. Rajasekhar Reddy, Elimination of harmonics in seven-level cascaded multilevel inverter using particle swarm optimization technique, in *Artificial Intelligence and Evolutionary Algorithms in Engineering Systems* edited by L.P. Suresh, S.S. Dash, B.K. Panigrahi (Springer India, New Delhi, India, 2015), pp. 265–274
13. H. Taghizadeh, M. Tarafdar Hagh, Harmonic elimination of cascade multilevel inverters with nonequal DC sources using particle swarm optimization. IEEE Trans. Ind. Electron. 57(11), 3678–3684 (2010)
14. Y. Zhang, S. Wang, G. Ji, A comprehensive survey on particle swarm optimization algorithm and its applications. Math. Problems Eng. 2015 (2015)
15. A. Kavousi, B. Vahidi, R. Salehi, M.K. Bakhshizadeh, N. Farokhnia, S.H. Fathi, Application of the bee algorithm for selective harmonic elimination strategy in multilevel inverters. IEEE Trans. Power Electron. **27**(4), 1689–1696 (2012)
16. S.D. Patil, S.G. Kadwane, S.P. Gawande, Ant colony optimization applied to selective harmonic elimination in multilevel inverters, in *Proceedings of the 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Bengaluru, India* (2016), pp. 637–640
17. K.P. Panda, S.S. Lee, G. Panda, Reduced switch cascaded multilevel inverter with new selective harmonic elimination control for standalone renewable energy system. IEEE Trans. Ind. Appl. **55**(6), 7561–7574 (2019)
18. M.G. Sundari, M. Rajaram, S. Balaraman, Application of improved firefly algorithm for programmed PWM in multilevel inverter with adjustable DC sources. Appl. Soft Comput. **41**, 169–179 (2016)
19. S. Mirjalili, A. Lewis, The whale optimization algorithm. Adv. Eng. Softw. **95**, 51–67 (2016)
20. S. Mirjalili, S.M. Mirjalili, A. Lewis, Grey wolf optimizer. Adv. Eng. Softw. **69**, 46–61 (2014)
21. A.F. Peña-Delgado, H. Peraza-Vázquez, J.H. Almazán-Covarrubias, N. Torres Cruz, P.M. García-Vite, A.B. Morales-Cepeda, J.M. Ramirez-Arredondo, A novel bio-inspired algorithm applied to selective harmonic elimination in a three-phase eleven-level inverter. Math. Problems Eng. (2020)
22. A.K.V.K. Reddy, K.V.L. Narayana, Optimal total harmonic distortion minimization in multilevel inverter using improved whale optimization algorithm. Int. J. Emerg. Electr. Power Syst. 21(3) (2020)
23. S. Mirjalili, The ant lion optimizer. Adv. Eng. Softw. **83**, 80–98 (2015)

Examine the Indian Tweets to Determine Society Emphasis on Novel Corona-Viruses (COVID-19)

Anil Kumar Dubey, Mala Saraswat, Raman Kapoor, and Rishu Gupta

Abstract The recent emerging coronavirus as novel corona virus (2019-nCoV) formed viral pneumonia based emergency not only in Wuhan but also Europe, Iran, North Korea, India, and many more countries. WHO has already declared this situation as pandemic of corona virus. The world has around twenty-four lakhs cases in the whole world with around one and half lakhs deaths as on 20 April 2020. The corona virus is the family member of Nidovirales, which occur in human body from animal-human interaction. Here, we deliver the basics of corona virus and illustrate the social impact of this emergency. The review will aid the knowledge of CoV with their family and understand the person for healthy life. Here we study the Indian Tweets to determine the people's emphasis on emerged Novel Coronaviruses (COVID-19), also compute the comparative tweets as concern of corona virus especially for Indian capital region for last six months January to June 2020 and January to March 2021 and find out the tactics of tweets for peoples concern about it.

Keywords Tweet · Virus · Coronavirus · nCoV · India · Delhi · Mumbai

1 Introduction

The Corona virus [1, 2] is an imperative pathogen to human body. It may infect the human respiratory system, gastrointestinal tract, hepatic, and central nervous system (CNS) of human body through Livestock, bats, mouse, birds, and several other wild animals. Corona virus is the RNA enveloped viruses, belong to Coronaviridae family. The two basic respiratory syndrome named as: Severe acute respiratory syndrome

A. K. Dubey (✉) · M. Saraswat · R. Kapoor · R. Gupta
ABES Engiineering College, Ghaziabad, Uttar Pradesh 201009, India

M. Saraswat
e-mail: mala.saraswat@abes.ac.in

R. Kapoor
e-mail: raman.kapoor@abes.ac.in

R. Gupta
e-mail: rishu.gupta@abes.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_44

(SARS) and Middle East respiratory syndrome (MERS), confirmed the bidirectional transmission of emerging Corona virus from animal to human and vice versa [4]. The mystery of newly formed corona virus in Wuhan, China (December 2019), illustrate the concentration of world around them [12–14]. The newly known corona virus in Wuhan declared as Novel corona virus (2019-nCoV) by WHO dated on 12 January 2020. The Coronaviruses (CoV) is the subfamily member of Corona virinae, within family of Coronaviridae in order to Nidovirales. Four generation of corona viruses are as follows: Alpha coronavirus, Beta coronavirus, Gamma coronavirus, and Delta coronavirus [3]. In 2002, about eight thousand persons infected through SARS-CoV in China, among these near about eight hundred died. As of now already the number of deaths around the world from the novel coronavirus cases has crossed more than one and half lakhs. Earlier only six types of corona viruses are known among several, and now one more as Novel corona viruses adding them and known list become to seven. As per the WHO, the MERS-CoV visible in human due to contact of infected dromedary camels.

2 Functions of Structural and Non-structural Proteins in CoV

Generally, the non-structural proteins (nsps) are synthesized with little volumes of infected cells [5]. There are sixteen non-structural proteins concern to corona virus as nsps1 to nsps16, in which three are unknown for their functioning as: nsps2, nsps7, and nsps11, remaining thirteen functioning as: nsps1 (coding region) specified to appreciate the viral RNA (ribonucleic acid), essential to effective viral repetition, ruin the cellular mRNA. The nsps2 is not recognized till and noted as unknown functioning (see Fig. 1).

The nsps3 (papain) is accountable for the cleavage from positions-1 to position-3 to ripen three non-structural proteins, indorsing the cytokine expression, and block the host innate immune responses. Fourth non-structural proteins nsps4 is responsible for transmembrane domain, viral replication, and membrane proliferation. The nsps5 is concern with viral survival and 3CLpro cleaves, impeding IFN signaling. Nsps6, Transmembrane domain, concern to the interaction with two nsps as nsp3 and nsp4. Nsps7 is the cofactor of nsp8 and nsp12. Similarly, nsps8 have the cofactor with nsp7 and nsp12 and responsible for Primase activity [6, 7]. The nsps9 is the essential for binding activity of protein and dimer with RNA/DNA. Nsps10 concern to scaffold protein of nsp14 and nsp16. The nsps11 is not clearly identified till date. The nsps12, RNA-dependent RNA polymerase, is responsible for viral replication and transcription. The nsps13 disturbs the tropism and virulence and belong to superfamily of helicase. The nsps14 perform the exoribonuclease activity in viral replication. Nsps15 do the dodging of dsRNA sensors, N7-methyltransferase. The last nsps16 confirm the methyltransferase inhibition and negatively regulating innate immunity [8]. Coronaviruses genomes is encrypted with five structural proteins as

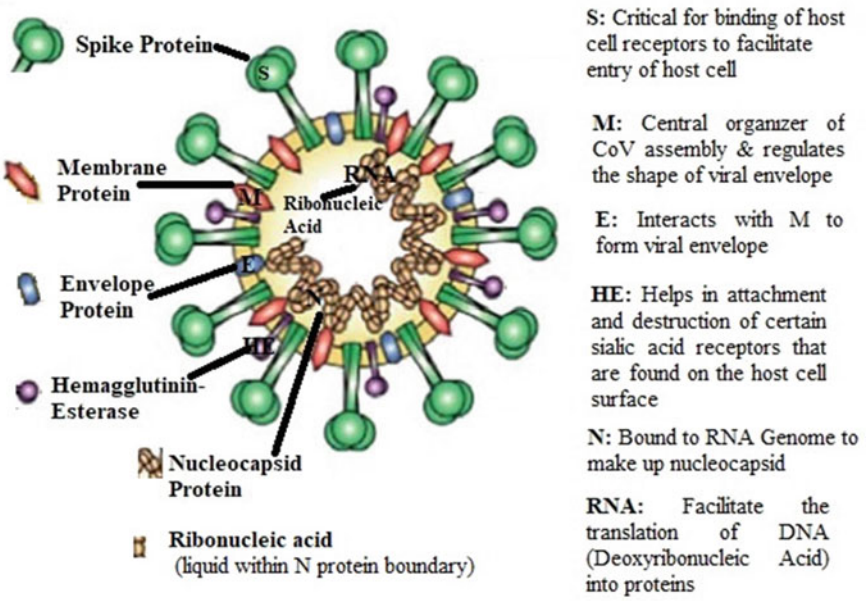


Fig. 1 Structural proteins in Coronavirus virion

shown in above figure. Where S Glycoproteins is positioned outside of virion and facilitate them for typical shape. The Spike proteins form homotrimers (composed protein of three identical polypeptide units), consent to construct morphologies to produce Coronaviruses. The M Glycoproteins contains the three transmembrane sections, which play an important role in the regenerating virions within the cells. The E Glycoproteins are the small proteins contained 76 to 109 amino acids. In coronavirus, the envelope proteins show the complex characteristics into muster and morphogenesis for virions inside the cells. The nucleocapsid proteins are phosphoproteins that proficient to bind the helix, having the lithe structure to the pathological genomic of ribonucleic acid, and shows the significant characteristics in the replication and transcription of coronaviruses.

3 Virus Origin

Several Corona virus exhibitions assorted multitude collection and tissue tropism [19–23]. Generally, the alpha and beta coronaviruses are responsible for the infection of human with interaction to mammals, camels. While the gamma and delta coronaviruses can cause infection to human from birds and fish, and also mammals. Before introducing the newly known novel corona virus, previous six known corona viruses are as: HcoV-229E, HcoV-NL63, HcoV-OC43, and HKU1. The novel corona

virus is possibly initiated by beta category of corona virus because the virus structure consist the changes occurred from beta category (see Table 1).

4 Global Effects

Several CoV have caused the diversity of diseases in birds and mammals fluctuating via pigs and cows, upper respiratory infection in chickens. The emerging corona virus 2019-nCoV, belongs to beta type of coronaviruses, may infect the lower respiratory tract, causes the pneumonia in human body, having slighter symptoms than SARS-CoV and MERS-CoV. Till January 2020, about to four hundred cases have found in China, in which more than seventy percent cases in Wuhan and some are in Beijing, Shanghai, Guangdong. Cases also predicted in other counties like India, Thailand, South Korea, Japan, US, etc. In most of the cases the patients had stayed or visited to Wuhan. Generally, patients had interacted with Seafood Wholesale Market in Wuhan, where there is sale of diverse forms of meat. As per the statistics the novel corona virus (2019-nCoV) will spread via man to man transmission due to interaction of infected people with close contact. As per WHO report dated on December 31, 2020, more than 82 million people were infected through COVID-19, in which about to 1.8 million peoples were dead globally. But pilot study compute that approximately the dead excess up to 3 million due to COVID-19 with direct or indirect impact. It means greater than the only sixty percent record is recorded for globally COVID-19 impact as official figure reported to death and remaining forty percent was missing [10]. Due to the higher global impact of corona virus to human health as well as wealth, citizens are afraid and discuss to difference parameters of COVID-19 via several transmission approaches as social media messaging, tweets [24–29], tiny audio/video and face to face communication etc.

5 Symptoms and Safety

The emerging novel coronavirus, 2019-nCoV, concern to breathe infection in human body [16–18] and having the following symptoms as: headache, fever, running nose, to feel sick, cough, sore throat, lungs infection, feeling tired, and sneeze. The medical research statistics illustrate that the corona virus infected patients have more than seventy percent chances of fever and cough. Up to twenty percentages may have the occurrence of shortness of breath and muscle ache. And remaining will have symptoms of confusion, headache, sore throat, chest pain, diarrhea, and nausea and vomiting, bilateral pneumonia. Till now, researchers are doing the investigation for vaccine to control it, but no single antiviral therapy for corona virus is designed. Since no vaccine till date, the best approach to safe the human body is to control the infection sources, early diagnosis, isolation, caring treatments, etc. The individuals are trying to prevent themselves from corona virus infectious through cover them salve from

Table 1 Characteristic of Beta-CoVs [9–11, 15]

S. no.	Beta- CoV	Non structured proteins	Spike protein (length of amino acids)	Structural proteins	Symptoms	Origin	Number of patients (globally)	Possible natural reservoir
1	SARS-CoV	At least 5	1255	4	• Fever of 100.5 F or higher • Dry cough • Shortness of breath	Guangdong province, China	8098	Bat
2	MERS-CoV	16	1353	4	• Fever • Cough • Shortness of breath • May be Pneumonia	Arabian Peninsula	2254	Bat
3	N-COV (Covid-19)	Expected 5	Till not finalize	Till not finalize	• Fever • Cough • Shortness of breath • Breathing difficulties • May be pneumonia	Wuhan, china	120,754,630+	May be wild animals Snakes, bat etc

fitted mask, avoiding crowded places, and circumvent get sick, pneumonia, cold and also do not touch the eye, nose, and mouth without proper soap wash.

6 Result and Discussion

In this section, author discusses the concerns of society in India especially for corona disease and their tweets from January to June 2020 for capital of India as Delhi (including NCR location) and Mumbai are as mentioned below. As per the statistics of tweets, tweeted from January 2020 to June 2020 in Delhi, NCR, after comparing the statistics of both, we found that the peoples are more impressed with viruses in March month, is greater than the previous consecutive months. The individual words tweets from January to June 2020 specially for Delhi citizen with compared to complete tweets of these words are commuted and found that Coronavirus words was more tweeted in compare to COVID-19, CoVs, and nCoV as shown in Figs. 2 and 3.

The COVID-19 related tweets are increasing up to 9% in March months from January and 8% from February, its decrease May and June by 5 to 7%. The WHO had declared name of novel corona virus as "2019-nCoV" for society pronunciation, and application in several computation eras. Most of the 2019-nCoV associated tweets posted in January month and it continuously reduces next to next month. The individual words tweets from January to June 2020 specially for Mumbai citizen with compared to complete tweets of these words are commuted and found that Coronavirus words was more tweeted in compare to COVID-19, CoVs, and nCoV as shown in Figs. 4 and 5.

Up to 47% of tweets of CoVs, in Mumbai is occurred in March month and 34% of nCoV is tweeted in January of 2020 (Figs. 6, 7, and 8).

Comparing first three months of tweet for corona related keywords for both Delhi and Mumbai in 2020 and 2021 (up to 20 March 2021), we found that approximately

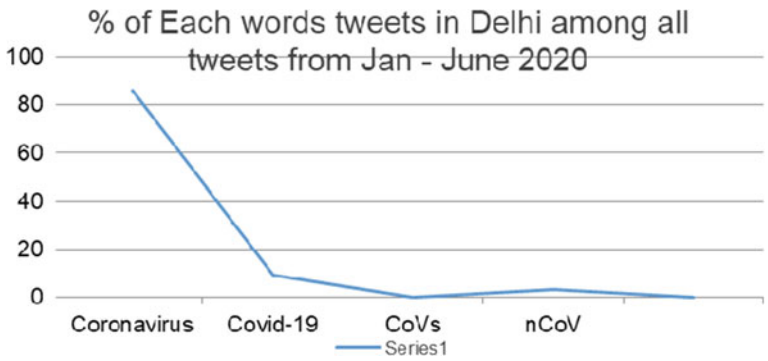


Fig. 2 Depict Delhi tweets of each word versus all tweets from January to June 2020

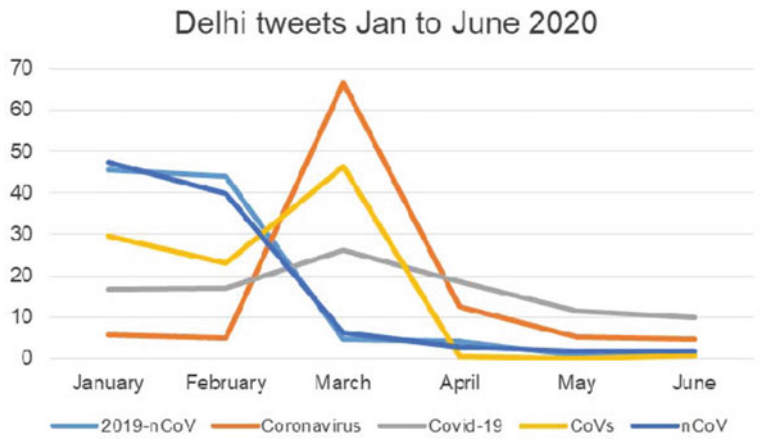


Fig. 3 Depict Delhi tweets from January to June 2020

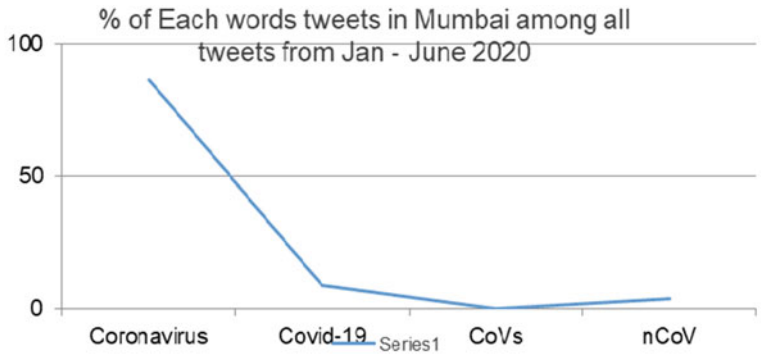


Fig. 4 Depict Mumbai tweets of each word versus all tweets from January to June 2020

three times more tweet of '2019-nCoV', tweeted by Mumbai Indians in compare to last year March tweets. Similarly, 'Coronavirus ' and 'COVID-19' tweet are continually increasing in Mumbai in compare to March 2020 tweets. The current year shows approximately two one and half times increases the tweet of '2019-nCoV', 'COVID-19'), tweeted by Delhi people's in compare to March 2020.

Comparing these two prominent city of India: Delhi and Mumbai for Corona related tweets from last six month January to June and found that approximately fifty-two percent of 'Novel Corona 'among 3,605,125 tweets, seventy-three percent of 'MERS' among 116,831 tweets, and sixty-four percent of 'SARS' among 333,654 tweets, is tweeted by Mumbai Indians. Delhi Indian is tweeted up to hundred percent of tweet in March for 'coronavirus'.

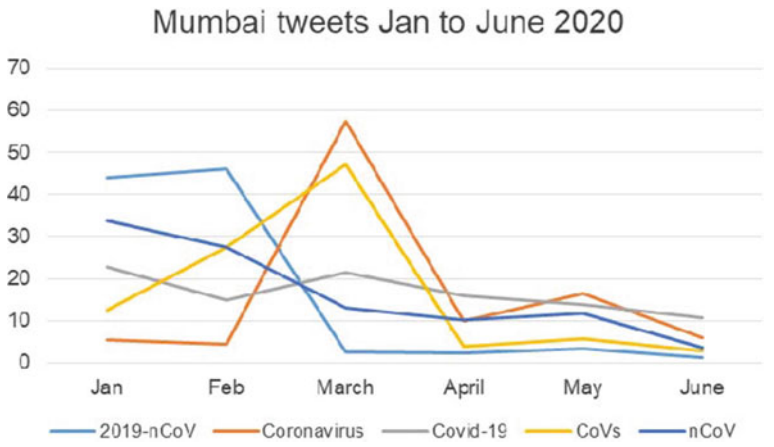


Fig. 5 Depict Mumbai tweets from January to June 2020

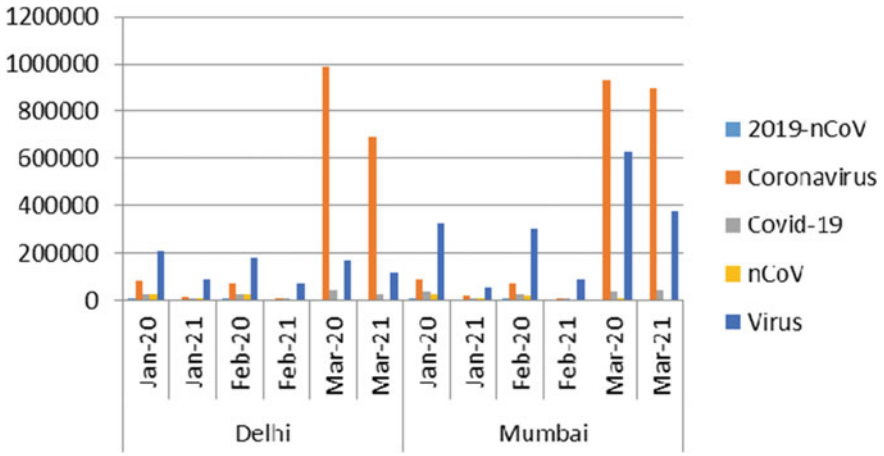


Fig. 6 Depict the comparative tweets of 2020 and 2021 for first three months for both Delhi and Mumbai

7 Conclusion

Viruses affect the normal behavior of body cells. Day to day as new medicines are introducing, the different characteristics of viruses are also developing. Recently world health organization declare the emerging for health issues related to coronavirus, a novel corona virus (2019-nCoV) initialize from Wuhan, central city of China. It is the family member of Nidovirales, having similar characteristics as occurred in human through animal. In this paper, authors illustrate the basic concept of newly initiated coronavirus and global impact. The review will aid the knowledge

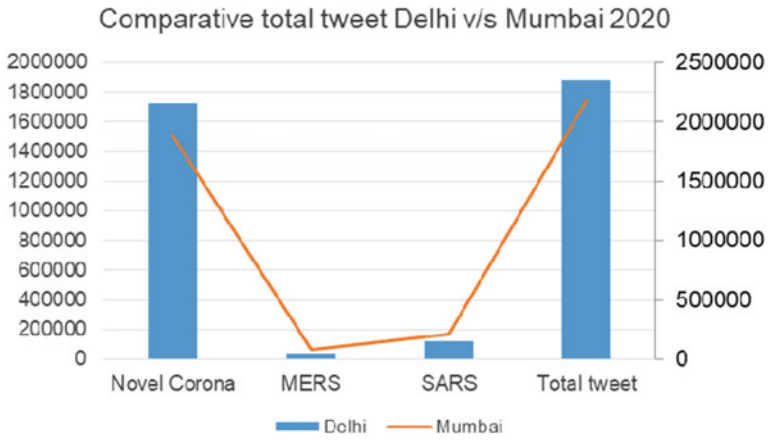


Fig.7 Depict the comparative tweets from January to June 2020 with total tweet

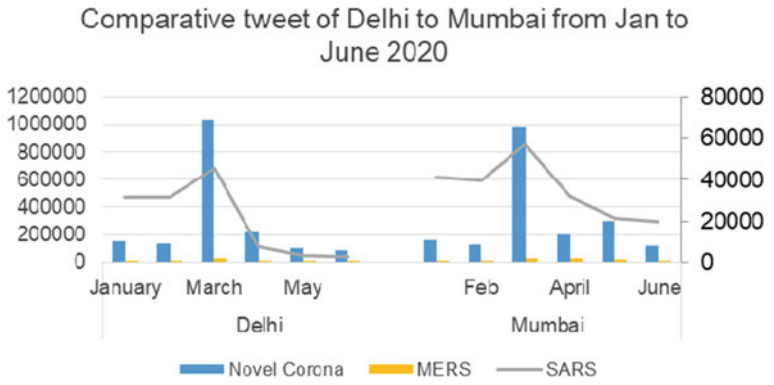


Fig. 8 Depict the comparative tweets from January to June 2020 for Delhi to Mumbai

of CoV with their family and understand the person focus on it. After analyzing the tweets of Delhi, NCR, and Mumbai for corona related keywords, we found that the current year tweet is increasing continually in compare to last year tweets for March, which means peoples of these two metro cities of India again feared for the corona disease. After analyzing these area tweets, it is clearly visible that Mumbai peoples are suffering with CoV physically as well as mentally and they are awaking others with the help of social media. From the result we can conclude that Mumbai, Delhi and after that NCR area is effecting with this virus and all are talking about the detailed symptoms of this virus. In perspectives of future, we will further analyze the social media tweets to find the impact of society emphasis and then we can differentiate the current scenario with previous year scenario. It will be helpful to identify the society mental thinking about the virus.

Acknowledgements ABES Engineering College, Ghaziabad, India is acknowledged for some of the facilities utilized in this research.

Conflict of Interest Statement The authors declared that there is no conflict of interest regarding this paper.

References

1. B.A. Johnson, R.L. Graham, V.D. Menachery, Viral metagenomics, protein structure, and reverse genetics: Key strategies for investigating coronaviruses. Virology **517**, 30–37 (2018)
2. F. Li, Structure, function, and evolution of coronavirus spike proteins. Annu. Rev. Virol. **3**, 237–261 (2016)
3. Y. Chen, Q. Liu, D. Guo, Emerging coronaviruses: genome structure, replication, and pathogenesis. J. Med. Virol. (2020)
4. S. Cauchemez, M.D. Van Kerkhove, S. Riley, C.A. Donnelly, C. Fraser, N.M. Ferguson, Transmission scenarios for Middle East Respiratory Syndrome Coronavirus (MERS-CoV) and how to tell them apart. Eurosurveillance 18(24). (2013). Bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin
5. T.T. Tok, G. Tatar, Structures and functions of coronavirus proteins: molecular modeling of viral nucleoprotein (2017)
6. Y.L. Siu, K.T. Teoh, J. Lo, C.M. Chan, F. Kien, N. Escriou, R. Bruzzone, The M, E, and N structural proteins of the severe acute respiratory syndrome coronavirus are required for efficient assembly, trafficking, and release of virus-like particles. J. Virol. **82**(22), 11318–11330 (2008)
7. C.A. De Haan, L. Kuo, P.S. Masters, H. Vennema, P.J. Rottier, Coronavirus particle assembly: primary structure requirements of the membrane protein. J. Virol. **72**(8), 6838–6850 (1998)
8. Y.H. Li, C.Y. Hu, N.P. Wu, H.P. Yao, L.J. Li, Molecular characteristics, functions, and related pathogenicity of MERS-CoV proteins. Engineering (2019)
9. Z. Song, Y. Xu, L. Bao, L. Zhang, P. Yu, Y. Qu, C. Qin, From SARS to MERS, thrusting coronaviruses into the spotlight. Viruses **11**(1), 59 (2019)
10. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200224-sitrep-35-covid-19.pdf
11. H.W. Chang, H.F. Egberink, R. Halpin, D.J. Spiro, P.J. Rottier, Spike protein fusion peptide and feline coronavirus virulence. Emerg. Infect. Dis. **18**(7), 1089 (2012)
12. N. Chen, M. Zhou, X. Dong, J. Qu, F. Gong, Y. Han, T. Yu, Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet (2020)
13. H. Lu, C.W. Stratton, Y.W. Tang, Outbreak of pneumonia of unknown ethology in Wuhan China: the mystery and the miracle. J. Med. Virol. (2020)
14. D.S. Hui, E.I. Azhar, T.A. Madani, F. Ntoumi, R. Kock, O. Dar, Zumla, A. (2020). The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—the latest 2019 novel coronavirus outbreak in Wuhan, China. Int. J. Infect. Dis. 91, 264–266
15. C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, Z. Cheng, Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet (2020)
16. D. Schoeman, B.C. Fielding, Coronavirus envelope protein: current knowledge. Virol. J. **16**(1), 69 (2019)
17. I. Seah, X. Su, G. Lingam, Revisiting the dangers of the coronavirus in the ophthalmology practice (2020)
18. S.L. Liu, L. Saif, Emerging viruses without borders: the Wuhan coronavirus (2020)
19. P.O. Gourinchas, Ş. Kalemli-Özcan, V. Penciakova, Covid-19 And Smes: A 2021 "Time Bomb" (2021)

20. E. Robinson, A. Jones, M. Daly, International estimates of intended uptake and refusal of COVID-19 vaccines: a rapid systematic review and meta-analysis of large nationally representative samples
21. E. Robertson, K.S. Reeve, C.L. Niedzwiedz, et al., Predictors of COVID-19 vaccine hesitancy in the UK household longitudinal study (2021)
22. B. MacKenna, H.J. Curtis, C.E. Morton, et al., Trends, regional variation, and clinical characteristics of COVID-19 vaccine recipients: a retrospective cohort study in 23.4 million patients using OpenSAFELY (2021)
23. C.A. Martin, C. Marshall, P. Patel, et al., Association of demographic and occupational factors with SARS-CoV-2 vaccine uptake in a multi-ethnic UK healthcare workforce: a rapid real-world analysis (2021)
24. S. Kumar, R. Kumar, M. Haider, A. Dubey, A comparative study of classifier algorithms for Twitter's sentiment based spam detection, in *1st International Conference on Computational Research and Data Analytics (ICCRDA 2020), 1, India*, vol. 1022, no. 1, IOP Conference on Series: Materials Science and Engineering (2020). https://doi.org/10.1088/1757-899X/1022/1/012016.
25. A.K. Dubey, M. Saraswat, R. Gupta, Swear tweets impact on society. Test Eng. Manag. J. 83(4), 17529–17539 (2020). ISSN: 0193-4120
26. P. Gupta, S. Kumar, R.R. Suman, V. Kumar, Sentiment analysis of lockdown in India during COVID-19: a case study on twitter. IEEE Trans. Comput. Soc. Syst. (2020)
27. U. Naseem, I. Razzak, M. Khushi, P.W. Eklund, J. Kim, COVIDSenti: a large-scale benchmark twitter data set for COVID-19 sentiment analysis. IEEE Trans. Comput. Soc. Syst. (2021)
28. A.K. Dubey, M. Saraswat, R. Gupta, SVC: swear words violate the Honored civilization. Int. J. Innov. Technol. Exploring Eng. (IJITEE) 8(7), 1601–1604 (2019). ISSN: 2278-3075
29. R. Harakawa, M. Iwahashi, Ranking of importance measures of tweet communities: application to keyword extraction from COVID-19 tweets in Japan. IEEE Trans. Comput. Soc. Syst. (2021)

Real-Time Rendering with OpenGL and Vulkan in C#

Dávid Szabó and Zoltán Illés

Abstract Real-Time rendering applications surrounds us in our everyday life, and this area of software just keep growing with newer systems, devices, and technologies. Starting from a large desktop PC down to small, even embedded devices, almost every computer device contains a graphics processor. These graphics units are programmable using Graphics APIs, and usually we use these libraries from C or C++ thanks to their low-level capabilities and because of the need for high performance. We want to present the starting steps of replacing the C or C++ language with .NET C# for developing multi-platform real-time graphical applications. Using the modern .NET environment, we can use Graphics APIs for rendering onto common .NET UI Frameworks while consuming all our previously implemented C# libraries and .NET technologies in the same application. To maintain compatibility with multiple platforms we are developing a library system to be able to use different Graphics APIs from the same C# source-code. In this paper, we are proposing some methods and considerations for implementing a library to be able to use the Vulkan and OpenGL APIs through a single C# codebase. We provide solutions for multi-platform rendering onto UI and dealing with the low-level challenges of using the two deeply different APIs to be able to deliver our unique real-time graphics into C# applications.

Keywords .NET C# · Real-Time · Vulkan · OpenGL

1 Introduction

With the modern .NET C# ecosystem, we are capable to build to multiple platforms including regular operating systems like Windows, Linux, or Mac [1], to mobiles like iOS or Android [2], and to other kind of platforms like embedded devices or

D. Szabó (✉) · Z. Illés
Eötvös Loránd University, Budapest, Hungary
e-mail: sasasoft@inf.elte.hu

Z. Illés
e-mail: illes@inf.elte.hu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_45

to the web. However, if You want to embed a real-time rendering module in Your application the possibilities are limited. There are no libraries or technologies in C# which You can use for implementing rendering solutions on the level of Graphics APIs in a multi-platform way. With Graphics APIs like Vulkan, DirectX, or OpenGL (which are basically C/C+ + libraries and interfaces) we can control the Graphics Processor in our device and use it for rendering or compute algorithms. The APIs themselves can be used in C# through binding libraries, but none of them provides complete compatibility with all the supported .NET platforms on their own. We need the combined use of these APIs to support as many platforms and devices as possible.

We have started the implementation of a library which presents a way for developing multi-platform real-time rendering applications using C#. The goal of the library is to provide a set of classes and functions for implementing low-level rendering algorithms in a cross-platform .NET application. The rendering code shall be implemented only once in a shared library, and all supported platforms can reference and consume this single codebase. Therefore, the library is an abstraction of Graphics APIs (currently Vulkan with vk.net and OpenGL with OpenTK). Based on the current platform and device we can choose which API we want to use in the compilation.

With our library's approach we aim to deliver a solution for writing low-level real-time rendering C# code in a shared library that can be consumed in multiple platforms during compilation. It will feature the manual initialization of resources like the Graphics Device and the Swap-Chain. Providing the possibility of writing custom shaders and render passes for custom rendering solutions. The data for these rendering algorithms can be uploaded to the GPU from C# through manually managed Buffers and Textures. The resulting frames of the rendering can be embedded into well-known UI frameworks like Windows Forms, WPF, Xamarin, and more. We are planning to implement a headless mode without a graphics presentation module in our library to provide the possibility of adding GPU Compute features to .NET applications as well. There are still many things left to be implemented in our library, but the starting steps have been done, the simplest rendering algorithms and their initialization can be already developed with it. On the following sections we present some of the implementation details, considerations, and challenges of developing such a library.

2 Graphics APIs in C#

2.1 OpenGL

OpenGL is Khronos' well-known old Graphics API which is still used in many applications, and most of the hardware supports some version of it. Its newest version at the time of writing is 4.6 released in 2017. OpenGL ES is a special version of OpenGL for mobile devices, and WebGL is developed for webpages. OpenGL is probably the easiest Graphics API to learn because with its simple commands

it delegates a lot of tasks to the GPU driver. Nevertheless, it is still a low-level API which will cause challenges when using it from a managed language like C#. OpenGL is supported on many regular operating systems like Windows, Linux, and Mac, though its support is decreasing especially on Mac thanks to its age. On Android and on embedded devices like the Raspberry Pi we can use OpenGL ES which is a lightweight version of OpenGL developed for smaller devices with smaller feature set. OpenGL and OpenGL ES's API set is similar, but OpenGL has more features, and some APIs may be under different function names or may accept the same parameters in a different format.

One of OpenGL's biggest disadvantage is that the Graphics Context (the access to the GPU and its state) can be active only on one execution thread. While we can change the active thread during runtime, this is inconvenient, dangerous, and very slow to do; therefore, we can say that we cannot use the OpenGL API in parallel, from multiple threads. OpenGL does not use Command Buffers directly, and its Rendering Pipeline executes based on the current state of the Graphics Context which is a big difference compared to Vulkan. Also, OpenGL hides the creation and configuration of the Swap-Chain (the resource to be able to render image sequences onto UI). The Swap-Chain is created by the underlying UI in binding libraries, and the Graphics Context is somewhat tied to it. We shall consider these differences during the development of the abstract library.

The most widespread binding library for using OpenGL in C# is the open-source OpenTK library [3]. OpenTK is consumable in .NET Framework and thanks to its 4.0 release .NET Core 3.1 is supported as well. However, to support .NET Standard [4] as a shared library we need to use a separate clone of OpenTK called OpenTK .NetStandard [5]. We can reference both libraries in a single .NET Standard project by editing the .csproj file directly. With the use of Conditions in the project file we can specify which package we want to reference for which platform and this way the build system will use them only during compilation of their supported platforms. This way the official OpenTK package will be used for .NET Core applications, and the third-party .NET Standard version is going to be compiled for other platforms like .NET Framework or Android.

2.2 Vulkan

Vulkan is the new Graphics API developed by Khronos and released in 2016 as a possible replacement for OpenGL. Vulkan fixes many of previously mentioned design flaws of OpenGL, and its specification is much more precise, developers have greater control over the GPU. There is no global state, and GPU resources can be used parallelly from multiple threads by compliance with some rules. The Swap-Chain creation is completely the developers' job and the tasks of the Rendering Pipeline, and the state transitions of the resources are strictly controlled by the developers.

Having the mentioned fine-grained management and more detailed control over the GPU renders the use of Vulkan much more complex and even more low-level

than OpenGL. Some of the state management and GPU memory access methods can be tricky to implement in C# due to the required low-level memory and pointer handling. Also, we need to develop a lot of features that in OpenGL were handled by the GPU driver and this will be a challenge to do in C# in a similarly performant way.

There are some C# binding libraries available for Vulkan. Usually, the main differences between these bindings are the version of the Vulkan specification that was used to generate the bindings, naming and namespace conventions and the way of parameter passing in the generated functions. We have used jpbruyere's vk.net package [6] because its way of handling pointers and structures in the generated library is the most compatible with our project's concept.

3 Implemented Architecture

3.1 Management, Device, and View

We made the structure of the library like Vulkan's structure. There is no global GPU state or Context, each resource is created through a parent resource, and therefore You must build up a tree of resources. At first You need to create a Management instance to start working with one of the APIs. We create the Management instance through a shared Factory class, and this is where we can decide about the used API. The Management will be created for the selected API, and all the resources we create through this instance will be wrappers to that API's resources. Management is a wrapper for the Vulkan Instance. The Management can create the Device which is a wrapper to the OpenGL Graphics Context or for the Vulkan Device.

For Device creation You need to pass in an implemented instance of the IGraphicsView interface. With any .NET UI library You can implement this interface with a UI Control. The Control should be able to return its native Handle and resolution. The Control should notify through events when it is initialized and become ready to use or when its size and resolution changed. This information is needed to create and manage the Swap-Chain. In Vulkan the Swap-Chain creation must be done manually, the SurfaceKHR is needed for its initialization [7], and it can be created by the Device from the View's Handle. In OpenGL for the Context creation, we need the View's Handle because it creates a default Swap-Chain automatically in the background. We should only use the native Handle after the UI Control is initialized by the UI Framework, and it has been placed onto the UI hierarchy. We should be notified about the initialization by some events of the Framework.

For all these resources we have created an abstract base class or interface. From a shared project the library can be used through these wrapper classes, therefore in the shared source-code the used Graphics API is not reflected. The base classes and interfaces are implemented in separate projects for Vulkan and OpenGL, and these

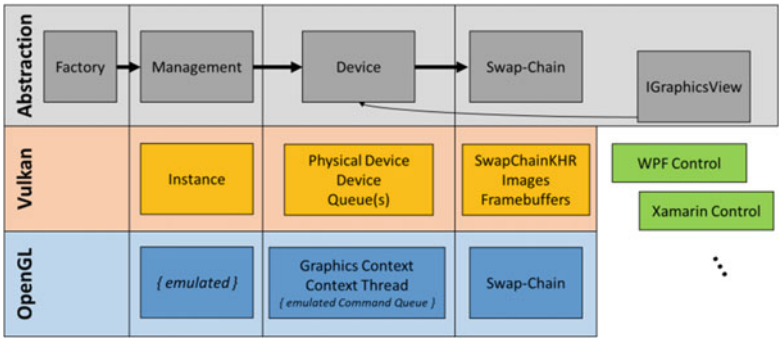


Fig. 1 Structure of classes of the abstraction layer and their implementations

implementations uses the correct Graphics API functionalities and resources behind the scenes (Fig. 1).

Because we are building the library similarly to Vulkan's structure, in OpenGL we need to emulate many functions that are handled by the driver or not accessible at all. This will introduce a performance penalty in the library when it is used with OpenGL compared to using OpenGL directly. This is expected because shaping the structure of our library closer to OpenGL's behavior would limit our possibilities of using Vulkan's (and other newer Graphics APIs) features and probably it would decrease their performance as well. The main goal of the library's design is to provide most of the features of the newer Graphics APIs with the smallest performance overhead as possible while maintaining compatibility with older and embedded machines with OpenGL and older Graphics APIs.

3.2 Commands and Threads

Both Vulkan and OpenGL collects rendering and state transition commands into Command Buffers. The difference is that in Vulkan we, the developers, are responsible for creating, filling, and submitting these Command Buffers while in OpenGL this mechanic is hidden from us. In the current state of our library, we are wrapping the Vulkan Command Buffers and emulating these buffers in OpenGL. This probably cause a performance penalty with OpenGL because the execution needs to go through the commands two times (however the library can do it somewhat parallelly).

We are working again with an abstract base class for a Command Buffer implemented both in Vulkan and OpenGL. In Vulkan the implementation is a wrapper above the Vulkan Command Buffer, therefore its functions append commands directly into the underlying Vulkan Command Buffer resource.

While it could be possible to create an immediate Command Buffer wrapper (like our Vulkan wrapper) where the OpenGL implementation calls the OpenGL functions directly, we have decided to emulate these Command Buffers for threading and safety

Fig. 2 Example of executing the list of delayed OpenGL commands

```
List<IOpenGLCommand> _commandList = new List<IOpenGLCommand>();

//...

foreach (IOpenGLCommand command in _commandList)
    command.Execute();
GL.Flush();
```

reasons mentioned later. We need to be able to delay the execution, to store these commands and their parameters in an ordered list. Commands are the instances of classes which implements an IOpenGLCommand interface. Command instances are added to a list and when we submit the list for execution the library calls the Execute method of all commands in that list. The Execute method is what will call the OpenGL function itself and the correct function to call is determined through polymorphism. Each supported OpenGL command is a class that implements this interface. If the command requires parameters, we store these parameters in a field of the command's class instance. We are storing parameters as immutable read-only structs therefore the data will be stored at the class's field, and it cannot change after the command has been submitted. This way the behavior of the two Graphics APIs would seem almost identical while using the shared library (Figs. 2 and 3).

```
internal interface IOpenGLCommand
{
    void Execute();
}
internal abstract class OpenGLParameterizedCommandBase<T> : IOpenGLCommand where T : struct
{
    private readonly T _parameters;

    protected OpenGLParameterizedCommandBase(in T parameters) => _parameters = parameters;

    protected abstract void Execute(in T parameters);
    public void Execute() => Execute(_parameters);
}

internal class OpenGLDrawCommand : OpenGLParameterizedCommandBase<OpenGLDrawCommand.DrawParameters>
{
    internal readonly struct DrawParameters
    {
        public readonly PrimitiveType geometryType;
        public readonly int first;
        public readonly int count;
        public DrawParameters(PrimitiveType geometryType, int first, int count)
        {
            this.geometryType = geometryType;
            this.first = first;
            this.count = count;
        }
    }

    internal OpenGLDrawCommand(in DrawParameters parameters) : base(parameters) { }

    protected override void Execute(in DrawParameters parameters)
        => GL.DrawArrays(parameters.geometryType, parameters.first, parameters.count);
}
```

Fig. 3 Example of classes to store OpenGL commands and delay their execution

We need the delaying of the OpenGL Commands because we cannot execute them right away. This is since we need to use multiple threads in our application even if our rendering algorithm itself is not multithreaded. We did the Swap-Chain initialization when the UI Control was created, and this means that all our initialization is executed on the UI thread, but the rendering algorithm should not be executed on it. We should always be careful and prevent the overwhelming of the UI thread because that could make the app unresponsive to the user. Therefore, we should avoid executing the Rendering Loop (the cycle that provides rendering commands to the GPU) on the UI thread. The filling of the Command Buffers is done in a separate thread which is started after initialization [8]. Inside its while cycle each iteration is a rendering of a frame by getting the next image from the Swap-Chain, filling a Command Buffer with the rendering commands, then lastly submitting the commands, and presenting the frame. We do not need to add any UI synchronization operations on this thread because the Swap-Chain resources will force waiting inside the library if needed.

In the case of OpenGL, we have started another thread after the Graphics Context's initialization. OpenGL has the limitation that its Context is only accessible from one thread [9]. This means that after the initialization of OpenGL we need to select a thread which is going to be the active thread and we must call all OpenGL commands on that thread. With the MakeCurrent function on the context we can change the active thread, by first calling it with a null parameter on the currently active thread then calling it again on the thread that we want to use. However, this is inconvenient to do and its also costs performance. It is safer if we create a dedicated thread for the Context and force the execution of OpenGL commands only on that thread. Emulated Command Buffers made this is easy to implement. When we submit the Command Buffer the list of commands is added to a BlockingCollection and the Context thread's job is to take and execute the commands from this collection. The synchronization between threads is handled by the Collection. This way we can implement multi-threaded rendering is the shared codebase which is going to work parallelly on Vulkan and will execute in single-threaded mode on OpenGL (Fig. 4).

3.3 Pipeline, Shaders, and Rendering Loop

The Graphics Pipeline is the main part of a rendering application, this is where the drawing commands are processed and their results are stored in the FrameBuffers (in our case, in the Window's Swap-Chain). The Pipeline has a state in which it executes the commands. This state accounts the used shaders, buffers, textures, other configurations, and settings that are required to execute a command on the GPU's Pipeline. The big difference between OpenGL and Vulkan is that OpenGL has one state which is constantly changing as we configure it (technically, this is the Context) and in Vulkan the state of the Pipeline is tied to the Command Buffers [10]. Therefore, in Vulkan every time when we create a Command Buffer it starts with a blank state and we need to configure the complete state before we can submit the commands for execution. In OpenGL it is possible to alter only parts of the state between

```
_context.MakeCurrent(null);

_contextThreadCancellationTokenSource = new CancellationTokenSource();
_contextThread = new Thread(() =>
{
    _context.MakeCurrent(_windowInfo);

    try
    {
        while (!_contextThreadCancellationTokenSource.IsCancellationRequested)
            _commands.Take(_contextThreadCancellationTokenSource.Token).Execute();
    }
    catch (OperationCanceledException) { }

    _context.MakeCurrent(null);
});
_contextThread.Start();
```

Fig. 4 Example of creating a dedicated thread for the OpenGL Context

drawing commands. While the using state management in OpenGL leads to faster development and less code, Vulkan's strict state handling has many advantages over it. The states and its transitions are controlled by the developer and a subsequent Command Buffer's state cannot get corrupted or altered by a previous Buffer. Also, in Vulkan most options of the Pipeline's state must be configured and compiled ahead. This will result in a Pipeline resource which contains all the required state information like shaders, used geometry type, blending modes and all the other configurations therefore the GPU knows everything about the Pipeline ahead and it can prepare better, and more efficient executions based on it (the concrete buffers and resources can be changed afterward in the Command Buffer, but their numbers, formats and relationships must be configured ahead). With OpenGL's ever altering state these kinds of optimizations are rare and hard to achieve.

Again, we are wrapping Vulkan's approach with the Pipeline and emulating it in our OpenGL abstraction. In our library the Pipelines must be created and must be configured ahead. With Vulkan the creation of our Pipeline object will lead to the compilation of shaders and the construction of the Vulkan Pipeline resource as expected. OpenGL is trickier again. Because there is no visible "state resource" to the developers we need our own way of storing it in a class with variables and fields. When we pass an emulated OpenGL Pipeline to our emulated OpenGL Command Buffer it will add a special internal command into the buffer which will call the needed OpenGL state configuration commands based on the stored values. This means that when we change the Pipeline or finished the Command Buffer, we need to automatically restore the original "default" state. The moment of the shaders compilation is an interesting question as well. We can use a "just-in-time compilation" in which we check if we have already compiled the shaders when we are binding the Pipeline. If not, then we can compile now because we are already executing on the OpenGL Context's thread. We can do the compilation ahead while creating our emulated Pipeline instance, but we still cannot compile them immediately, because the shader

compilation must be executed on the Context's thread. This means that during the creation of the emulated Pipeline instance our library will insert a special internal command into the Context's list of commands to compile the shaders on the Context's thread.

The shader source files can be added to the project as an Embedded Resource [11], and it can be loaded from the Assembly during runtime as a stream (later, even textures or other resources can be included this way as well). In the current state of our library both the OpenGL GLSL shader sources and the Vulkan SPIR-V bytecode sources must be included and specified, to be able to compile the correct shaders for the used Graphics API. This is a temporal solution; we are in the early stages of researching a way of writing shaders in C# in a shared way and compiling them onto the required format for all the supported Graphics APIs during development time.

To implement the Swap-Chain in the library we needed to prepare the Frame-Buffer implementation as well. The results of drawing commands will be stored in FrameBufffers which are groups of textures (like, color texture, depth texture etc.). To present something onto the screen we need to create a FrameBuffer using the Swap-Chain's color texture and render into that. In OpenGL, this mechanic is hidden from us and it is called Default FrameBuffer. However, in Vulkan we need to manually create and manage the FrameBuffer's for each image of the Swap-Chain (double/triple buffering requires multiple images). FrameBuffer creation is relatively cheap in Vulkan, we could do it at the beginning of every frame, but we are only creating a new one when the previous FrameBuffer for that image is not usable (for example when size of the window changed). To get the next image from the Swap-Chain we need to do CPU side synchronization with Vulkan's Semaphore resource to avoid the CPU rushing ahead the GPU and to prevent using of an image that the GPU is still using for a previous command. For this we need to create a Vulkan Semaphore for each image in the Swap-Chain and as the API requests we need to provide the image's Semaphore when we are trying to get the next image or when we are ready to present one. These Semaphores will pause our thread if necessary. While this is handled in the background by OpenGL as well, the necessary waiting would only happen on the Context's thread, our Rendering Loop's thread would keep going. To solve this, we are inserting special Semaphore-like commands into the Context's command list. Inside this command there is a CountdownEvent which is a C# threading synchronization resource. When the Context reaches the Semaphore command and sets the CountdownEvent's signal (exiting the Semaphore). The function for getting the next frame from the OpenGL Swap-Chain waits on this CountdownEvent; therefore, it can force the waiting on the Rendering Loop's thread if the Context's thread is not yet finished with the image. In our current solution we are working with a single Semaphore which presents a hard frame-by-frame synchronization that is bad for performance. We are planning to refactor this by using multiple Semaphores for letting the Rendering Loop's thread 2–3 frames ahead of the Context's thread if it is necessary.

Leveraging the abstractions and implementations of these resources we are ready to use the library for the simplest real-time graphics rendering algorithms. After the initialization of the GraphicsManagement and GraphicsDevice resources we can start

```
if (_graphicsDevice.SwapChain.TryBeginFrame(_renderPass, out GraphicsSwapChain.Frame frame))
{
    frame.commandBuffer.Clear(_graphicsDevice.SwapChain, Vector4.Zero);

    frame.commandBuffer.BeginRenderPass(_renderPass, frame.frameBuffer);
    frame.commandBuffer.BindPipeline(_pipeline);

    frame.commandBuffer.SetViewportAndScissor(frame.frameBuffer.Size);

    frame.commandBuffer.Draw(4u);

    frame.commandBuffer.EndRenderPass();

    _graphicsDevice.SwapChain.PresentFrame();
}
```

Fig. 5 CPU side rendering code using our library for a simple Raytracing application implemented with Fragment Shaders. The code is written in a shared library, can be compiled to Windows, Linux, and Android with Vulkan and OpenGL

a thread with an infinite cycle in it (infinite until the user tries to close the application, which we can handle with a CancellationToken [12] even in waiting operations like with CountdownEvent). After the resource initialization (Pipeline, Shaders etc.) each iteration should handle the preparing and rendering of a new frame. It should start by requesting the next image from the Swap-Chain, then appending commands and state transitions to the frame's Command Buffer then lastly the presentation of the frame. This code can be placed into a shared library and the rendered images can be embedded into several C# UI Frameworks (Fig. 5).

4 Performance

We have implemented a simple Raytracing application [13] in Fragment shader to test the state of our library and to measure the initial performance of our abstraction. The application does not need long CPU side rendering algorithms; therefore, the execution will be mostly GPU limited. We have tweaked the parameters so we have got a reasonable framerate and load on the GPU, but we can also detect the performance hit on our CPU code. We have implemented this Raytracing application with our C# library and compiled an OpenGL and Vulkan version using .NET Core 3.1 for Windows. Also, we have implemented it in both Vulkan and OpenGL just by using the C# binding libraries without any other abstraction layers. This way we can compare the performance penalty of our abstraction to the direct usage of Graphics API binding libraries in C# (Fig. 6).

The test results reinforced our early expectations. OpenGL performance in our library is 10–15% slower compared to using OpenGL without abstraction. This performance hit is probably going to grow larger with more complex rendering algorithms. One reason for this penalty is that we have a strict synchronization between

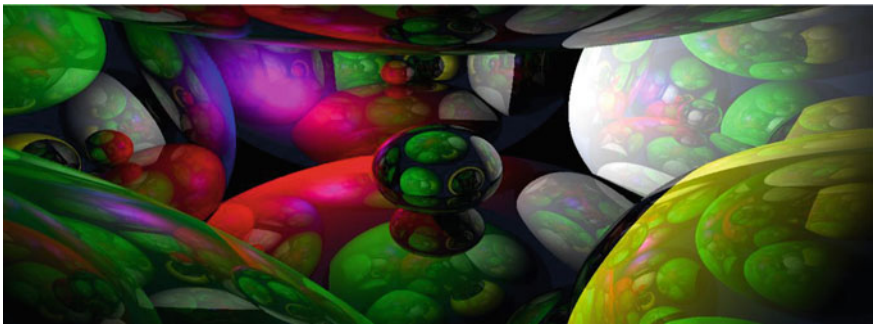


Fig. 6 A rendered frame of the Raytracing application implemented with our library

the Rendering Loop's thread and the OpenGL Context's thread during the presentation of a rendered frame. We have introduced this waiting because it is the safest way of handling the built-in Swap-Chain of OpenGL. We are planning to refactor this to a still safe, but more flexible synchronization. The other reason is the abstraction itself, the emulation of the Command Buffers. This incompatibility between concepts will always cause a performance hit.

The library's Vulkan implementation has unmeasurable performance difference compared to the directly used Vulkan binding library. While this can still change in the future when we are implementing more features from the APIs, at this state it is foreseeable, that the newer Graphics APIs like Vulkan could be used with relatively small performance penalty while older or non-compatible devices can be still supported through OpenGL (Fig. 7).

Frame pacing in real-time graphics is also as important as average framerates. During the first 20–30 s of rendering both APIs, but especially OpenGL had some spikes in its performance (50–100 ms of breaks). This could result in immersion breaking stuttering on the rendered image sequences. The rest of the execution has a

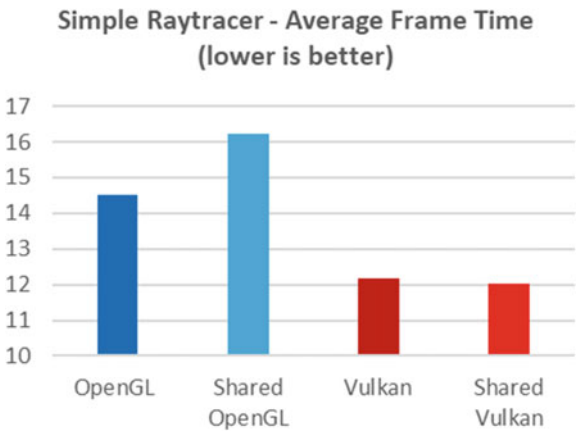


Fig. 7 Average frame rendering time of raytracing tests with OpenGL and Vulkan used in C# compared to the implementations of these APIs in our shared library (lower values are better)

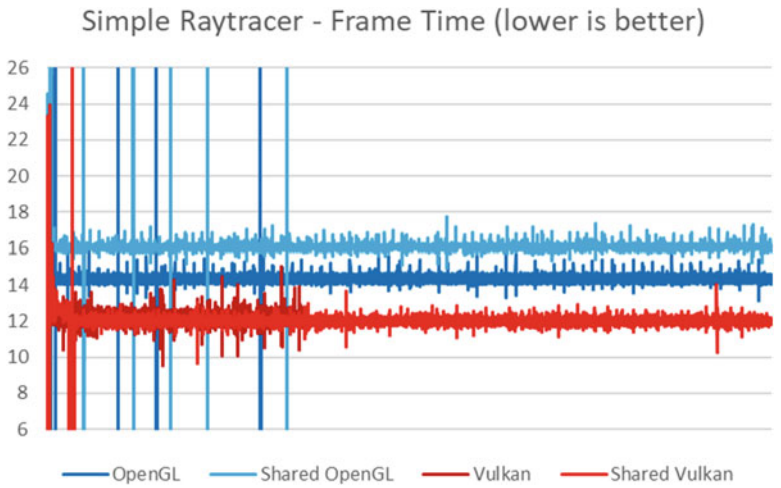


Fig. 8 Frame Pacing (Average frame rendering time) of raytracing tests with different OpenGL and Vulkan used in C# compared to the implementations of these APIs in our shared library (lower values are better)

consistent frame pacing and frames are delivered before their deadlines [14]; therefore, the performance problems at the beginning are caused by the startup of the .NET executing environment and the JIT compiling and preparing of the application (Fig. 8).

5 Conclusion

While our library is still in its early state it shows that the implementation of a low-level real-time rendering library in C# is possible which provides a unique opportunity to deliver custom rendering solutions with the .NET environment. The results of the rendering can be embedded into well-known .NET UI frameworks like WPF and Xamarin. The supported features and frameworks are limited yet, but now that the base of the architecture is built it can be extended with further functionalities and to support more APIs and platforms.

We would like to recommend this paper and our approach both to .NET and graphics developers. We are presenting a new kind of development environment for both regular C# programmers and graphics programmers as well. .NET developers could power their application with multi-platform real-time graphics modules and graphics developers could reach other platforms and new interesting development environments with our approach.

Acknowledgements EFOP-3.6.3-VEKOP-16-2017-00001: Talent Management in Autonomous Vehicle Control Technologies–The Project is supported by the Hungarian Government and co-financed by the European Social Fund.

References

1. A. Troelsen, P. Japikse, Pro C# 7: With .NET and .NET Core, 8th edn. (Apress, New York, 2017)
2. C. Bilgin, *Mastering Cross-Platform Development with Xamarin* (Packt Publishing, Birmingham, 2016)
3. OpenTK, https://github.com/opentk/opentk. Last Accessed 05 May 2021
4. Introducing .NET Standard, https://devblogs.microsoft.com/dotnet/introducing-net-standard/. Last Accessed 05 May 2021
5. OpenTK.NetStandard, https://github.com/emmauss/opentk. Last Accessed 05 May 2021
6. vk.net, https://github.com/jpbruyere/vk.net. Last Accessed 05 May 2021
7. D. Szabó, Dr. Z. Illés, Real-time vulkan graphics in C#, in *Proceedings of XXXIII. DidMatTech 2020 Conference* (2020)
8. Z. Illés, Programozás C# nyelven (Jedlik Oktatási Stúdió, Budapest, 2005)
9. D. Szabó, Dr. Z. Illés, Real-time OpenGL graphics in modern C#, in *XXXII DIDMATTECH* (Trnava, 2019).
10. G. Sellers, J. Kessenich, *Vulkan Programming Guide.* 1st edn (Addison-Wesley, 2016)
11. Xamarin Samples, github.com/xamarin/mobile-samples/blob/master/EmbeddedResources. Last Accessed 05 May 2021
12. B. Watson, *Writing High-Performance .NET Code.* 2nd edn. (Ben Watson, 2018)
13. E. Haines, T. Akenine-Moller, *Ray Tracing Gems* (Apress, 2019)
14. D. Szabó, Dr. Z. Illés, V.B. Heizlerné, *Real-Time functionality in Windows*, in *INFODIDACT* (Budapest, 2018)

Prediction of Students' Performance with Artificial Neural Network Using Demographic Traits

Adeniyi Jide Kehinde, Abidemi Emmanuel Adeniyi, Roseline Oluwaseun Ogundokun, Himanshu Gupta, and Sanjay Misra

Abstract Many researchers have studied student academic performance in supervised and unsupervised learning using numerous data mining techniques. Neural networks often need a greater collection of observations to achieve enough predictive ability. Due to the increase in the rate of poor graduates, it is necessary to design a system that helps to reduce this menace as well as reduce the incidence of students having to repeat due to poor performance or having to drop out of school altogether in the middle of the pursuit of their career. It is, therefore, necessary to study each one as well as their advantages and disadvantages, to determine which is more efficient in and in what case one should be preferred over the other. The study aims to develop a system to predict student performance with Artificial Neutral Network using the student demographic traits to assist the university in selecting candidates (students) with a high prediction of success for admission using previous academic records of students granted admissions which will eventually lead to quality graduates of the institution. The model was developed based on certain selected variables as the input. It achieved an accuracy of over 92.3%, showing Artificial Neural Network's potential effectiveness as a predictive tool and a selection criterion for candidates seeking admission to a university.

A. J. Kehinde · A. E. Adeniyi
Department of Computer Science, Landmark University Omu Aran, Omu-Aran, Nigeria
e-mail: Adeniyi.jide@lmu.edu.ng

A. E. Adeniyi
e-mail: adeniyi.emmanuel@lmu.edu.ng

R. O. Ogundokun
Center of ICT/ICE, Covenant University, Ota, Nigeria
e-mail: ogundokun.roseline@lmu.edu.ng

H. Gupta
Birla Institute of Technology Pilani, Hyderabad, India
e-mail: f20150339h@alumni.bits-pilani.ac.in

S. Misra (✉)
Department of Computer Science and Communication, Ostfold University College, Halden, Norway
e-mail: sanjay.misra@covenantuniversity.edu.ng

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_46

Keywords Admission · Student · Quality education · Demographic traits · ANN

1 Introduction

The scholastic accomplishment of students in the university is the most significant criterion to determine the nature of university students [1–3]. Studying the academic achievements of students is therefore of great importance for fostering student improvements and enhancing the standard of higher education [4–6]. However, this accomplishment is influenced by demographic factors, psychometric factors, and previous results of the student [7]. Usually, most higher institutions use the previous academic record of a candidate for admission [8]. While this is okay, the previous result is still influenced to a degree by certain demographic factors. The quality of admitted students has a big influence on the institution's level of academic performance, research, and training. The principal goal of the admission framework is to decide the applicants who might perform well in the wake of being admitted into the school. Failure to make an accurate decision on admission can result in the admission of an unsuitable student to the institution. Hence, it is essential to analyze the academic potential of students [9, 10].

The majority of the prevalent studies have usually relied majorly on the academic factors for performance prediction [4, 11–14, 36]. While this might be true for a student already admitted into the university, prospective candidates only have a previous academic level result. To circumvent this setback, there is a need to examine the influence of demographic factors on academic performance (since it influences a student's academic performance). This study approaches student's performance prediction with the use of a machine learning technique. Machine learning is mostly used to study complex relationships between data [9, 15]. Machine learning can learn without necessarily being pre-programmed. Among the machine learning techniques is Artificial Neural Network [16]. Artificial Neural Network (ANN) is growing and gaining recognition in data analysis. It is capable of analyzing complex data sets as well as identify relationships between variables [17, 18].

This study aims to implement a neural network that predicts student performance concerning demographic factors (arguments) such as gender, age, family background, and so on. This is aimed at reducing the incidence of admitting a poor student. It also points out students who qualified for admission but are likely to require attention to avoid failure [19]. These factors have been carefully studied and coordinated to be suitable for computer coding using the Artificial Neural Network model. The system was trained and tested using data of students who have already graduated intending to enhance predictive device accuracy.

The remaining sections of this paper are organized as follows: section two examined recent and relevant pieces of literature to the study, section three expatiate the methods used in the study, and the fourth section shows the result obtained from the proposed methods.

2 Literature Review

Some studies have examined student performance prediction, though it has been using the previous educational result to predict the future result in the same educational level. Some of these studies are reviewed below.

In Vamshidharreddy [18], a comparison between several classification algorithms which include decision tree (DT), support vector machine (SVM), and Naïve Bayes was presented. Naïve Bayes performed best with an accuracy of 77%. Li et al. [20] used the internet for student performance prediction. Online learning records for project courses and network logs were used in the prediction. Spatial Prediction based on Deep Network (SPDN) was used for predicting student performance at a course. They reported accuracy of 73.51%. A technique to improve scholar performance prognosis using the collaborative electronic pool and further considering student features interaction was proposed by Wei et al. [21]. New features such as time, initial try, and initial struggle and drop were introduced. Yahaya et al. [22] attempted to envisage the performance of undergraduate scholars in chemistry courses using a multilayer perceptron. An accuracy of 92% was recorded. The use of an ensemble model for the improvement of student graduation prediction was studied by Lagman et al. [23]. This is aimed at identifying students who have a high chance of not meeting the graduation time. With the identification proposed, such students can be given attention in areas that they are deficient. Average accuracy of 88.30% was recorded at best.

Quinn and Gray [24] approached student's academic performance prediction with Moodle data using a further education setting. Their study investigated if data from the learning management system Moodle can be used in predicting student's academic performance. This is aimed at predicting whether a student would pass a course or not. The classifier on all course data had an accuracy of 65%. Predicting whether a student would pass or fail very well has an accuracy of 92.2%. Data built on the first six weeks performed poorly, and there was a need to extend the data gathered. Li et al. [20] proposed the use of Graph Neural Network (GNN) for better scholar performance prognosis. Collaborative electronic questions were presented to students. A new GNN was further presented, that is R2- GNN. Rakic et al. [25] presented a paper aimed at examining how useful and impactful digital technology is on an e-learning platform. Data from online sources were used to examine key pointers of the performance of students in several courses. Social Network Analysis [35], K-Means clustering, and Multiple Linear Regression were employed to evaluate scholar's success. The result demonstrates a huge connection amid the performance of the student and the utilization of electronic informative materials from the online education platform. There are several other related studies [4, 26, 27, 36] available in literature related to student's performance which we are not considering due to the limit of the work.

3 Material and Method

It has been observed that several variables (including demographics, extra-curricular activities, environment as well as biological factors) are considered to affect the performance of a student. These variables were carefully researched and modified to form a detailed equivalent number appropriate for computer coding. These variables are taken as input to the system. Figure 1 displays the proposed system flow diagram.

3.1 Data Collection

The input variables used are those gotten from the dataset of the student database. The dataset was obtained from the UCL Machine Learning Repository (https://www.kaggle.com/dipam7/student -grade-prediction). The dataset includes several variables, among which using educational data mining the features that are most related to and have more effect on the output variables were selected and they include: The output variable reflects a student's success at the end of a term/semester. However, for this study, the variables used include G1(assignments), G2(tests), G3(Final exam) which are numeric values within the range of 0–20.

Since the classifier accepts input in numeric form, there is a need to converting categorical variables to numerical in preparation for training. Table 1 shows the transformed data.

Figures 2, 3, and 4 below is a probability plot of sample data against the quantiles of a given theoretical distribution (the default normal distribution). It calculates optionally a best-fit line for the data and plots the results. From Fig. 1, the skewness was determined to be 0.240613 while the Kurtosis was -0.693830. The mean, mu = 10.91, and the standard deviation, sigma = 3.31. The skewness in the probability plot of G2 in Fig. 2 was determined to be -0.431645, while the Kurtosis was

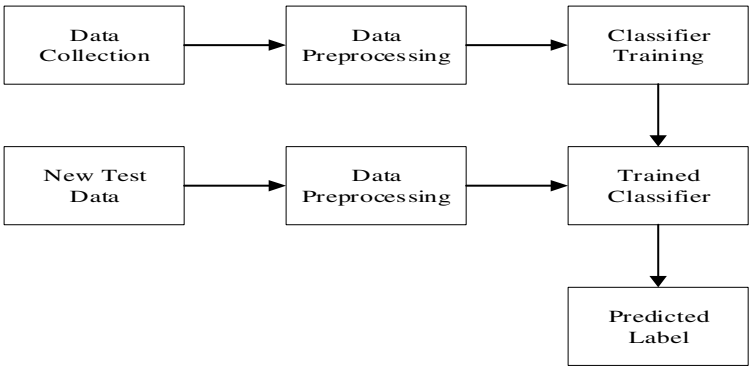


Fig. 1 Proposed system flow diagram

Table 1 Most correlated features

0	G1
1	G2
2	G3
3	Medu
4	Fedu
5	Studytime
6	Famrel
7	Free time
8	Absences
9	Age

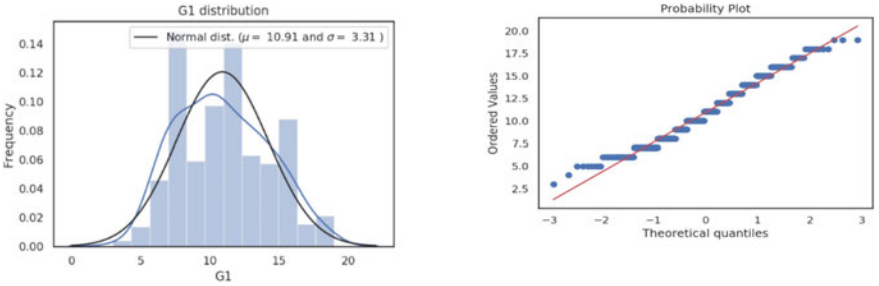


Fig. 2 Probability plot of G1

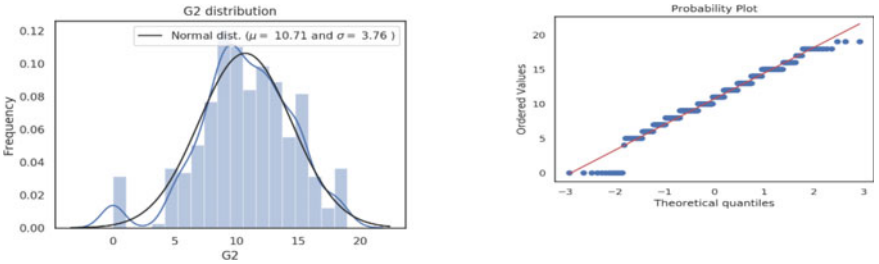


Fig. 3 Probability plot of G2

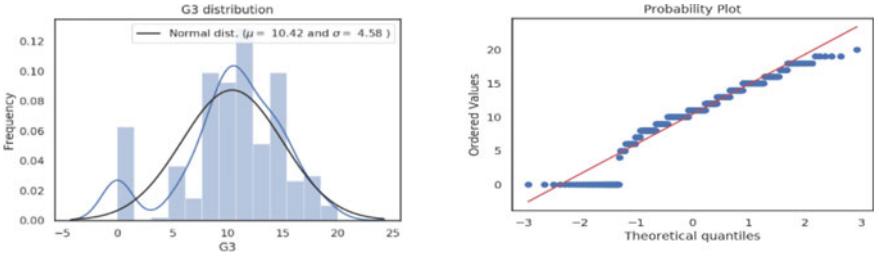


Fig. 4 Probability plot of G3

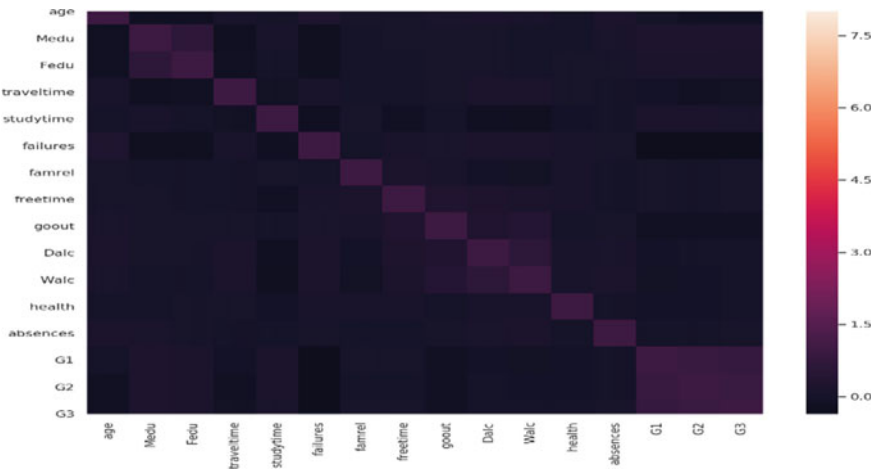


Fig. 5 Correlation matrix

0.627706. Finally, the skewness was determined to be -0.431645, while the Kurtosis was 0.627706 for G3 in Fig. 3. These skewness and kurtosis values for G1 imply that the distribution is fairly symmetrical and lacks outliers. However, the value of skewness and kurtosis for G2 and G3 shows that they a slightly negatively skewed and have outliers. Hence, the correlation matrix was used to eliminate the outliers in the distribution.

Features Selection using the Correlation Matrix

Correlation examines the relationship and association that exist between variables. The relationship and association refer to how much a variable is affected by a change in another variable. The correlation could be simple, partial, or multiple depending on the number of variables being examined at once. A zero correlation typically implies that no relationship exists between the variables. The correlation matrix was applied to determine the features that affect the results G1, G2, and G3. This weeds out the variables (features) that are not relevant to the performance of the candidate (feature selection) [7]. The correlation matrix in Fig. 5 generated through the exploratory data analysis finds which features are most related to the G1, G2, or G3 scores.

Where G1, G2, and G3 stand for assignment scores, test scores, and final exam scores respectively. The most correlated features are shown in Table 1.

3.2 Prediction

For forecasting student's performance, ANN was applied. ANN is a system that is made up of several neurons (taken as input to the system), hidden layers, and output layers [28]. Weights are used to connect the neurons. The process of designing the

network is a tedious one as there is a dependency between the various variables like the structure of the network and the values of the weights [19, 29, 30]. The learning phase and the prediction phase are the two phases of the Network. The learning phase plays an important role as this is where the weight is adjusted to produce the required output [28]. The adjustment of the weight is repeated until a termination criterion is met [31, 32]. These termination criteria could be the acceptable mean square or number of evolutions to achieve the target value. After training an ANN, the trained network is tested and validated. To test and validate the trained ANN, unknown samples are used to test the Network. To express an ANN problem, we can use the set of inputs in (1) [33],

$$X = \{x = x_i | x \in \mathbb{R}^n, i = 1, 2, \ldots n\}, \tag{1}$$

and outputs in (2)

$$Y = \{y = y_o | y \in \mathbb{R}^m, o = 1, 2, \ldots m\}. \tag{2}$$

From (1) and (2), we can observe that the input and output layers of the ANN have n and m neurons respectively. Given a chosen $\mathcal{F}$ number of hidden layers, to design an ANN for the data in (1) and (2) the objective function can be given as in (3),

$$\mathcal{F} = f(X, Y, W). \tag{3}$$

W being the weight connections between the layers. It could be represented as in (4),

$$W = \{W_{12}, W_{23}, ..., W_{l-1,l}\}. \tag{4}$$

l is the sum of the layers, this includes the input, hidden, and output layers. Minimizing $\mathcal{F}$ in (3), the mean square error (MSE) gives an optimized synaptic weight of ANN.

In [33], from the schematic diagram of an ANN, it is observed that the system performs the addition of all the inputs first, and then the output of the addition is passed to the transfer function. The output of the transfer function serves as an input to the neuron in the preceding layer. Each neuron is formed with the representation in (5) [34],

$$y_p = f\left(\sum_{i=0}^{n} w_i^p x_i^p + b^p\right), \tag{5}$$

where x_i^p in (5) is the ith input of the pth neuron, w_i^p is the weight value of the connection between ith input and pth neuron, and $\mathcal{F}$ is the transfer function. A bias

is associated with each neuron, and it is represented with b. y_p is the output of the pth neuron which is passed as input to the next neuron?

4 Results and Discussion

This section describes the statistical results, the configuration of the Artificial Neural Network, and performance based on the academic performance educational data obtained of 376 students. Python programming language was used as well as the TensorFlow backend.

4.1 Datasets

In supervised training, the data set is divided into three groups: the training set, validation or verification test, and testing set. The training set allows the system to detect the relationship between the input data and the given outputs, to create a relation between the input and the expected outcome. Altogether, 396 student records were used in this analysis where the training and testing set was split in the ratio 70:30.

The dimension reduction is essential for feature extraction, to take out the features that are important in the dataset. The params denote parameters that show the number of available data in the dataset. After every dropout layer, the number of parameters reduces.

The network was trained with batch size set to eight and the Epoch set to terminate at 500. Figures 6 and 7 show the histogram of the training and validation in terms of its Mean Absolute Error (MAE) and loss.

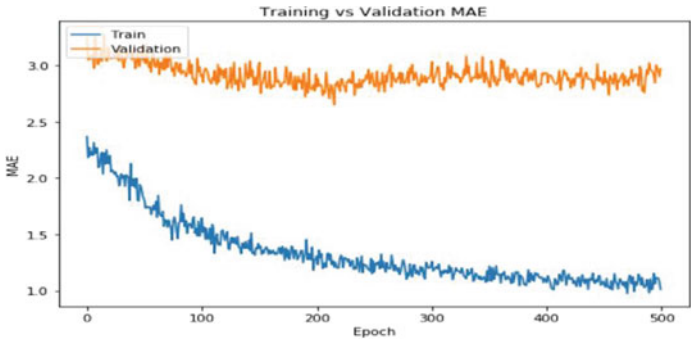


Fig. 6 Training and validation MAE

Fig. 7 Training and validation loss

Table 2 Comparison of Similar Student Performance Prediction Systems

Authors	Study focus	Accuracy (%)
Vairachilai and Vamshidharreddy [18]	Comparison of methods–decision tree, SVM and Naïve bayes	77
Li, Zhu, Zhu, Ji and Tang [20]	Performance at a course	73.51
Yahaya et al. [22]	Undergraduate who will pass chemistry courses	92
Quinn and Gray [24]	Pass or fail on All courses	65
Pekuwali [26]	Student performance in the final year	94.24
Proposed System	The student will graduate or not	92.26

4.2 Discussion

An accuracy of 92.26% was obtained after testing the system. Table 2 demonstrates a comparative analysis of the result obtained with other similar systems.

Considering Table 2, it can be noted that the performance of the system is encouraging. In the instance where the accuracy was higher than the proposed system, previous records of academic performance were being used to predict the subsequent result. This makes it difficult to use the system to predict student's performance at the time of admission. The designed system considers only the demographic factors and its performance shows high accuracy. The performance of the system was greatly influenced by the methods used in examining the correlation between the input data and the predicted outcome.

5 Conclusion

This paper presented an approach to student performance prediction using demographic features only. The relationship between the demographic features and the result was examined with the skew and kurtosis of the dataset. Outlier from the statistical evaluation of skew and kurtosis were further removed in the feature selection stage using the correlation matrix. The resulting dataset was then used to train an artificial neural network. The demographic dataset was obtained from the UCL Machine Learning Repository. The student performance prediction showed an accuracy of 92.26%. The proposed approach is not restricted to demographic features only. Hence, examining the performance of the system with both demographic and previous records of students for predicting student's performance at the next educational level could be an area of study in the future.

References

1. J.B. Awotunde, R.O. Ogundokun, F.E. Ayo, G.J. Ajamu, E.A. Adeniyi, E.O. Ogundokun, Social media acceptance and use among university students for learning purpose using UTAUT model. Adv. Intell. Syst. Comput. **1050**, 91–102 (2020)
2. R.O. Ogundokun, M.O. Adebiyi, O.C. Abikoye, T.O. Oladele, A.F. Lukman, A.E. Adeniyi, A.A. Adegun, B. Gbadamosi, N.O. Akande,Evaluation of the scholastic performance of students in 12 programs from a private university in the south-west geopolitical zone in Nigeria. F1000Research **8** (2019)
3. A. Soni, V. Kumar, R. Kaur, D. Hemavathi, Predicting student performance using data mining techniques. Int. J. Pure Appl. Math. **119**(12), 221–227 (2018)
4. O.T. Omolewa, A.T. Oladele, A.A. Adeyinka, O.R. Oluwaseun, Prediction of student's academic performance using k-means clustering and multiple linear regressions. J. Eng. Appl. Sci. **14**(22), 8254–8260 (2019)
5. A.M. Shahiria W. Husaina, N.A. Rashida, A review on predicting student's performance using data mining techniques, in *The Third Information Systems International Conference; Procedia Computer Science*, vol. 72 (2015), pp. 414–422. https://doi.org/10.1016/j.procs.2015.12.157
6. M. Imran, S. Latif, D. Mehmood, M.S. Shah, Student academic performance prediction using supervised learning techniques. iJET **14**(14) (2019)
7. H. Altabrawee, O.A.J. Ali, S.Q. Ajmi, Predicting students' performance using machine learning techniques. J. Univ. Babylon, Pure Appl. Sci. **27**(1) (2019)
8. H. Agrawal, H. Mavani, Student performance prediction using machine learning. Int. J. Eng. Res. Technol. (IJERT) **4**(03) (2015)
9. J. Sultana, M.U. Rani, M.A.H. Farquad, Student's performance prediction using deep learning and data mining methods. Int. J. Recent Technol. Eng. (IJRTE) **8**(1S4) (2019)
10. E. Wakelam, A. Jefferies, N. Davey, Y. Sun, The potential for student performance prediction in small cohorts with minimal available attributes. Br. J. Edu. Technol. (2019). https://doi.org/10.1111/bjet.12836
11. S. Bergin., A. Mooney, J. Ghent, K. Quille, Using machine learning techniques to predict introductory programming performance. Int. J. Comput. Sci. Softw. Eng. (IJCSSE) **4**(12), 323–328 (2015)
12. J.R. Ebenezer, R. Venkatesan, K. Ramalakshmi, J. Johnson, P.I. Glen, V. Vinod, Application of decision tree algorithm for prediction of student's academic performance. Int. J. Innov. Technol. Explor. Eng. (IJITEE) **8**(6S) (2019)

13. L.M.A. Zohair, Prediction of Student's performance by modelling small dataset size. Abu Zohair Int. J. Educ. Technol. Higher Educ. (2019), pp. 16:27. https://doi.org/10.1186/s41239-019-0160-3.
14. D. Buenaño-Fernández, D. Gil, S. Luján-Mora, Application of machine learning in predicting performance for computer engineering students: a case study. Sustainability **11**(2833) (2019). https://doi.org/10.3390/su11102833
15. B. Sekeroglu, K. Dimililer, K. Tuncal, Student performance prediction and classification using machine learning algorithms, in *ICEIT 2019, March 2–4, 2019* (Cambridge, United Kingdom, 2019). https://doi.org/10.1145/3318396.3318419
16. J.L. Rastrollo-Guerrero, J.A. Gómez-Pulido, A. Durán-Domínguez, Analyzing and predicting students' performance by means of machine learning: a review. Appl. Sci. **10**(1042) (2020). https://doi.org/10.3390/app10031042
17. A.K. Hamoud, A.M. Humadi, Student's success prediction model based on Artificial Neural Networks (ANN) and a combination of feature selection methods. J. SouthWest Jiaotong Univ. **54**(3) (2019). https://doi.org/10.35741/issn.0258-2724.54.3.25
18. V.S. Vamshidharreddy. Student's academic performance prediction using machine learning approach. Int. J. Adv. Sci. Technol. **29**(9s) 6731–6737 (2020)
19. F. Ofori, E. Maina, R. Gitonga Using machine learning algorithms to predict students' performance and improve learning outcome: a literature based review. Stratford Peer Rev. J. Book Publ. J. Inf. Technol. **4**(1), 33–55 (2020)
20. X. Li, X. Zhu, X. Zhu, Y. Ji, X. Tang, Student academic performance prediction using deep multi-source behavior sequential network. PAKDD **2020**, 567–579 (2020). https://doi.org/10.1007/978-3-030-47426-3_44
21. H. Wei., H. Li., Xia M., Wang Y., Qu H. (2020). Predicting Student Performance in Interactive Online Question Pools Using Mouse Interaction Features. LAK 2020, March 23 - March 27. https://doi.org/10.1145/3306307.3328180.
22. C.A.C. Yahaya, C.Y. Yaakub, A.F.Z. Abidin, M.F.A. Razak, N.F. Hasbullah, M.F. Zolkipli, The prediction of undergraduate student performance in chemistry course using multilayer perceptron, in *The 6th International Conference on Software Engineering & Computer Systems; IOP Conf. Series: Materials Science and Engineering*, vol. 769 (2020). https://doi.org/10.1088/1757-899X/769/1/012027
23. Lagman A. C., Alfonso L. P., Goh M. L. I., Lalata J. P., Magcuyao J. P. H., and Vicente H. N. (2020). Classification Algorithm Accuracy Improvement for Student Graduation Prediction Using Ensemble Model. International Journal of Information and Education Technology, Vol. 10. https://doi.org/10. https://doi.org/10.18178/ijiet.2020.10.10.1449.
24. R.J. Quinn, G. Graya Prediction of student academic performance using Moodle data from a Further Education setting. Irish J. Technol. Enhanc. Learn. **5**(1) (2020)
25. S. Rakic, N. Tasic, U. Marjanovic, student performance on an e-learning platform. mixed method approach. iJET **15**(2) (2020)
26. L. Fernandez-Sanz, J.A. Medina, M. Villalba, S. Misra, A study on the key soft skills for successful participation of students in multinational engineering education Int. J. Eng. Educ. **33**(6(B)), 2061–2070 (2017)
27. L. Fernández, J.A. Medina, M.T. Villalba de Benito, S. Misra, Lessons from intensive educational experiences for ICT students in multinational settings. Tech. Gaz. **24**(4) (2017)
28. C.E. Choong, S. Ibrahim, A. El-Shafie, Artificial Neural Network (ANN) model development for predicting just suspension speed in solid-liquid mixing system. Flow Meas. Instrum. (2020). https://doi.org/10.1016/j.flowmeasinst.2019.101689
29. J. Jabamony, G.R. Shanmugavel, IoT Based bus arrival time prediction using Artificial Neural Network (ANN) for Smart Public Transport System (SPTS). Int. J. Intell. Eng. Syst. **13**(1) (2020). https://doi.org/10.22266/ijies2020.0229.29
30. D. Bui, T.N. Nguyen, T.D. Ngo, H. Nguyen-Xuan, An artificial neural network (ANN) expert system enhanced with the electromagnetism-based firefly algorithm (EFA) for predicting the energy consumption in buildings. Energy **190** (2020). https://doi.org/10.1016/j.energy.2019.116370

31. R.B. Solanki, H. Kulkarni, S. Singh, P.V. Varde, A.K. Verma, Artificial Neural Network (ANN)-based response surface approach for passive system reliability assessment Reliability and Risk Assessment in Engineering. Lecture Notes in Mechanical Engineering (2020). https://doi.org/10.1007/978-981-15-3746-2_29
32. A. Raya, T. Halderb, S. Jenaa, A. Sahooa, B. Ghoshb, S. Mohantyc, N. Mahapatrad, S. Nayaka, Application of artificial neural network (ANN) model for prediction and optimization of coronarin D content in Hedychium coronarium. Ind. Crops Prod. **146** (2020). https://doi.org/10.1016/j.indcrop.2020.112186
33. N. Gupta, M. Khosravy, N. Patel, S. Gupta, G. Varshney, Artificial neural network trained by plant genetic-inspired optimizer, frontier applications of nature inspired computation (2020). https://doi.org/10.1007/978-981-15-2133-1_12
34. N. Upadhyay, P.K. Kankar, Integrated model and machine learning-based approach for diagnosis of bearing defects reliability and risk assessment in engineering. Lecture Notes in Mechanical Engineering (2020). https://doi.org/10.1007/978-981-15-3746-2_20
35. N. Sharma, G. Sikka, Multimodal sentiment analysis of social media data: a review, in T*he International Conference on Recent Innovations in Computing* (Springer, Singapore, 2020, March), pp. 545–561
36. C. Verma, V. Stoffová, Z. Illés, D. Kumar, Toward prediction of student's guardian in the Secondary Schools for the real time, in *The International Conference on Recent Innovations in Computing*. (Springer, Singapore, 2020, March), pp. 755–765

Real-Time Interaction Tools in Virtual Classroom Systems

Viktória Bakonyi, Zoltán Illés, and Tibor Szabó

Abstract For years our research goal is how to integrate ICT in teaching process to activate and involve our students and get better and better results. We implemented a Classroom Response System which fitted perfectly to our university requirements. During the emergency situation caused by COVID-19 most of the education changed to remote teaching choosing between hybrid, online synchron or online mode. In the time of writing the paper we already started the third semester of emergency situation. Now we got to know thoroughly different IT tools and experiences how to enhance our education quality which key factor is activity. In this paper we should like to focus on interaction possibilities in virtual classroom systems and present some benefits of E-Lection system implemented by us. Online synchron teaching period with students' feedbacks lead us to plan new additional features to our system. This newness will produce a more useable, complex, and integrated system to support both students and professors' requirements.

Keywords Virtual classroom system · Real-time · Interaction · Education · CRS · Classroom response systems

1 Introduction

Several years ago, we faced the fact that students are not so active comparing to their behavior ten or twenty years ago and sometimes it seems they are boring during the lessons. This was noticed by many educators all over the world. In 2014 Ferdinand von Prondzynski, of Robert Gordon University in Aberdeen, spoke about the old-fashioned lectures which may cause a higher dropping rate. The problem declared by him is the following: "*Truly interactive lectures are still rare, and nowadays*

V. Bakonyi (✉) · Z. Illés
Eötvös Loránd University, 1117 Pázmány s. 1/c, Budapest, Hungary
e-mail: hbv@inf.elte.hu

T. Szabó
Faculty of Central European Studies, Constantine the Philosopher University, Nitra, Slovakia

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_47

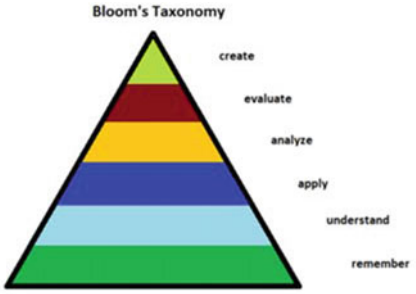


Fig. 1 Blooms' taxonomy—on the top level the creation (active learning)

many students don't turn up at these events at all," (https://bit.ly/37DnkNO) though activity is one of the key elements of modern teaching and learning.

At about 60 years ago, E. Dale the American pedagogist measured the efficiency of learning during different passive and active learning environment and the result proved his idea that activity enhance the efficiency of learning. Here we can mention the Bloom taxonomy which makes classification of learning objects stating that remembering is on the lowest and creation is on the highest level, which highlights the importance of active learning (Fig. 1).

But it is not the only viewpoint because classical lectures lost their popularity. Prensky [1] spoke about **digital natives** first. It seems to be at first only a simple word "digital natives" describing the generation which meet digital devices already in their early ages and use them as evidence. It is far much than this! The usage of devices, the usage of internet browsing for several hours per day changed the physical activity of their brain. In an experiment they measured the **brain activity** of persons who usually read books and others who uses browsers. The difference of the active area is conspicuous. Researches state that this is the result of the must of frequent and quick decisions what we are forced when reading a homepage with a lot of links—which one to follow. This way digital natives need bigger stimulus not to be bored. They are surrounded by a lot of different devices in their daily life, so they are accustomed to colorful media flow. If they are boring, they start to use their smart devices for non-academical purposes like checking the social network. "Students perceived that their non-academic information and communication technology use had costs, especially distraction" due to a new research [2]. This problem was detected earlier by Kathrine Hayes who started to use a new notion: the **hyper attention** [3]. Hyper attention means that the person who use several devices parallel must switch between the stimulus and it can be managed properly on 2% of people—at the others efficiency will decrease [4]. Professors may choose between two things to avoid these unneeded effects "restrict or integrate?" [2].

These facts lead that we really must change our teaching methods and we voted for integration.

2 Interaction Using CRS (Classroom Response Systems)

In the 60 s appeared the first audience response systems (ARS) to map the satisfaction of participants in marketing of products or in politics. As any other tool it began to find its place in education as well. They were expensive systems with built in clickers, so they did not become a well-spread possibility. At about 15 years ago smart phones started their carrier in market. A lot of workplaces and leaders all over the world realized that they could involve these devices into the working process. That was the way how BYOD (Bring Your Own Device) systems became a real possibility in a lot of cases, e.g., in the world of ARS-s. More and more new applications appeared for special education goals too and their name became CRS (classroom response system, also called SRS = Students Response System), mainly they are implemented as BYOD systems based on the student's own devices.

You can find ARS-s on link https://www.capterra.com/audience-response-software/ where you can compare their features too to choose the best fit for your purposes.

The main question is whether they are really useful in teaching or they are only a new fashion which will disappear soon due to its uselessness. We can read several publications relating to this topic [5–7].

"Results indicate that mobile-based CRS technology is a useful and effective tool for facilitating interaction among learners and content, enhancing students' engagement with entrepreneurial knowledge acquisition, and improving students' motivation toward increased entrepreneurial capability" [8].

The fact is that it may be used in a way that it helps to get real-time feedbacks mapping the actual knowledge and interest of the audience and make the lecture, the lesson more interactive with activating the students. Active learning techniques fits to all students but especially helps students with low-income therefore it may be promoting equality [9]. That is why we cannot resign their usage in modern education. For today more and more educators use such systems in their daily process, e.g., Kahoot is used by 50% of American teachers in public schools but these types of applications appear in university education too.

3 E-Lection—Our CRS

In our teaching process we also faced with the problem, mentioned in the introduction, the lack of activity, the lack of interaction so we decided to use a CRS in our practice too (in the time of face-to-face lectures). The first idea got to our mind 9 years ago, when we examined the free, BYOD systems and realized there is nothing which is totally good for our goals. We asked students opinion about BYOD system usage and their attitudes toward a CRS System. After getting positive answers we decided to implement an own system [10]. Now we have the 3rd version of it, refining always

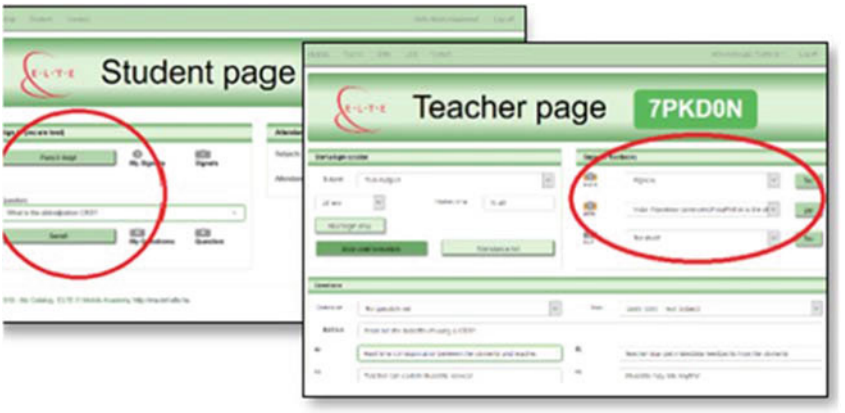


Fig. 2 Teachers' and students' pages with communicational panels

to the feedbacks of students and the requests of our faculty. The system is available at https://election.inf.elte.hu.

The system was created as a web-based bidirectional, real-time lecture management system. We use university LDAP authentication to avoid extra human resources for handling user database.

Anybody with any smart devices (phones, laptops, tablets) is able to login the system, though students have to give a Captcha code too, created by the professor. Bidirectional behavior means that not only the teachers may send questions to the joined devices, but students also may send questions to the professors. Extra specialty of our system that not only questions but do not understand signals may be sent by the students which are presented in real-time of teachers' interface, and they can catch them at a glance (see Fig. 2). The professors may send questions from a ready-made question bank or on the fly and later the answers and activity of students can be analyzed due to the built-in database.

After the test semesters in which we refined the solution due to students' feedbacks we made experiments. In 2018 we measured the efficiency of it, and we can state that using it increases the exam result of students. Moreover, the students are more active, and it helps the professors to detect the problematic topics on the fly. We presented the results in several papers, conferences [11–13].

This was the situation when COVID-19 blocked almost everything in our normal lifestyle.

4 COVID-19—New Problems in Education

Immediately thousands of researchers all over the world started experiments to understand the new problems and find solutions. Biologists, chemists, and physicians made

experiments to reveal the behavior of the virus to be able to defeat it [14]. Mathematicians made models "*to evaluate the behavior of the virus spreading*" [15]. Sociologists, psychologists researched the effect of social distance to the human being [16].

What about education which usually based on personal connections? "This impact both the way tutors teach and students learn. This also transforms how all these stakeholders interact with each other's" [17]. At first educators tried to modify their previously honored teaching process by their own ideas. Soon more and more papers became published analyzing education during Covid [18] and Software Industry offered a lot of free solutions to help schools [19]. As in many other countries in Hungary and in Slovakia too we stated the third remote semester. There were different possibilities to use:

- LCMS (Learning Content Management System) where educators have to upload the documents, ppt-s, video tutorials, create online tests, and quizzes; on the other hand, the students have to upload the homework, performed the online tests.
- VCS (Virtual Classroom Systems) and give lessons online in real-time to keep the personal contact with our students as much as we can.
- Hybrid solution, some lessons are online, some has got only documents or another version of hybrid learning to attend personally only one third of the students at the same time and they are rotating.

In the twenty-first century there are a lot of applications which we can use and integrate into education but "*the teacher should be careful not to use too many interfaces at once (Facebook, WhatsApp, Discord, MS Teams, Google docs, Skype)*" [20]. It is very important that the institution should determine the information systems, software, etc. that can be used.

Let us see what the situation at our universities was!

4.1 Faculty of Central European Studies, Constantine the Philosopher University in Nitra, Slovakia

Constantine the Philosopher University in Nitra recommended the use of combinations of LMS (**Moodle**) and a video conferencing software (**Jitsi Meet**), but MS Teams were also available. At university, we use the Moodle LMS with many years of experience, furthermore the teachers often participate on courses on how to use this system. The FCES (Faculty of Central European Studies) is mainly engaged in humanities, regional tourism, and teacher training. Of course, according to lock down we must provide an easy, convenient system to support online education. Jitsi Meet complied with these requirements.

4.2 Faculty of Informatics, ELTE, Hungary

In Informatics Faculty at ELTE, Hungary, the leader's decision was to go on with the schedule using real-time virtual classroom system (VCS), **Microsoft Teams**. Nobody new whether how many students will be able to follow it from home due to possible network problems or due to missing devices; therefore, the dean's instruction was to create more written documents as well and make recordings available from our LMS system, **Canvas**.

At the end of the first emergency semester, we asked the students opinion about it.

4.3 Students' Opinions

We made a survey which was repeated in Spring and in Autumn semester in 2020 asking students opinion about the used VCS see Table 1. We collected altogether 577 data from Hungarian and English students. These surveys were created by Google Forms to ensure anonymity [19]. Most of them liked it, they graded it quite well, but they do not think it may substitute face-to-face learning in each case. (We must add that first semester student's likeness was lower than experienced students, which average was grade 4.0 in Autumn.)

Besides their likeness toward VCS-es, we wanted to understand what the advantages and disadvantages of this type of teaching by their opinion are. They mentioned a lot of things; we analyzed the results and published in papers [21–23]. Among others the result proved that they need more interactivity what is equal to our personal experiments as teachers and strengthened by other researchers.

"*Interaction is crucial to student satisfaction in online courses*" [24].

That is why we began to focus the possibility of interactions built into VCS systems like Ms Teams and Jitsi Meet.

5 Virtual Classroom Systems—Tools for Interaction

The popularity of virtual meetings increased within the last decade in industry mainly in multinational firms where employees may be staying in different countries. That is why there are professional video conference applications in the market like Zoom

Table 1 Students' opinions

University	Likeness	May substitute	Number of students
ELTE FI	3.75	3.04	473
UKF	3.71	3.21	61

(https://zoom.us/), Ms Teams (https://www.microsoft.com/en-ww/microsoft-teams/online-meetings), Cisco Webex (https://www.webex.com/), Google Classroom with Google Meet (https://meet.google.com/), etc. As usual education soon try to use any new possibilities in teaching, in this context these applications are called virtual classroom systems (VCS). Before the emergency situation there were already offered functionalities for using them in virtual classrooms, though they were not used too frequently in education, in a normal situation. [25]. After Covid everything will be changed, and we shall turn toward the new possibilities [26].

Their standard functionalities are as follows: Video streaming, calling; Chat possibility; Content sharing; Recording; Assignments; Gradings.

5.1 Ms Teams

Microsoft Teams has got professional features. There is desktop, smart devices downloadable, and a version for using in most of the browsers, though this last-mentioned has not all the features. The owner can add members or accept or decline requests to be a member. In a meeting you may have channels and private channels to work in. Besides video streaming and calling there is a built-in recording possibility too. The recorded videos are stored in the system for 21 days automatic. Limitations of the meeting member numbers are really high (1000), though it can show parallel only 49 webcams together in gallery.

Users can post or chat in any time and a notification will be send to the partner. According to our experience students preferred communication way is this as it is proved earlier in experiences [27]. In the application we have quiz and assignment possibilities with grading. In the class notebook the teacher can publish files (similar to an LMS), watch the work of single or groups of students and offer a living share content for each participant. We may download the attendance register, with date and time of login and logout, though a bit blemish that it is available only during the lesson. Professor face to a big problem if he wants to use full interactivity with Teams. Students may mute or kick out each other using "member" permissions. Otherwise, using a weaker permission they are not able to cyber bullying each other, but the mean time they are not able to share their desktop and show their problems.

5.2 Jitsi Meet

Jitsi Meet is a very easy to use, free, open-source, video conferencing software which is fully encrypted. Does not require installation, can be run from a web-browser, but we can find this application in App Store and Google Play for iOS and Android operating systems too. Without registration, we can create video chat rooms, in which you can restrict access with a password. We can share the view of the app, the web browser tab, the full screen, or a YouTube video. Other important features of

Table 2 Comparison of features due to interaction tools

Appl. name	Talk	Live share	Chat	Raise hands	Polls
MS Teams	X	X	X	X	X
Jitsi Meet	X	X	X	X	-

the software are live streaming and video recording. Host of the room can manage participants (for example: kick out or mute it). It is clear that for the full support of the online educational process, it is necessary to have additional software. Previously we can read about the Jitsi Meet problems with connections and the limited number of users [28] though in our experience it is much better now due to the fact that it is running on university own server.

Comparison—advantages, disadvantages

We summarized our experiences of the used CRS-s in Table 2**.** The basic features of the video conference applications were almost the same. Naturally MS Teams has got several additional features which are like an LMS. Despite all this we think that they do not offer all tools a teacher need! Let us make the situation clear!

6 Why to Use a Parallel System?

Professional VCS systems support students' activity in different ways besides the ability of add their web-camera or use their microphones to join to a discussion like sharing content, live real-time shared editing, chatting, raising hand and creating polls which we really need. In the case of several hundreds of students instead of 20–30 students' professors need more tools to follow the interactions.

- Educators want to keep on their eyes on student's activity and the VCS systems usually do not have built in tools for saving such activities like raising hands, likes (see Fig. 3) or chatting. Chatting is an unstructured flow of messages which is hard to follow. Sometimes we get more than 50 questions during a lesson (see Fig. 4).

- It is impossible to answer on the spot—it would be great to be able to download for later datamining. In our CRS everything is available in a structured form, data is stored in a database and downloadable for datamining contrary to Teams chat or likes possibilities (see Fig. 5).

- Built in polls or Forms are great with a lot of settings to give before sending them, so they are not really good for real-time, on the fly feedbacks. To send a question you have to click on menu assignments, then choose quiz, then click on Create button, then give the settings, then comes the Form to write questions and at last send them. You are not able to stop answering when you want—you have to give it previously. In our solution on the fly questions are one of the greatest

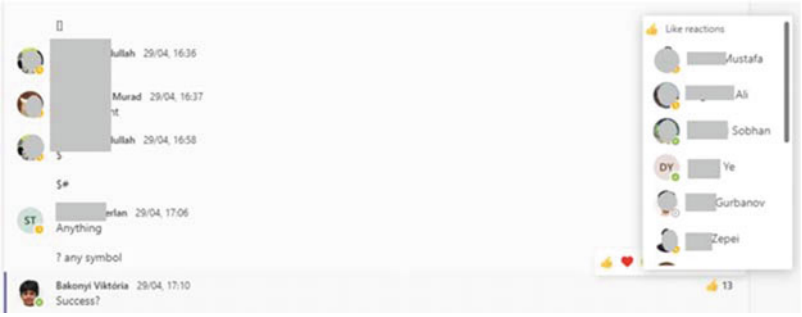


Fig. 3 Likes—hard to follow and impossible to download in a huge class

Fig. 4 Chat s during a lesson

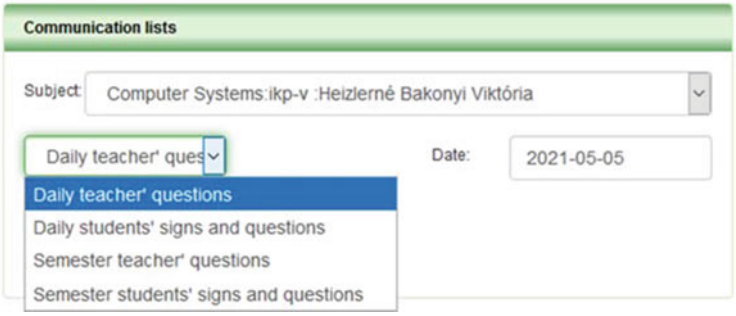


Fig. 5 Downloadable Communication lists in E-Lection

features, professors may see on the main page the textboxes where they can type question (or overwrite questions from question bank and send it with one click. With another click they can stop answering possibility! (see Fig. 6).

Sometimes teachers want to get an attendance register just to know who participated in a meeting. VCS systems has got such possibilities though they are not so comfortable or may show improper list due to possible anonymity. Contrary to our E-Lection system where we can see or download a lecture or whole semester attendance register with summarization.

That is why we use today our E-Lection parallel, which logs every student's action, use faculty authentication system and produces downloadable lists to analyze student's activity [10–13, 29–31].

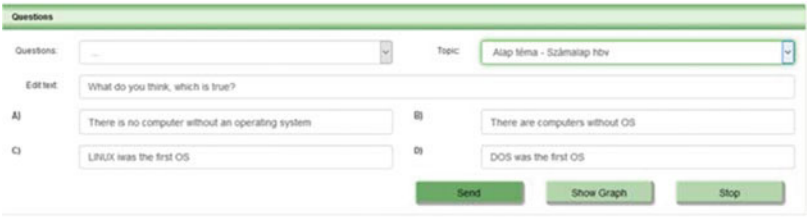


Fig. 6 E-Lection, on the fly questioning panel

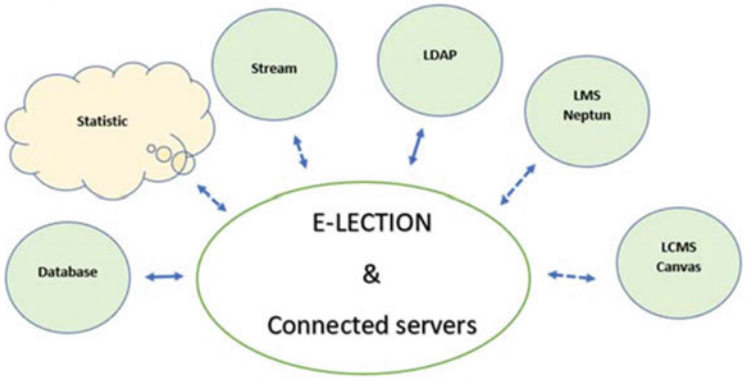


Fig. 7 E-Lection, planed model (dotted lines mean under testing)

6.1 New Plans

In survey [19] we got feedbacks from students who complaint about using several university systems parallel—like Neptun, Canvas, Teams, Faculty' stream server, some homework checking web applications etc. Some of them has got only a mobile phone to connect therefore it is hard to follow (Fig. 7).

That is why we should like to create a more complex system adding streaming functionality, exam and practice module and a statistical module to our E-Lection. To handle not answered questions we need to join the system to Neptun (LMS) and send feedbacks to students' practice teachers too. We already tested the streaming module and there is the first version of connecting E-Lection to Canvas (university LCMS).

7 Summary

We state together with other researchers that active learning and interactivity is one of the key elements of successful learning and teaching. Before COVID-19 period we implemented a lecture management system to involve students better with activating them. We state during the emergency situation interactivity plays an even more important part of education. Professional VCS systems make the deal but remains some unsolved problems. Due to our experiences and students' feedbacks we used our system parallel and planned a more complex one. The new version of E-Lection is to be connected to other university IT systems. The work never stops, there are so many ideas we should implement to improve educational quality.

Acknowledgements EFOP-3.6.3-VEKOP-16-2017-00001: Talent Management in Autonomous Vehicle Control Technologies–The Project is supported by the Hungarian Government and co-financed by the European Social Fund.

References

1. M. Prensky, Digital Natives, Digital Immigrants. Horizon **9**(5), 1–6 (2001)
2. Z. Vahed, L. Zanella, S.C. Want, Students' use of information and communication technologies in the classroom: uses, restriction, and integration. Active Learn. Higher Educ. Sage J. (2019)
3. N.K. Hayes, Hyper and deep attention: the generational divide in cognitive moods. Profession 187–199 (2007)
4. T Bradberry, Multitasking damaging your brain and career, new studies suggest. https://www.forbes.com/sites/travisbradberry/2014/10/08/multitasking-damages-your-brain-andcareer-new-studies-suggest/2/#6088a80642ef. Accessed 7 Apr 2021
5. J.L. Brown, Quick, click: Student response systems evolve in higher ed, New student response systems offer increased versatility, University Business. https://universitybusiness.com/quick-click-student-response-systems-evolve-in-higher-ed. Accessed 7 Apr 2021
6. H. Dangel, C. Wang, Student response systems in higher education: Moving beyond linear teaching and surface learning. J. Educ. Technol. Dev. Exch. **1**(1), 93–104 (2008)
7. S. Mader, F. Bry, Audience response systems reimagined, in *Advances in Web-Based Learning–ICWL 2019*, ed. by M. Herzog, Z. Kubincová, P. Han, M. Temperini. Lecture Notes in Computer Science, vol. 11841 (Springer, Cham, 2019), pp. 203–216
8. Y.J. Wu, T. Wu, Y. Li, Impact of using classroom response systems on students' entrepreneurship learning experience. Comput. Hum. Behav. **92**, 634–645 (2019)
9. E.J. Theobald, M.J. Hill, E. Tran, S. Angrawal, E.N. Arroyo, S. Behling, Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math, in *PNAS (Proceedings of the National Academy of Sciences of the United States of America)*, ed. by S.T. Fiske, vol. 117, no. 12 (Princeton University, Princeton, NJ, 2020), pp. 6476–6483.
10. R. Zitny, T., Szabó, Z., Illés, H.V. Bakonyi, I. Pšenáková, Education using mobile technologies, in *ICETA 2016* (IEEE Computer Society Press, Denver, 2016), pp. 115–120
11. H.V. Bakonyi, Z. Illés, Real-time tool integration for lectures, in *15th IEEE International Conference on Emerging Elearning Technologies and Applications: ICETA 2017* (IEEE Computer Society Press, Denver, 2017), pp. 31–36
12. V.H. Bakonyi, Z. Illés, Real time classroom systems in teachers training, in Lecture Notes in Computer Science 11169. (2018), pp. 206–215

13. Z. Illés, H.V. Bakonyi, Z. Jnr. Illés, Supporting dynamic, bi-directional presentation management in real-time, in *11th Joint Conference on Mathematics and Computer Science*, ed. by E. Vatai, vol. 2046, CEUR-WS.org (2016), pp. 113–118
14. B. Orrù, J. Szekeres-Bartho, M. Bizzarri, A.M. Spiga, V. Unfer, Inhibitory effects of Vitamin D on inflammation and IL-6 release. A further support for COVID-19 management? Eur. Rev. Med. Pharmacol. Sci. **24**(15), 8187–8193 (2020)
15. P. Szabó, M. Ferencová, E. Szabóová, Assessment of pandemics impact on public health and aviation sector by using mathematical modelling. Acta Avionica J. **23**(2), 14–21 (2020)
16. Á. Kövér, A. Antal, I. Deák, Civil society and COVID-19 in Hungary: the complete annexation of civil space. Nonprofit Policy Forum **12**(1), 93–126 (2021)
17. J.I. Olszewska, The virtual classroom: a new cyber physical system, in *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*. IEEE (2021), pp. 187–192
18. E. Kővári, G. Bak, University Students' online social presence and digital competencies in the COVID-19 virus situation, in *Bridges and Mediation in Higher Distance Education. HELMeTO 2020*, ed. by L.S. Agrati et al. Communications in Computer and Information Science, vol. 1344 (Springer, Cham, 2021), pp. 158–169
19. Distance Learning surveys: https://forms.gle/cDf7McuVmDaMxvfUA. English version, https://forms.gle/SzDhE7gUsu4XRqFM8. Hungarian version. Accessed 7 Apr 2021
20. I. Pšenáková, P. Pšenák, U. Kováč, Skúsenosti a poznatky z on-line vzdelávania počas pandémie covid-19, in *New Methods and Technologies in Education, Research and Practice: Proceedings of XXXIII. Didmattech 2020 Conference Budapest*, ed. by A. Abonyi-Tóth, V. Stoffa, L. Zsakó. (ELTE Informatikai Kar, Budapest, 2020), pp. 110–118
21. V. Bakonyi, Z. Illes, C. Verma, Key element in online education to activate students with real-time tools, in *Institute of Electrical and Electronics Engineers 2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM) Conference: Dubai, United Emirates 2021* (Curran Associates, Red Hook (NY), 2021), pp 326–331
22. V. Bakonyi, Z. Illés, Real-time and digital solutions in education during emergency situation in Hungary, in New Methods and Technologies in Education, Research and Practice: Proceedings of XXXIII. Didmattech 2020 Conference Budapest, ed. by A. Abonyi-Tóth, V. Stoffa, L. Zsakó. (ELTE Informatikai Kar, Budapest, 2020), pp. 231–240
23. V. Bakonyi, Z. Illés, Real-time online courses during emergency situation in Hungary, in *ICETA 18th International Conference on Emerging Elearning Technologies and Applications November 12 Stary Smokovec*, Slovakia, (2020), under press
24. F. Martin, M. Parker, D. Deale, Examining interactivity in synchronous virtual classrooms. Int. Rev. Res. Open Distrib. Learn. **13**(3), 227–261 (2012)
25. M. Bower, Virtual classroom pedagogy. ACM SIGCSE. Bulletin **38**(4), 148–152 (2006)
26. S. Alhat, Virtual classroom: a future of education post-COVID-19. Int. J. Educ. **8**(4), 101–104 (2020)
27. R. Li, Communication preference and the effectiveness of clickers in an Asian university economics course. Heliyon **6**(4), 1–5 (2020)
28. J. Byrne, M. Furuyabu, J. Moore, T. Ito, The unexpected problem of classroom video conferencing: an analysis and solution for google hangouts and Jitsi meet. J. Foreign Lang. Educ. Technol. **5**(2), 302–320 (2020)
29. V. Bakonyi, Z. Illés, C. Verma, Real-time education in emergency situation, in *Institute of Electrical and, Electronics Engineers 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)* (Amerikai Egyesült Államok: IEEE, Piscataway (NJ), 2021), p. &
30. V. Bakonyi, Z. Illes, C. Verma, Towards the real-time analysis of talks, in *International Conference on Computation, Automation and Knowledge Management Dubai, United Arab Emirates*, ed. by A.K. Chauhan, G. Singh (Amity University, 2020), pp. 322–327
31. V. Bakonyi, Z. Illes, C. Verma, Analyzing the Students' attitude towards a real-time classroom response system, in *Institute of Electrical and Electronics Engineers 2020 International Conference on Intelligent Engineering and Management (ICIEM)* (IEEE, London, 2020), pp. 69–73

Cost-Efficient BAT Algorithm for Task Scheduling in Cloud

Yagya Malik, Daanish Goyal, Abhiti Sachdeva, and Punit Gupta

Abstract Cloud computing is transforming the way small companies and big corporations operate in the coming generations. If the need for cloud computing grows, so does the need for effective resource management in the cloud world to satisfy customer needs. The aim of traditional resource allocation algorithms is to reduce the overall cost and time spent on all tasks. However, in cloud computing systems, computing capability varies depending on the resource (public vs private clouds), and therefore, the expense of resource use varies as well. As a result, it is critical to consider the resource consumption expense. As a result, in this paper, a cost-effective Bat algorithm for job scheduling in cloud computing architecture is proposed. The proposed algorithm is put to the test against current algorithms in terms of execution cost and resource use. As opposed to current algorithms, the proposed algorithm has a lower cost.

Keywords Cloud computing · Optimization · Virtual machines · Resource optimization

1 Introduction

Cloud computing is a quickly expanding paradigm of information engineering that provides a variety of resources. Managing excessive resource demand is quickly becoming a crucial issue. There is often a trade-off between resource usage and efficiency: higher resource utilization normally leads to better performance, but at a higher cost. Multiple types of software process activities can necessitate different resources in terms of computing power and/or implementation options. These specifications include the cloud provider (if using public clouds), the virtual machine image,

Y. Malik · D. Goyal · A. Sachdeva · P. Gupta (✉)
Department of Computer Communication & Engineering, Manipal University Jaipur, Jaipur, India

A. Sachdeva
e-mail: Abhiti.179301015@muj.manipal.edu

Y. Malik · D. Goyal · A. Sachdeva
Department Computer Science, University of Florida, Gainesville, Florida, USA

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_48

the machine type, and the number of machines (in case of a distributed activity).To tackle the challenges mentioned above on the execution cost of task scheduling in cloud computing environments, this paper proposes a cost-aware Bat algorithm for task scheduling in cloud infrastructure.

In recent decades, the cloud has arisen as a technological model in which all types of applications are virtually delivered to consumers. It is no longer necessary to provide physical facilities because the cloud will have resources ranging from a tiny piece of software to a massive system. As the business is migrating to the cloud so the challenges to maintain performance, integrity, and security are also increasing. There is no conclusive definition of cloud as everyone has their own definition. But we can say cloud computing is a scalable on-demand service mechanism that works on pay-as-you-use terminology. Resources are always available online so it is usable 24×7. Cloud is categorized according to the use and the services it provides which is discussed later. Cloud growth is evolutionary as most organizations are moving onto the cloud and this number is expected to increase in the upcoming years. With moving on the cloud, complexities are also increasing like migration issue, proper resource utilization, security risks. So it is also a big challenge for the service providers to maintain SLA while providing QoS. The provisioning of tasks on virtual machine performs using static, dynamic and meta-heuristic approaches is the main issue. Broad classification of scheduling in a scalable cloud environment performs virtual machine level. The first category includes user request mapping on virtual machine and the second category includes virtual machine mapping on the host machine. The primary focus of the scheduling of tasks on virtual machines and virtual machines on host improve the performance evaluation parameters time and cost. The optimal assignment of the tasks-vm and vm-host mapping comes under the category of Np-hard problem.

In cloud is based on pay-per-use model for delivery of services which plays an important role to attract big and small business to use cloud services rather that deploying own IT infrastructure. In this work, cost-aware scheduling is proposed, which is inspired by nature-inspired bat behavior. The proposed model is inspired by the search behavior of bat which helps to optimize the problem of finding the least cost solution in the cloud. The proposed model is compared with existing work. The experiments show that the proposed model gives better performance than the existing model.

The rest of this manuscript is organized as follows. Section II covers the related works. Section III includes the motivations of task scheduling optimization. Section IV presents the proposed ANN-WOA approach and its implementation details. The performance evaluation and analysis is included in Section V, and finally, section VI concludes the work.

2 Related Work

There have been many studies related to cost optimization, energy efficiency, resource scheduling, and scaling with different solutions; some of the related works are stated in this paper.

Yang et al. [1] predicted workload in their regression model. The predicted workload is later used to auto scale virtual resources at different levels of different service clouds. The proposed mechanism combines real-time scaling and pre-scaling. Ahn et al. [2], has stated that most of the scaling methods are supporting only CPU utilization and data transfer frequency. So, they proposed an auto scaling method that considers the execution deadline and characteristics of the application. Their method categorized two kinds of job patterns; Bag of task jobs or workflow jobs. Run Time Scaling and Performance-oriented scheduling algorithms are proposed for the bag of task jobs and a Workflow scheduling algorithm is proposed for workflow jobs. Sakthi Saravanankumar et al. [3], presented an approach to increase the capability of VM, while in traditional approach, it is very challenging to maintain scaling the virtual machine up to the physical machine capability. In this vertical scaling, a variety of CPUs are grouped together and pinned for peak conditions.

Wang et al. [4], proposed an availability-aware approach to determine the computing resource allocation using vertical and horizontal scaling. They proposed one availability approach model for both horizontal and vertical scaling. Kirthica et al. [5], In their approach they split the request dynamically according to the availability of resource. In their residue-based technique, they performed horizontal scaling. They enhanced the elasticity of the cloud by proposing a greedy technique to rank clouds which improved the resource provisioning technique. Priya et al. [6] combined Min–min and Max–min strategy and proposed a load balancing technique RASA (Resource-Aware Scheduling algorithm). At the first phase, virtual nodes are created, and then the response time of each Vm is calculated. When the least loaded node is found, it is allotted to the client. Max–min is applied if numbers of resources are even, else Min–min strategy is used.

Tesfatsion et al. [7] in their feedback controller mechanism determined optimal configuration to minimize the energy consumption while maintaining the SLA. The controller parameters are adjusted at run time to achieve objective performance under different workload conditions. Their approach combined VM count, number of cores, and scaling the frequencies to control cloud data center in an energy-efficient manner. Karthikeyan et al. [8]. To reduce the energy consumption while VM migration, they have proposed the Artificial Bee Colony–Bat Algorithm (ABC–BA) with Naive Bayes classifier. The performances were compared using the success and failure index and energy consumption. The proposed model was able to minimize the energy consumption and failure rate. Kansal et al. [9] proposed a model based on the firefly approach. Their energy-aware virtual machine migration approach transfers the maximally loaded virtual machine to the minimum active node while maintaining the performance and energy (Table 1). Other similar work proposed of

Table 1 Comparative study

Research paper	Input parameters	Parameters optimized/conclusion
[13]	CPU, memory, bandwidth	Memory utilisation and response time, turn around time
[14]	Tasks and lists of VM	ACO is better than RR and FCFS
[15]	CPU, memory, bandwidth	Cost of processing and communication time
[16]	user privilege, expectations, task length, pending time in queue	Makespan, average latency, load balancing index
[17]	Tasks and VM	Ma + B7 + B2:E7 + B4:E7 + B2:E7 + B7 + B2:E7 + B7 + B2:E7

task scheduling are proposed using auto scaling, and energy efficiency [10–12] in current scenario.

3 Proposed Model

In this section, we have proposed a Bat algorithm based on the behavior of "Microbats". Microbats use their ability of echolocation to navigate and "see" the world around them. They use echolocation to hunt, locate other bats, avoid various obstructions such as predators and trees during flight.

This is done by emitting a very loud and short sound pulse. When this sound pulse bounces off other elements, bats analyze the feedback. In context to the cloud model, we simulate virtual bats and update their parameters by keeping the bats in an array and iterating through it. These parameters include position, velocity (which are performed first), loudness, and pulse emission rate. After each iteration, the frequency of echolocation, position, and velocity is denoted by

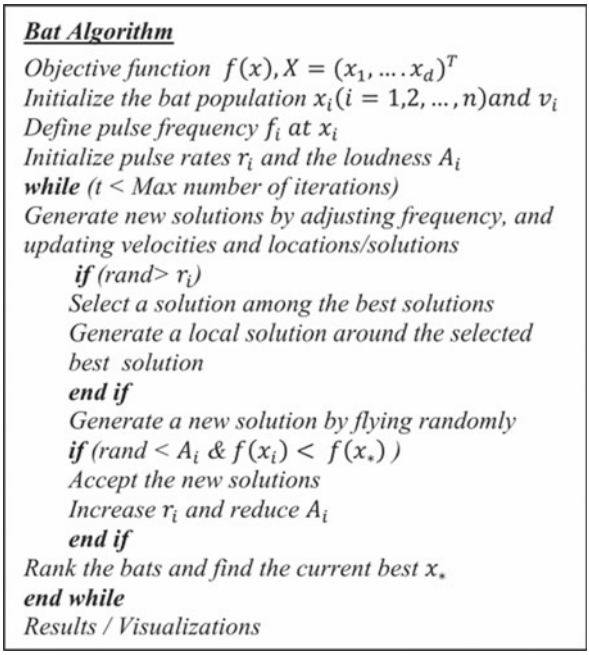
$$Frequency(i) = Frequency_{\min} + (Frequency_{\max} - Frequency_{\min})\beta \tag{1}$$

$$velocity_i(t) = velocity_i(t-1) + (Position_i(t) - position_{globalbest}\, Frequency_i) \tag{2}$$

$$Position_i(t) = Position_i(t-1) + velocity_i(t) \tag{3}$$

$$fitness\ value_i = \sum_{j=1}^{j=n} \alpha * \cos t_j + \beta * task\ completion_time_j \tag{4}$$

Fig. 1 Proposed bat algorithm



where $\alpha + \beta = 1$.

β can be either 0 or 1, it is picked randomly from a uniformly distributed vector. Position $_{\text{Global Best}}$ Location is the current local best location. It is determined after comparing all the solutions among all the bats. This means that the bats will adjust their loudness and rate of pulse emission with respect to the distance between the bats and their prey. This happens in the form of iterations (Fig. 1).

4 Simulation Results

In this section, the comparative study of the proposed algorithm is compared with the existing work. The simulation is done using cloudsim 3.0 and the dataset is used from a parallel workload repository. The comparison is done with GA-EXE (genetic algorithm with execution time as fitness function) and GA-Cost (genetic algorithm with cost as fitness function) (Table 2).

The simulation is done using the following parameters:

In Fig. 2, a comparison of the proposed cost-aware bat algorithm (BAT-Cost) with the existing algorithm is shown with increasing task load. The comparison shows that the proposed algorithm process provides better execution time as compared to existing algorithms.

In Fig. 3, a comparison cost of the proposed cost-aware bat algorithm (BAT-Cost) with the existing algorithm is shown with increasing task load. The comparison

Table 2 Simulation Parameters

Parameter	Value
Iteration	100
Population	100
Crossover rate	0.5
Learning rate	0.2
Velocity	0- VM count
Alpha	0.2
Gama	1

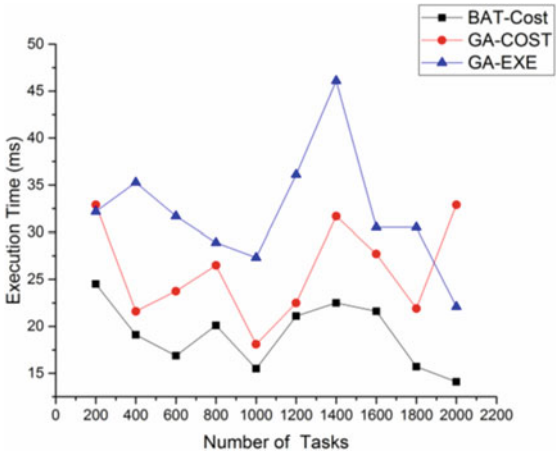


Fig. 2 Comparison of execution time

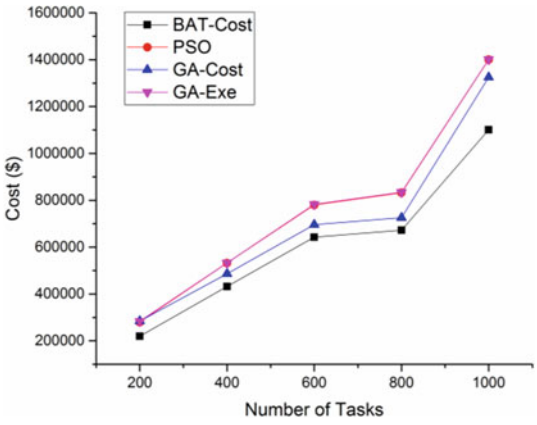


Fig. 3 Comparison of cost time vs increasing task load

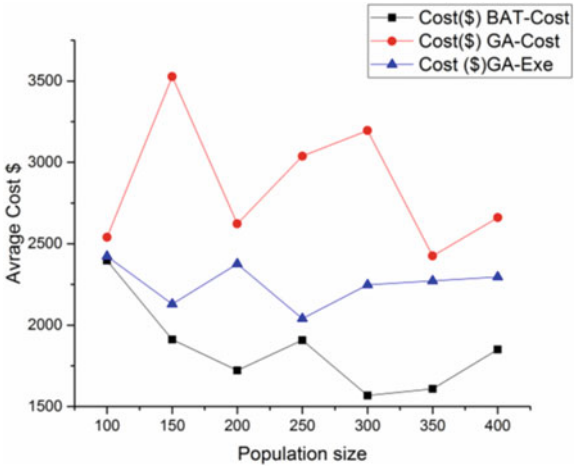


Fig. 4 Comparison of cost time vs population size

shows that the proposed algorithm process provides better processing cost time as compared to the existing algorithms.

Figure 4 shows a comparison cost of the proposed cost-aware bat algorithm (BAT-Cost) with the existing algorithms, which is shown with an increasing number of populations. The comparison shows that the proposed algorithm process to provide better processing cost time as compared to existing algorithms.

5 Conclusion

It is clear that many of the solutions are available with different approaches and more need to be found yet. The algorithms are approaching a single factor like energy, migration, or scaling. The results are quite effective and the performance of the cloud has been significantly increased with cost-efficiency. Resource cost optimization is a key factor in cloud computing and a more improvised way to manage the resources is yet to be found. The rapid change in technologies and managing the performance with client satisfaction will be a challenging task. The proposed model provides a better computational cost than the existing models. In the future, the proposed algorithm can be used to improve energy efficiency.

References

1. J. Yang, C. Liu, Y., Shang, Z. Mao, J. Chen,. Workload predicting-based automatic scaling in service clouds, in *2013 IEEE Sixth International Conference on Cloud Computing* (2013, June), pp. 810–815
2. Y. Ahn, J. Choi, S. Jeong, Y. Kim, Auto-scaling method in hybrid cloud for scientific applications, in *The 16th Asia-Pacific Network Operations and Management Symposium*, (2014, September), pp. 1–4
3. P. Sakthi Saravanankumar, M. Ellappan, N. Mehanathen, CPU resizing vertical scaling on cloud. Int. J. Future Comput. Commun. **4**(1), 1–12 (2015)
4. W. Wang, H. Chen, X. Chen, An availability-aware virtual machine placement approach for dynamic scaling of cloud applications, in *2012 9th International Conference on Ubiquitous Intelligence and Computing and 9th International Conference on Autonomic and Trusted Computing* (2012, September) pp. 509–516
5. S. Kirthica, R. Sridhar, A residue-based approach for resource provisioning by horizontal scaling across heterogeneous clouds. Int. J. Approx. Reason. **101**, 88–106 (2018)
6. S.M. Priya, B. Subramani, A new approach for load balancing in cloud computing. Int. J. Eng. Comput. Sci. **2**(5), 1636–1640 (2013)
7. S.K. Tesfatsion, E. Wadbro, J. Tordsson, A combined frequency scaling and application elasticity approach for energy-efficient cloud computing. Sustain. Comput. Inf. Syst. **4**(4), 205–214 (2014)
8. K. Karthikeyan, R. Sunder, K. Shankar, S.K. Lakshmanaprabu, V. Vijayakumar, M. Elhoseny, G. Manogaran, Energy consumption analysis of Virtual Machine migration in cloud using hybrid swarm optimization (ABC–BA). J. Supercomput. **76**(5), 3374–3390 (2020)
9. N.J. Kansal, I. Chana, Energy-aware virtual machine migration for cloud computing-a firefly optimization approach. J. Grid Comput. **14**(2), 327–345 (2016)
10. J. Zheng, T.E. Ng, K. Sripanidkulchai, Z. Liu, Pacer: a progress management system for live virtual machine migration in cloud computing. IEEE Trans. Netw. Serv. Manag. **10**(4), 369–382 (2013)
11. Y. Ahn, J. Choi, S. Jeong, Y. Kim,. Auto-scaling method in hybrid cloud for scientific applications, in *The 16th Asia-Pacific Network Operations and Management Symposium* (IEEE, 2014, September), pp. 1–4
12. S.K. Tesfatsion, E. Wadbro, J. Tordsson, A combined frequency scaling and application elasticity approach for energy-efficient cloud computing. Sustain. Comput. : Inf. Syst. **4**(4), 205–214 (2014). (Author, F.: Article title. Journal 2(5), pp. 99–110)
13. M.B. Gawali, S.K. Shinde, Task scheduling and resource allocation in cloud computing using a heuristic approach. J. Cloud Comput. **7**(1), 1–16 (2018)
14. M.A. Tawfeek, A. El-Sisi, A.E. Keshk, F.A. Torkey, Cloud task scheduling based on ant colony optimization, in *2013 8th International Conference on Computer Engineering & Systems (ICCES)*, (2013, November), pp. 64–69
15. L. Guo, S. Zhao, S. Shen, C. Jiang, Task scheduling optimization in cloud computing based on heuristic algorithm. J. Netw. **7**(3), 547 (2012)
16. X. Wu, M. Deng, R. Zhang, B. Zeng, S. Zhou, A task scheduling algorithm based on QoS-driven in cloud computing. Proc. Comput. Sci. **17**, 1162–1169 (2013)
17. M. Abdullahi, M.A. Ngadi, Symbiotic organism search optimization based task scheduling in cloud computing environment. Futur. Gener. Comput. Syst. **56**, 640–650 (2016)

Systemic Thinking in Programming Education

Szilárd Korom and Zoltán Illés

Abstract The programming education focuses on algorithmic thinking competence, while it ignores many other informatics competencies. Most programming problems cannot be solved just using this competence, because information systems are much more complicated, and require many other fields and knowledge. In this article, we would like to clarify how systemic thinking competence comes in, how it can be useful during programming education. We believe that systemic thinking could be improved by changing the methodology over the already existing educational fields. This article provides a general structure on how to formulate our curricula so that the students will be able to get closer to understanding the big picture.

Keywords Systemic Thinking · Systems Thinking · Informatics Competences

1 Introduction

One of the most important questions in programming education is what informatics competencies we would like to improve. In elementary schools and in high schools, it is definitely about the competencies, but in software developer-oriented education, the main goal is still to pick up the mandatory competencies. Fortunately, they are already well defined [1].

If we have a look at some university course syllabuses [2–4], or if we read some educational research projects, we could experience that they are focusing on the algorithmic thinking competence. Furthermore, the tools, games, and smart devices that can be used during informatics education are focusing on this particular competence too [5–7]. Sometimes, they change the tone to a technology or a development environment, but we could say that the schema is the same:

S. Korom (✉) · Z. Illés
Faculty of Informatic/Department of Media & Educational Tech, Eötvös Lóránd University, Budapest, Hungary
e-mail: szicsa@inf.elte.hu

Z. Illés
e-mail: illes@inf.elte.hu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_49

1. Learn the basics of the development environment, the most important phrases, and definitions.
2. Build a simple project from Hello World application to something more interesting.
3. Solve more and more difficult problems in the environment by learning advanced details about the technology, language, problem-solving strategies, algorithms.

In this article, we would like to show that something is missing from this structure: a competence that could be really useful in many other segments of life, and inevitable for a developer. This one is the systemic thinking competence.

The main goal is to have a general concept of how we should present, teach particular items of informatics education to improve systemic thinking as well since it can be learned and practiced through the normal projects that we already use. Another interesting area is how to measure if we could successfully improve competence.

2 Systemic Thinking

There are many articles that tried to define systems thinking [8–13]. Arnold and Wade [14] tried to collect and generalize it in one single definition as follows:

Systems thinking is a set of synergistic analytic skills used to improve the capability of identifying and understanding systems, predicting their behaviors, and devising modifications to them in order to produce desired effects. These skills work together as a system.

In short, the authors identified systemic thinking as a system. The issue is that the collection of the individual components, solutions for the sub-problems do not indicate the right behavior of the whole. The algorithmic thinking in a complex system that does not always lead to a solution (or not fast enough), because the outcome of the events is often determined by the structure and not by the states of the individual components.

Ackoff is very clear about it [14]:

The systems approach to problems focuses on systems taken as a whole, not on their parts taken separately. Such an approach is concerned with total-system performance even when a change in only one or a few of its parts is contemplated because there are some properties of systems that can only be treated adequately from a holistic point of view. These properties derive from the relationships between parts of systems: how the parts interact and fit together. In an imperfectly organized system even if every part performs as well as possible relative to its own objectives, the total system will often not perform as well as possible relative to its objectives.

What does it mean in a concrete software system? Let us have a real-time chat solution, where no matter how many clients are running, each of them can communicate with one another through a real-time database. From an algorithmic point of view, the problem is not difficult, since the client application just provides an interface for text inputs and outputs (maybe authentication), reads the incoming message

and sends the user input to a given recipient. However, the most important feature, which is the real-time chat functionality itself is a property of the whole. None of its sub-groups have this property and the reason why it is working is that the parts fit together. The clients are using the right endpoint for reading and writing, the message structure is well defined, so it can be parsed by the applications, and the interfaces are working, so a user can enter the system.

2.1 Problems

We cannot speak about systemic thinking without having a particular problem. Systemic thinking is a way how we think about a problem. It is the ability to see and understand a system in a mess [15]. Most of these problems require a simulation, to be able to understand and observe the whole. As Gallón L. pointed out [16] after Forrester [17]:

"*Systems thinking + simulation = systemic thinking*"

Thus, in order to improve systemic thinking, we will need examples and simulations of real problems, that require a software or technology system, in hope that we can practice this "new" way of thinking.

We will need a general form, a model that can be used when we think about how a curriculum should be structured. The model must be a helping tool, that defines layers of how we think about a problem, therefore, we will know exactly how a "learner" can reach the "systemic level" in a specific part of his/her studies. We will use the DIKUW model [18].

3 The DIKUW Model

The DIKUW hierarchy is a representation of how we think about a mess. The layers are: Data, Information, Knowledge, Understanding, and Wisdom (Fig. 1).

The diagram is very clear about the hierarchy. Furthermore, the students will have to stay between the Complicated-Complexity border and Mundane-Complexity border. The hierarchy and the borders provide a pattern of how we should organize something that we would like to teach. We need to drive the students higher and higher with every taught element to improve the skill to understand context independent systemic problems. The meaning of each layer must always be defined. In the following chapter, we try to explain what they mean in informatics education.

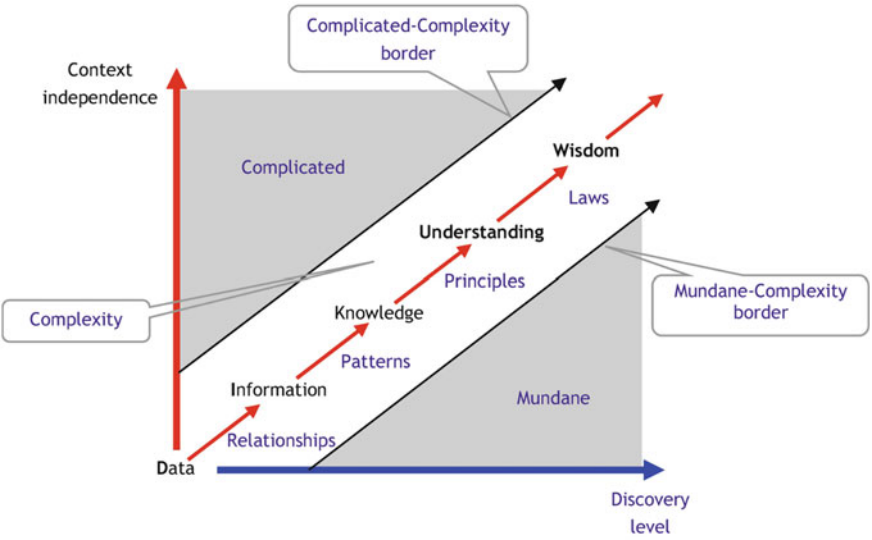


Fig. 1 DIKUW hierarchy and complexity [19]

3.1 Data

"*Data are symbols that represent the properties of objects and events*" [18].

Imagine a student who would like to learn programming on her/his own, they buy a book that contains the keywords, commands of a programming language, and some sample codes. In that situation, the code and the commands are just data for the student since they cannot do anything with it. They do not understand the relationship between the linguistic elements. The code, which is a mess, remains as an incomprehensible text, nothing more than any other strings. The data is a code, a programming environment, the rules of a programming paradigm or a technology.

3.2 Information

"*Information is…endowed with meaning, relevance and purpose*" [20].

Once data has been processed, we have the information. In programming education, it is usually the point where the program becomes useful for a student, i.e., when they can execute it. The description above the code is the information.

3.3 Knowledge

"*Knowledge…exists in the mind of the knower…*" [21]

When information is processed by the student, it turns into knowledge. The information becomes knowledge after answering "how to…" questions related to the programming environment or the code. The product of knowledge is a skill, what we usually call programming. When a student can write or modify a code based on the learned patterns, they have the knowledge of programming.

3.4 Understanding

"*Understanding is related to causes (why) by knowledge principles*" [16].

During programming, the environment is always connected to something else. For example, shell scripting is based on the principles of the operating system. The understanding layer is about the connection between the elements of the information system. When the student can answer questions like: Why do we use this programming environment? Why is this the most efficient algorithmic solution?

3.5 Wisdom

Wisdom is about the future, rather than the other four categories. It is definitely systemic because, on this layer, we have the ability to see what we do not understand. We can formulate new questions that we cannot immediately answer, we can predict the behavior of the system.

By wisdom, we can design new systems, where the current model is applicable and figure out new situations where we can use our solution. Wisdom is the high-level perspective above the original problem.

3.6 Why the DIKUW Model?

Without a model like DIKUW, it would be really hard to create a curriculum, that guides the students from the information level to the full understanding by demonstrating the new element in a whole context. A DIKUW-based curricula helps the students put the new data into the big picture. The model represents a specific structure beyond a curriculum and a restriction at the same time since it must include a system around the taught component. The next chapter shows that this is an actual problem, that we need to give an answer to.

4 Current Curricula and DIKUW Hierarchy

When we teach the features of the Excel program, for example, the important thing is not the program itself, because in real life, there is no such thing as an Excel problem (like drawing a simple diagram for data). What is more important is to make the students be able to recognize when a specific problem can easily be solved by using this application.

All in all, we could say that the context is essential. The best we can do is to demonstrate a whole system, where the particular knowledge item is just one part with meaningful examples for the students to see where a new tool is applicable.

In one of our previous articles [22], we wanted to prove that by missing the context and putting the knowledge into a system, the students are not able to have the understanding. In that research, we asked university students (205 first semester computer science) why a specific technology is useful, why is it better than another. We would like to show the results of some of our questions.

The first question was: Why do we store data in databases and not in files? There were no options, they had to answer the question in a free text format. We categorized the answers, and the results can be seen in Fig. 2.

Our conclusion was that the students have absolutely no idea about the advantages of a database. The issue here is that they are probably not able to decompose complex problems or systems, where a database is absolutely critical.

The second question was: What is the difference between a table and a database? There were no options either, so we had to categorize the answers. It was pretty hard since they were really mixed. Our result [22] can be seen in the following figure:

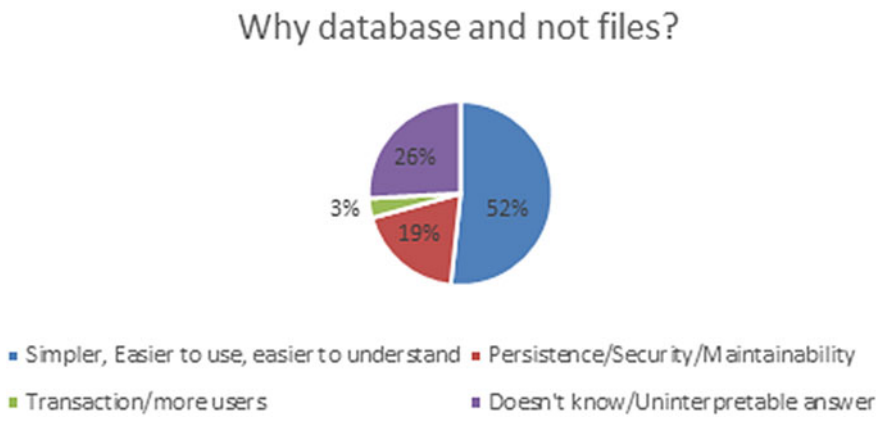


Fig. 2 Counted categorized answers for the first question

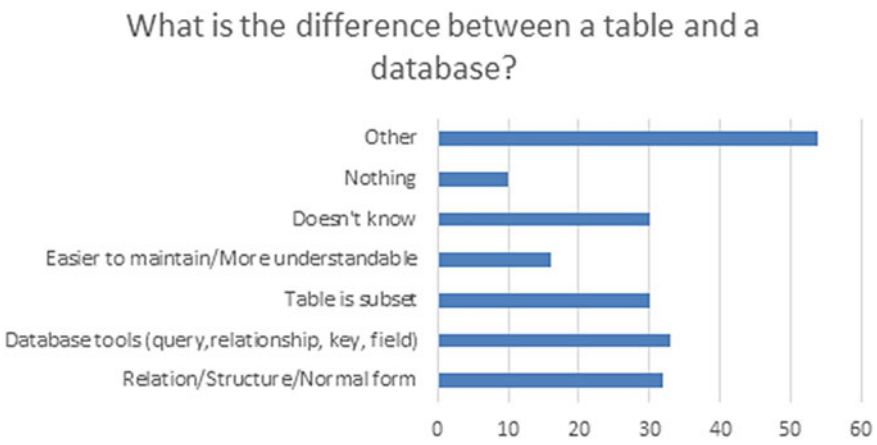


Fig. 3 Counted categorized answers for the second question

5 Overall Example

Imagine a high school student who learns something about databases through the Access program. What does their "knowledge" look like in the DIKUW model? In the following, we try to explain some of the items that a curriculum can reach. This is not a complete list, just a short example.

1. *Data*: We have a program called Access, there is a thing called "database" which looks something like a table.
2. *Information*: Access has a "query" tool to get a subset of the data, which is more like a function in Excel.
3. *Knowledge*: We can make queries to solve the exercises.
4. *Understanding*: As you can see on Figs. 2 and 3, they don't really reach this level, but it should be the point where they can answer a similar question
5. *Wisdom*: Of course, the students are not here. With wisdom, they could decompose a problem, where an element of a solution system is a database, or a database management tool.

6 IoT Systems for Systemic Thinking Improvements

Of course, not every part of the IoT education can be used to improve systemic thinking competence. For example, when we create a console-based arithmetic calculator, we can not really speak about systems, like when we do something in a document editor, or in a raster graphics editor. So, what kind of curricula is applicable for this competence? In our opinion, a programmable smart device-based methodology.

By "programmable smart devices" we mean robots, embedded systems, mobile phones, minicomputers, or microcontrollers. If we use them together, we get a system

by definition, because we have multiple components that interact with each other. All in all, we have to create machine-to-machine communication over a network (over the internet), so we have "Internet of Things" or IoT in short. As an example, we can have a smart home, an autonomous system (like autonomous car driving system) project.

The good thing about these systems (from a programming teacher's perspective) is that the components do not require complicated source code. For example, a LED driver microcontroller is not large or complex at all. Thus, the students are not facing an algorithmic problem, but a system problem (plus technological challenges, that can be difficult or easy, depending on the project). In one of our previous articles, we created a figure, that shows the components of an IoT-based curriculum [22], which can be seen in Fig. 4.

Of course, this does not mean, that every single IoT curricula must include all of these elements but collects the possibilities. Figure 4 shows that an IoT-based curriculum can be understood without dealing with systemic problems. They are scalable, spiral, the exercises/problems are about designing and implementing increasingly difficult (but overlapping) systems. The systems can be "difficult" with having the same components (for example, if more and more complex software is included)

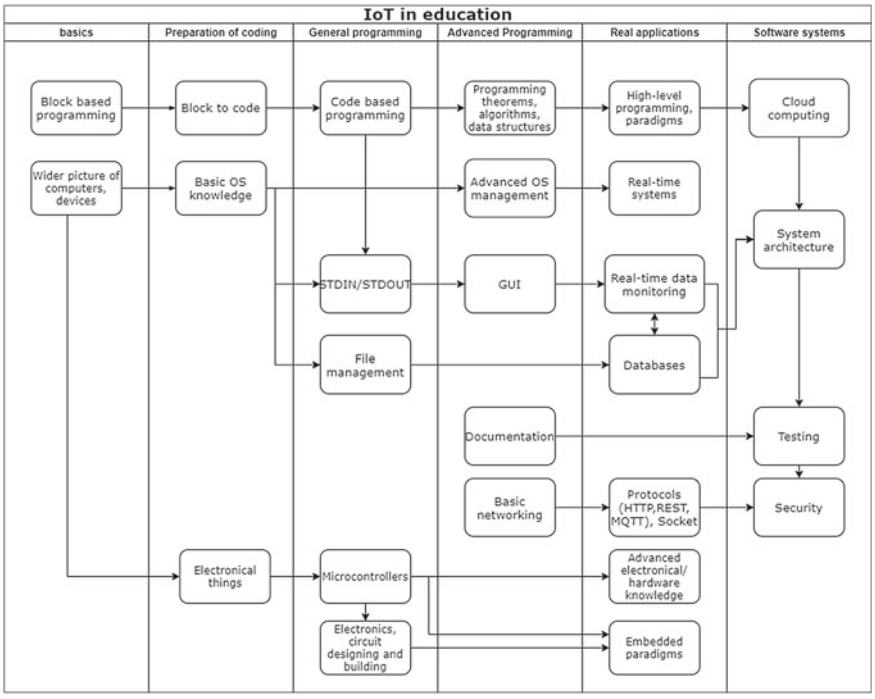


Fig. 4 Components of an IoT system from teaching perspective

or with including new ones. The more components we touch, the more complex system we get. On the other hand, the technological challenge will increase too.

As an example, we would mention the "end requirement" of a university course of ours, which was a fully operational IoT system. Due to the nature of the course (Embedded and Real Time Systems), the emphasis was on the "real-time" part. It is really important, that no specific project idea was given so that all phases of planning were left to the students. The requirements were the followings:

1. The system must be dynamic: The system can be easily extended with a new feature or device, or one can be removed without affecting the system.
2. Has a monitoring solution: It can be a data visualization, a user interface, a complex device control feature, etc.
3. Has a data gathering solution: The system is able to gather information from the real world, from the environment.
4. Only real-time communications: Real-time communication protocols are implemented and used without unnecessary data or bandwidth connections
5. The final system is a simulation of a real problem
6. The project must be demonstrated

Using Fig. 4, such a solution includes the following elements:

A system like that can be easily extended, for example by processing different types of data, or by using a not ready-to-use board (like Raspberry Pi), or by automated tests, deployment strategy etc. (Fig. 5).

7 Measurement

As mentioned above, the context, simulations, and demonstration are very important to deal with systemic thinking. Thus, we must build a whole working system into our curricula to mark the direction, to have a goal, or as a demonstration to show how the learned elements could be built into a wider perspective.

Imagine that we have many Raspberry Pis and sensors with which we can build a system that gathers information from the real world. If a student is working with one device, he/she should write one single software, which is pretty simple from an algorithmic point of view. However, in the classroom, we have many other devices, so in case they can transfer their data to a monitoring solution, we could have a whole working system. Of course, the monitoring solution must be done by the teacher, but it will be perfect for a demonstration. After it has been done, we could discuss the details, the differences between the two solutions. The properties of the system may not be fully explored by the students, but they have a context at least, which is really close to a real-life problem (simulation of a smart home system). Eventually, after many projects, the system can be more and more complex, including many other fields of their studies such as a database, graphical solutions, distributed systems, etc.

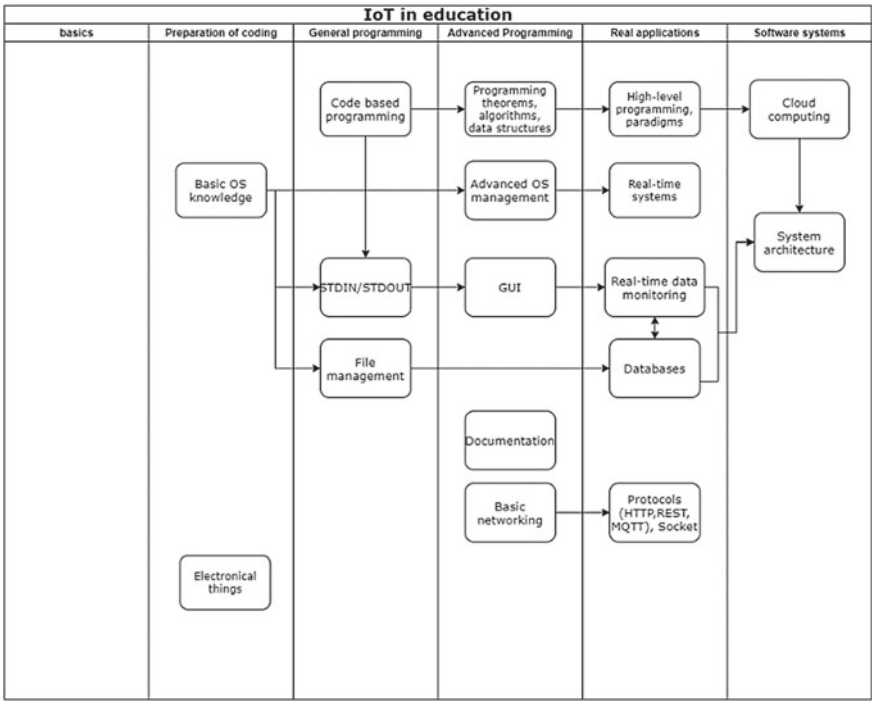


Fig. 5 Elements of a specific IoT solution

The algorithmic thinking competence is usually improved by increasingly more difficult algorithms, tasks or games. The measurement of our success is that the students can understand those difficult algorithms. The learning path of these algorithms is already well defined, for example, by Nikházy [23]. We should follow the same pattern for systemic thinking as well, which means we need increasingly more difficult and complex systems. Since we do not have something like algorithms, we call the elements of systems thinking as a helpful hand [14]. To make sure that we succeed, discuss the details, the properties with the students by following these elements and questions:

Recognizing Interconnections

1. How many elements do we have in the system?
2. What software is in the system?
3. What other IT technologies do we have in the system?

Identifying and Understanding Feedback

1. Who communicates with whom in the system?
2. How are they connected?
3. How do they communicate with one another?
4. How should we integrate a new component into the system?

Understanding System Structure

1. What does the system look like as a whole?
2. What are the inputs of the system?
3. What are the outputs of the system?
4. What intermediate states may exist?

Differentiating Types of Stocks, Flows, Variables

1. What are the properties of the system? (not the properties of a part of the system)
2. What are the external/internal dependencies of the system?
3. How does a particular input go through the system?
4. How could we visualize the different flows for different inputs?

Identifying and Understanding Non-Linear Relationships

1. Do we have a parallel behavior somewhere in the system?
2. Can we change it to a sequential behavior? What is the difference?
3. Could this relationship cause a problem? (for example, multiple file handling)

Understanding Dynamic Behavior

1. How dynamic is the system?
2. What are the consequences if we remove a particular element from the system?
3. What are the consequences if we add a new element into the system?

Reducing Complexity by Modeling Systems Conceptually

1. Can we replace a component with a more efficient solution?
2. Are there any unnecessary parts of the system?
3. How could we redesign the system?
4. How could we visualize the system and its properties?

Understanding Systems at Different Scales

1. Is this a scalable system?
2. Are there any sub-systems?
3. Why it is a sub-system and not a single component?

The categorized questions above help us to measure the systemic thinking level of the students. However, it is pretty hard to score the answers. What matters is whether the student can answer a particular question and how complete the answer is and how adequate it is. For example, for the "Are there any unnecessary parts of the system?" question, the important thing is how many elements are listed by the student (compared to the full list).

Thus, these questions can be used in our everyday teaching life as a kind of self-examination. Furthermore, it indicates a research topic as well. The hypothesis is that programmers at different levels would achieve different results. The only thing that is needed is to design a sufficiently complex and general IT system that is understandable only via the description (i.e., not too technology-specific). The system

should also include inaccuracies, design errors that can be corrected, eliminated, or replaced with a more efficient solution.

The above questions are of course general that need to be specified for the current project. For example, "How many elements do we have in the system?" can be reasonable for software units, but also for hardwares, technological components (database, message queue, servers, clusters, etc.)

8 Conclusion

Programming education is based on informatics competencies, especially algorithmic thinking. The issue is that most of the problems cannot be solved by this competence, so curricula must be extended with systemic thinking.

It is not that easy to define systemic thinking and most importantly, not that easy to teach. It seems that it should be taught through examples and demonstrations and by showing the context of the element. Simulations of the problems will be really useful even if the thing that we would like to teach is much smaller than that.

For a curriculum, the teacher must make sure that the students get higher and higher in the DIKUW hierarchy, so they will have the ability to understand a wider point of view later when they will meet a more complex problem. After a while, they will be able to put together the elements and will understand how they fit together. That wisdom will become a skill to design new systems or decompose an already existing one.

Acknowledgments EFOP-3.6.3-VEKOP-16-2017-00001: Talent Management in Autonomous Vehicle Control Technologies – The Project is supported by the Hungarian Government and co-financed by the European Social Fund.

References

1. G. Horváth, P. Szlávi, L. Zsakó, Informatics (ICT) competencies. In: ICAI2010—8thInternational Conference on Applied Informatics Eger, Hungary (2010, January 27–30).
2. Universitat Pompeu Fabra Barcelon, Subject syllabus: The Internet of Things. https://www.upf.edu/pra/en/3376/22580.pdf. Accessed 30 Mar 2021
3. University of Virginia: Angela Orebaugh, Ph.D Securing the Internet of Things. https://collab.its.virginia.edu/syllabi/public/bcca51cc-9823-4bd4-8045-55b7b8f098cf. Accessed 30 Mar 2021
4. Uppsala Universitet: Syllabus for Internet of Things. http://www.uu.se/en/admissions/master/selma/kursplan/?kKod=1DT094. Accessed 30 Mar 2021
5. M Kölling, Educational programming on the raspberry Pi. Electronics **5**, 33 (2016). https://doi.org/10.3390/electronics5030033
6. BBC micro:bit introduction. https://makecode.microbit.org/courses/csintro/introduction. Accesses 30 Mar 2021

7. N. Rusk, M. Resnick, R. Berg, M. Pezalla-Granlund, New pathways into robotics: strategies for broadening participation. J. Sci. Educ. Technol. USA (2007). https://doi.org/10.1007/s10956-007-9082-2
8. B. Richmond, Systems dynamics/systems thinking: let's just get on with it, in *International Systems Dynamics Conference. Sterling, Scotland* (1994)
9. H. Benson, J. Borysenko, A. Comfort, L. Dossey, B. Siegel, Economics, work, and human values: new philosophies of productivity. J. Conscious. Change **7**(2), 198 (1985)
10. L.B. Sweeney, J.D. Sterman, Bathtub dynamics: initial results of a systems thinking inventory. Syst. Dyn. Rev. **16**(4), 249–286 (2000). https://doi.org/10.1002/sdr.198
11. K.A. Stave, M. Hopper, What constitutes systems thinking? A proposed taxonomy, in *25th International Conference of the System Dynamics Society. Boston, MA* (2007)
12. B. Kopainsky, S.M. Alessi, P.I. Davidsen, Measuring knowledge acquisition in dynamic decision making tasks, in *The 29th International Conference of the System Dynamics Society* (Washington, DC,.(2011), pp. 1–31
13. A. Squires, J. Wade, P. Dominick, D. Gelosh, Building a competency taxonomy to guide experience acceleration of lead program systems engineers, in *9th Annual Conference on Systems Engineering Research (CSER)* (Redondo Beach, CA, 2011), pp. 1–10
14. R.D. Arnold, J.P. Wade, A definition of systems thinking: a systems approach. Proc. Comput. Sci. **44**, 669–678 (2015)
15. R.L. Ackoff, The future of operational research is past. J. Oper. Res. Soc. **30**(2), 93–104 (1979). https://doi.org/10.2307/3009600
16. L. Gallón, Systemic thinking, in *Quality Education. Encyclopedia of the UN Sustainable Development Goals*, ed. by W. Leal Filho, A. Azul, L. Brandli, P. Özuyar, T. Wall (Springer, Cham, 2019). https://doi.org/10.1007/978-3-319-69902-8_58
17. J.W. Forrester, System dynamics–a personal view of the first fifty years. Syst. Dyn. Rev. **23**(2–3), 345–358 (2007). https://doi.org/10.1002/sdr.382
18. R.L. Ackoff, From data to wisdom. J. Appl. Syst. Anal. **16**, 3–9 (1989)
19. B. Castellani, F.W. Hafferty, M. Ball, E-social science from a systems perspective: applying the SACS toolkit (2009)
20. J. Rowley, The wisdom hierarchy: representations of the DIKW hierarchy. J. Inf. Sci. **33**(2), 163–180 (2007)
21. H.M. Dr. Gladney, H.M.G. Consulting, P.A. McDonald, Definition 19 on page 483 of Zins, C. Conceptual approaches for defining data, information, and knowledge. JASIST **58**, 479–493 (2007)
22. S. Korom, Z. Illés, Competence Improvements with IoT systems. Central Eur. J. New Technol. Res. Educ. Pract. **3**(1) (2020). https://doi.org/10.36427/CEJNTREP.3.1.474
23. L. Nikházy, A problem-based curriculum for algorithmic programming. Central Eur. J. New Technol. Res. Educ. Pract. **2**(1), 76–96 (2020). https://doi.org/10.36427/CEJNTREP.2.1.399

Technology Based University Identification Model for Real-Time

Chaman Verma, Zoltán Illés, and Veronika Stoffová

Abstract The educational study place (school, college, or university) is a student's prime demographic identity. Automation of the demographic features of the student toward the technology available is a promising task. So, considering the problem, this paper used a two-group Linear Discriminant Analysis (LDA) approach to classify the educational university based on the technology parameters. This research was conducted on 302 actual samples collected from Indian and Hungarian Universities. Out of a total of 37 features, 13 features were found important as suggested by the Analysis Of the Multivariate Variance (MANOVA), Tolerance, and Wilk's lambda (Λ). The paper's findings suggested our selected features acquired validated accuracy of 89.4% to recognize the student's university. This approach has been used for binary classification. It might help the university's response system, such as Google classroom, Microsoft Forms, Google Forms, E-lection, etc., to quickly identify the international study environment's technology opinion.

Keywords Canonical discriminant function · LDA · MANOVA · Technology · Wilk's lambda

C. Verma (✉) · Z. Illés · V. Stoffová
Department of Media and Educational Informatics, Eötvös Loránd University, Budapest, Hungary
e-mail: chaman@inf.elte.hu

Z. Illés
e-mail: illes@inf.elte.hu

V. Stoffová
e-mail: NikaStoffova@seznam.cz

V. Stoffová
Department of Mathematics and Computer Science, Trnava University, Trnava, Slovakia

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_50

1 Introduction and Related Work

Nowadays, artificial intelligence plays a vital role in identifying educational patterns. Several machine learning algorithms have been applied to compare various kinds of datasets from academic settings. A discriminant analysis technique is very famous for classification. It has been used for learning discriminative feature transformations in statistical pattern recognition [1]. It has been used for the last six decades to analyze the data patterns for classification. The LDA has identified the girl student's attrition from their school with an accuracy of 78.33% [2]. It outperformed other machine learning algorithms: K-nearest neighbors, classification and regression trees, Gaussian Naive Bayes, support vector machines to identify the student academic grades in the bachelor courses with an accuracy of 90.74% [3]. It also supported three machine learning algorithms to filter out significant features to recognize the student final examination result [4]. The teacher's growth based on their course student's opinion identified with LDA and attained a good accuracy of 90.5% [5]. It has also been used as dimension reduction to determine the student's attention during engineering education classes [7]. Recently, few demographic features like educational standard and nationality have been predicted with other machine learning algorithms [8, 9].

The earlier literature did not apply LDA to identify the university of the student based on the technology. Also, no one enlightens the automated demographic features of students to evaluate their viewpoints. Hence, this paper is an experiment based on the samples collected from universities to exhibit a new scientific evidence approach.

Figure 1 shows the pictorial view of this paper's idea considered as a preliminary concept. Students need to fill their responses on the computer system about

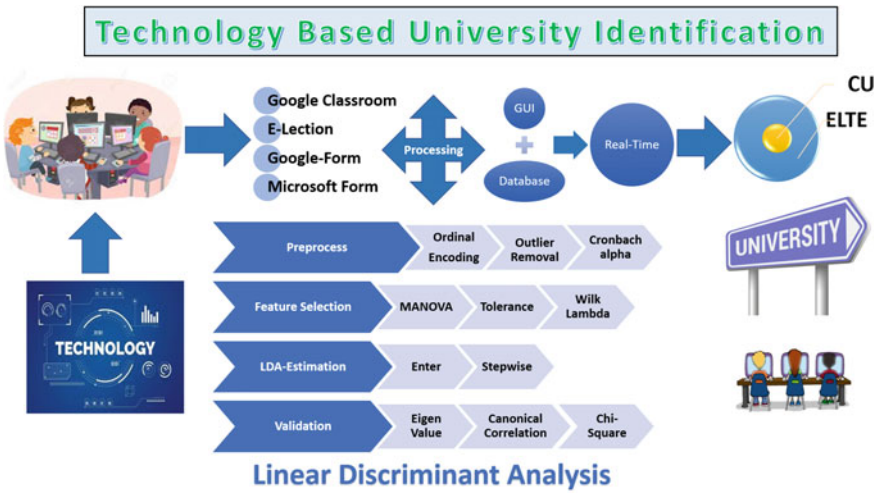


Fig. 1 University identification system

the technology provided them at their institutions. For this, they can opt university web-based system or any other response system like Google and Microsoft Forms, Classroom rooms, and E-lection [10, 11]. Recommended the LDA model for existing student response system fetches student's responses from the database. After performed classification, it will provide the name of the student's university on the GUI interface through the real-time server.

2 Contribution

This paper explored the preliminary novel technology features to identify the university of students. The university's identification accuracy has also been improved and feature reduction applied with the LDA approach. The outcomes of the study might be helpful for educational institutions in the international study environment. Using the present predictive model, the university's real-time response system would predict the opinion of students. Presently, We have a stable web-based real-time web & mobile application called E-lection that has been developed in ASP.NET WebForms in C# language [10, 11]. The real-time functionality has been added with SignalR library [12]. Hence, the present model could be deployed to support university leadership to predict the previous university based on their perceptions concerning the technology provided. Similarly, other demographic features like age, gender, locality, study-level might be identified as well.

3 Methods

3.1 Dataset

The dataset holds 302 samples and 38 technology features. These samples were gathered randomly from the two famous Universities that resided in Indian and Hungary. The percentage of student's participation from the public university Eötvös Loránd University (ELTE) has 48.01% and Chandigarh University (CU) has 51.99%. All students belonged to the computer science program. The dataset's features are depicted in Fig. 2. The hybrid scale of measurement was: development and availability (Yes = 1, No = 2, Don't Know = 3), Attitude and Educational Benefit (Strongly Disagree = 1, Disagree = 2, Undecided = 3, Agree = 4, Strongly Agree = 5), Usability (Never = 1, Rarely = 2, Sometimes = 3, Often = 4, All the time = 5). The reliability of the selected 13 features of the instrument was 0.849, computed with Cronbach's alpha (α) [13].

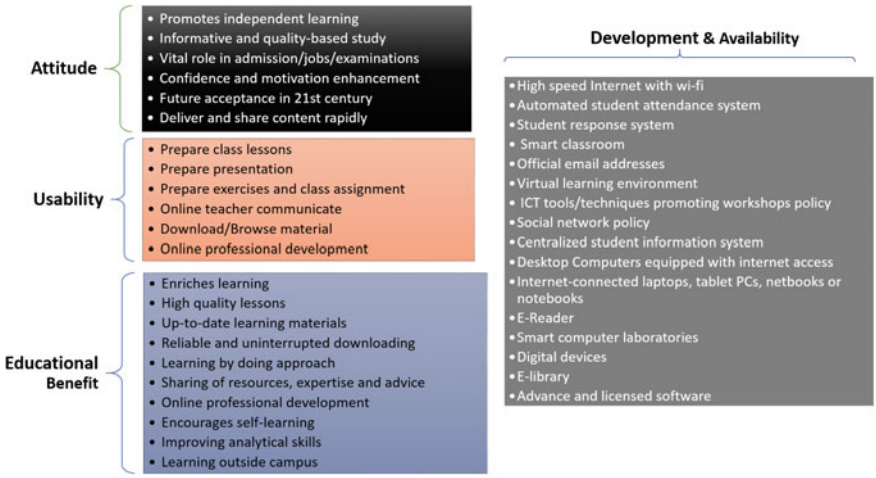


Fig. 2 Technology features (*i*)

4 Preprocess and Feature Selection

To analyze the dataset and execute the experiment, we used IBM SPSS statistics version 25. All the features are appropriately encoded with a natural Ordinal Encoding scheme. The Fisher's LDA has been applied on a dataset in violation of the important assumption (normality) [14]. Also, the Box's M test found that samples have a lack of equal covariance in matrices (homogeneity of covariance in MANOVA). Table 1 shows the insignificant p value that violated the second assumption (Box's M = 428, F (91, 4.5) = $p < 0.05$) and proved that the estimation of discriminant function, which is unequal covariance matrices. Therefore, LDA is robust and it can tolerate this assumption [5, 6].

The multicollinearity of selected features is tested with Tolerance (T). It needs to be maximum to reveal the low multicollinearity among independent features. The T is associated with each independent feature and ranges from 0 to 1. According to [15], there is no hard and fast rule for tolerance thresholds, but better to consider it above 0.4. A strong correlation always exists among independent features if it goes below 0.20 [16]. Equation 1 draws the equation to estimate the Variation Influence Factor (V) and T. The T is 1-R^2, where R^2 depicts the coefficient of determination for regressing the predictors (i). The T is the reciprocal of V of the features, and it is also found $>$ 1.

Table 1 Box's test of equality of covariance matrices

Box's M	F	df	Sig. p
428.05	4.5	91	0.000

$$V = 1 - \frac{1}{1 - R_i^2}, T = \frac{1}{V} \tag{1}$$

During the MANOVA, we used Wilk's Λ statistics that conform the mean equality in predictors based on the target feature [17]. Based on it, most contributed/significant ($p < 0.05$) features are elected to discriminate against the student's university. It probably tests the badness of fitting of the independent features, and it exhibits a way to discriminate records into classes. Equation 2 is the general equation to estimate the Λ by subtracting the proportion of the total variance η from 1. Higher the value of Λ indicates lower discriminatory strength of the function.

$$\Lambda = 1 - \eta^2 \tag{2}$$

Figure 3 graphs the included important features that played a meaningful act in discriminating the target feature university. The Y-axis holds the estimated values of the T and Wilk's lambda (Λ), and the x-axis shows the selected features. We decided the predictors having Wilk's lambda (Λ) values $<$ 0.5, and with high Tolerance (T) (> 0.7). The straight green line shows how all features fit the LDA model for the discrimination purpose. Further, the curved blue line shows a low correlation among the selected features. Therefore, these features are found significant to identify the university.

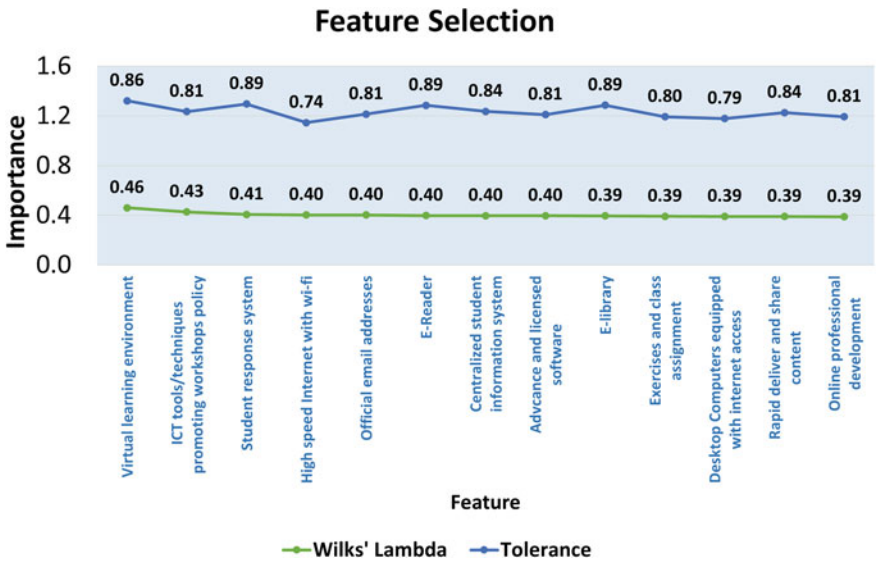


Fig. 3 Feature (i) selection

4.1 Linear Discriminant Analysis

The LDA is a classifier used to classify objects such as individuals, firms, products, etc. This paper applied the two-group LDA approach with ENTER and STEPWISE method. It supports estimating the optimum discriminant Model and assesses overall fit. During simulation, the dependent feature (k) is a university. It is a type of categorical (Non-metric and nominal). To identify the university, 37 independent features (i) are considered (metric or scale). Also, the sample size is 10 : 1 ratio against the feature, and the leave one out (cross-validation) method is used to test these samples. Equation 3 shows the canonical discriminant function of the LDA for the university (k) discrimination.

$$Z_{jk} = a + W_1 X_{1k} + W_2 X_{2k} + + W_i X_{ik} \tag{3}$$

where Z_{jk} is discriminant Z score of discriminant function j for object k, a = Intercept, W_i = Discriminant weight or coefficient for independent feature i, X_{ik} = Independent feature i for object k.

5 Results

This section elaborated on the results of the two experiments with validation approaches. During the first experiment, the full features are passed through the LDA Model using ENTER method, and 86% of cross-validated accuracy is noted. Later, the second experiment used the STEPWISE method of LDA. During the STEPWISE approach, the thresholds of partial F are set to 2.71, and the maximum number of steps is found to be 74.

Figure 4 graphs the improvement in the discrimination accuracy provided with 13 features. It can be seen that training accuracy remains the same, but cross-validation accuracy boosted is by 3%. Hence, we discussed the results of the second experiment having maximum accuracy.

Table 2 shows the LDA model classification results on the original and cross-validated samples. It can be seen that on average, 90.6% of actual grouped cases are correctly classified, and similarly, on average, 89.4% of cross-validated grouped cases are correctly classified. During cross-validation testing, a minor error of 1.2% error is increased. It proved that stabilization in the two types of accuracies. Also, the massive number of instances are predicted accurately of the CU compared to the ELTE (92.4% $>$ 86.2%).

Table 3 exhibits the validation metric of the Canonical discrimination function of the presented LDA model such as Eigenvalue λ, Chi-Square (χ^2), and Wilk's lambda Λ. The percentage of explained variance is 100% and proved with >1 Eigenvalue λ which is the ratio of explained and unexplained variances. Also, the model has a lower Wilk's lambda Λ value of 0.38 that proved the goodness of fitting the features

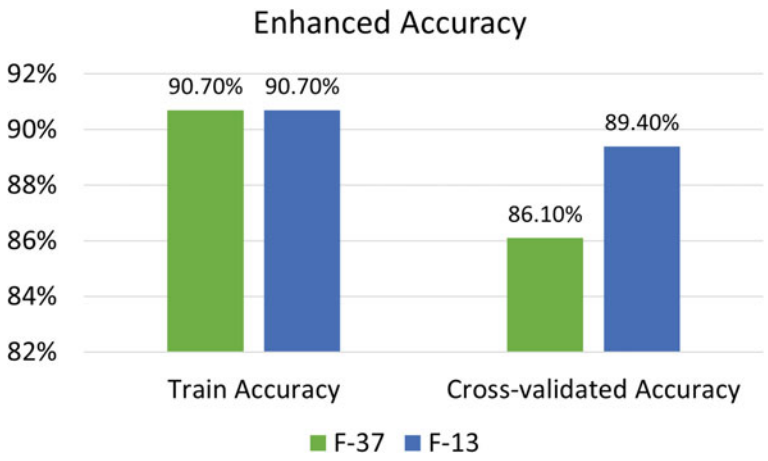


Fig. 4 Feature (i) based discrimination accuracy

Table 2 Fisher's LDA classification results

		Predicted		Accuracy and error		
Count	Name	CU	ELTE	Accuracy (%)	Average accuracy (%)	Error
Original	CU	148	9	94.3%	90.6%	9.4%
	ELTE	19	126	86.9%		
Cross-validated	CU	145	12	92.4%	89.4%	10.6%
	ELTE	20	125	86.2%		

Table 3 Metrics of canonical discriminant functions

λ	% Variance	r_c	Λ	χ^2	Sig. p
1.6	100	0.786	0.38	282.3	0.000

(i) in the model. Additionally, the strong association among categorical feature(k) and predictors is evident with the Chi-Square (χ^2) test (χ^2 (13) = 282.3, $p < 0.5$) in Eq. 4, where O_i is the observed value and E_i is the expected value of the feature under investigation.

$$\chi^2 = \sum_{i=1} \frac{(O_i - E_i)^2}{E_i} \tag{4}$$

Figure 5 visualizes the discriminant scores confirmed the group membership of the target feature k. It can be seen that the LDA score less than 0 (–ve) confirmed the membership of the CU, and the ELTE is recognized with a score above 0 (+ve). The canonical discriminant functions were evaluated at the centroid of CU with –1.22

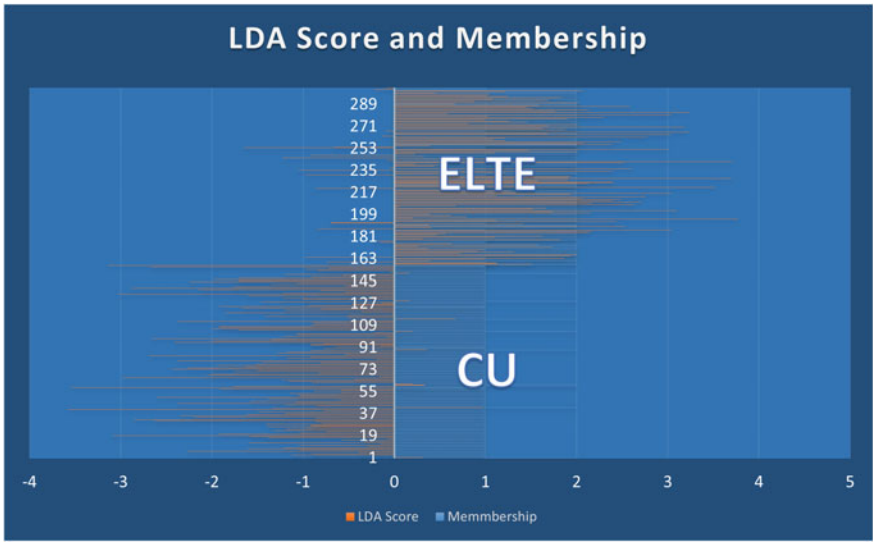


Fig. 5 Discriminant score of dependent feature (k)

and ELTE with 1.32. The Canonical correlation (r_c) of 0.786 depicts the strength of association of the discriminate function score with group membership and computed with the Eq. 5 and it is found highly appropriate for discrimination [5].

$$r_c = \sqrt{\frac{\lambda}{1+\lambda}} \tag{5}$$

$$\begin{aligned} Z_{jk} = -0.4.70 + (-0.513) + 0.450 + (-4.51) + (-0.927) + 0.636 + (-0.477) + 0.586 \\ +0.436 + 0.296 + (-0.435) + 0.202 + 0.230 + (-0.166) + (-0.470) \end{aligned} \tag{6}$$

Equation 6 inserted the unstandardized coefficients of 13 features (i) in the canonical discriminant function to estimate the Z score of dependant feature k. The value of constant α is –4.7. Additionally, Fig. 6 displays the Fisher's discriminant coefficients that promised the membership of dependant feature k.

6 Limitation

The present study is confined to a limited number of samples, population, number of universities, technology parameter, and feature reduction technique (LDA), The presented model is not yet to be implemented. The selected features that predicted the university are also confined.

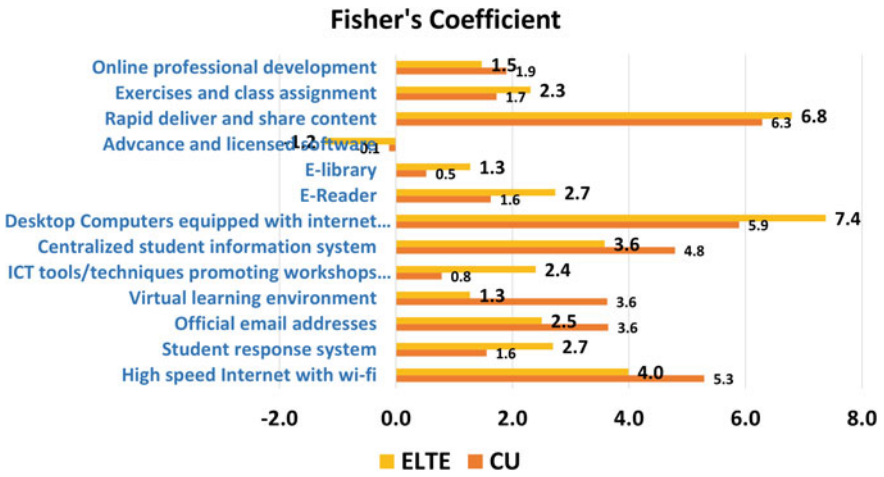


Fig. 6 Fisher's coefficient of feature (i)

7 Conclusion

This paper presented a Fisher's LDA model with a selection of 13 prominent features (i) out of 37 features in the dataset. Based on these features (i), the student's university has been identified. For this, there were two experiments conducted and compared the classification accuracy to discriminate the university groups. The presented model can identify the study university of students with validated accuracy of 89.4%. Further, the feature's fitting in the model suggested with minimum Wilk's lambda Λ value of 0.38. Also, the canonical correlation (r_c) found high (0.786) that proved the strong bonding of 13 predictors with the university group, and the same was also validated with the significant Chi-Square (χ^2) ($p < 0.5$).

The study's limitations are: implementation of the single classifier, abnormal samples, confined feature selection approaches. These shortcomings could be overcome to gather more samples by applying more feature selection methods with several classification algorithms. Moreover, implementing the present model on the university could be helpful to categorize the technology opinion. To make the sample normal, a data transformation can be used. The dataset's feature can also be transformed into a new feature with Principal component analysis [18, 19] and the chi^2 association can also be used [20].

Acknowledgements This work of Chaman Verma and Zoltán Illés was sponsored by the Hungarian Government and Co-financed by the European Social Fund under the project "Talent Management in Autonomous Vehicle Control Technologies (EFOP-3.6.3-VEKOP-16-2017-00001)".

References

1. M. Agaoglu, Using discriminant analysis for multi-class classification: an experimental investigation. Knowl. Inf. Syst. **10**(4), 453–472 (2006)
2. M.N. Hasan, A comparison of logistic regression and linear discriminant analysis in predicting of female students attrition from school in Bangladesh, in *4th International Conference on Electrical Information and Communication Technology (EICT)* (2019), pp. 1–3
3. H. Gull, M. Saqib, S.Z. Iqbal, S. Saeed, Improving learning experience of students by early prediction of student performance using machine learning, in *2020 IEEE International Conference for Innovation in Technology (INOCON)* (2020), pp. 1–4
4. S. Poudyal, M. Nagahi, M. Nagahisarchoghaei, G. Ghanbari, Machine learning techniques for determining students' academic performance: a sustainable development case for engineering education, in *2020 International Conference on Decision Aid Sciences and Application (DASA)* (2020), pp. 920–924
5. M. Agaoglu, Predicting instructor performance using data mining techniques in higher education. IEEE Access **4**, 2379–2387 (2016)
6. R.W. Klecka, *Discriminant Analysis* (Sage Publications, Newbury Park, 1980)
7. S. Poudyal, M.J. Mohammadi, J.E. Ball, Data mining approach for determining student attention pattern, in *2020 IEEE Frontiers in Education Conference (FIE)* (2020), pp. 1–8
8. C. Verma, V. Stoffová, Z. Illés, S. Tanwar, N. Kumar, Machine learning-based student's native place identification for real-time. IEEE Access **8**, 130840–130854 (2020)
9. C. Verma, Z. Illés, V. Stoffová, V. Bakonyi, Artificial neural network for educational standard of university student for real-time, in *Future of Information and Communication Conference (FICC)*, vol. 1364 (2021), pp. 323–334
10. E-Lection, https://election.inf.elte.hu/ (Accessed 03 April 2021)
11. V. Bakonyi, Z. Illés, Real-time classroom. Cent. Eur. J. New Technol. Res., 13–20 (2019)
12. V. Bakonyi, Z. Illés, Using real-time application in education. Int. J. Adv. Electron. Comput. Sci., 33–39 (2019)
13. L.J. Cronbach, Coefficient alpha and the internal structure of tests, in *Psychometrika*, vol. 16, No. 3 (Springer Science and Business Media, 1951), pp. 297–334
14. R. Fisher, The use of multiple measurements in taxonomic problems. Ann. Eugen. **7**, 179–188 (1936)
15. P. Allison, *Multiple Regression: A Primer* (Pine Forge Press, 1999)
16. D. Weisburd, C. Britt, *Statistics in Criminal Justice* (Springer Science & Business Media, 2013)
17. B.S. Everitt, G. Dunn, *Applied Multivariate Data Analysis* (Edward Arnold, London, 1991), pp. 219–220
18. C. Verma, Z. Illés, V. Stoffová, P.K. Singh, Predicting attitude of Indian student's towards ICT and mobile technology for real-time: preliminary results. IEEE Access **8**, 178022–178033 (2020)
19. K. Sachdev, M.K. Gupta, Predicting drug target interactions using dimensionality reduction with ensemble learning. Proceedings of ICRIC **597**(2020), 79–89 (2019)
20. D. Kumar, C. Verma, P.K. Singh, M.S. Raboaca, R.A. Felseghi, K.Z. Ghafoor, Computational statistics and machine learning techniques for effective decision making on student's employment for real-time. Mathematics **9**(11), 1–29 (2021)

Comparison of Multi-Criteria Decision-Making Techniques for Cloud Services Selection

Neha Thakur, Avtar Singh, and A. L. Sangal

Abstract Cloud Computing is the latest model that supports the idea of computing as a utility. Fundamentally this is not the latest idea, but for users (industries and researchers), it's now a viable truth to pay for the cloud services according to their usage, with no upfront cost and with effectually unlimited adaptability. Although the future of cloud services looks very promising, there are still some challenging problems out of which the selection of cloud services is one of them. This problem has fascinated huge attention both from industry experts and academicians. To help the users to make wise decisions, researchers have developed various techniques by conducting different studies. Considering various criteria during the cloud services selection process, Multi-Criteria Decision-Making (MCDM) techniques are widely adopted to assist the decision-makers to make acute choices. In this paper, a comparison of MCDM-based approaches for cloud service selection has been performed. This paper has identified the challenges and highlighted the limitations of MCDM-based cloud services selection techniques after contrasting and compiling the reviewed approaches. This study is anticipated to bring profits to both academia and industry professionals.

Keywords Cloud services · MCDM · Cloud computing · MADM · MODM · Cloud services selection

N. Thakur (✉) · A. Singh · A. L. Sangal
Dr. B. R. Ambedkar National Institute of Technology, Jalandhar, Punjab, India
e-mail: nehat.cs.18@nitj.ac.in

A. Singh
e-mail: avtars@nitj.ac.in

A. L. Sangal
e-mail: sangalal@nitj.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_51

1 Introduction

Cloud computing is "new-generation information storage and computing systems with virtualization as the central, facilitate technology to communicate and manage distributed computers that are vigorously provisioned on-demand as a custom-made catalog to meet a specific service-level agreement" [1, 2]. In simple words, Cloud computing is a new technology for distributing computing resources over the web as a service to cloud users similar to other utilities (Dish TV, gas, water, and electricity). These services are distributed based on three cloud service models. These are 1. Software as a Service (SaaS) 2. Infrastructure as a Service (IaaS) 3. Platform as a Service (PaaS) [3]. As cloud offers various business benefits, many organizations start building their applications on the cloud platform to make their organization brisk by using scalable and flexible services of cloud. But then again shifting applications on the cloud is not a piece of cake. There are many challenges to grasp the complete potential of cloud computing. These challenges are associated with the fact that definite requisites and features are essential to be met before moving different applications to the cloud [4]. Also, with the proliferation of a hefty number of services over the cloud, it has to turn out to be even more problematic for cloud consumers to choose which service provider can meet their needs. Each cloud service provider offers a similar type of service with a different set of features at dissimilar prices and performance levels. While a cloud service provider may be cheaper to rent virtual machines, but offering gigabytes of storage space from them can be affluent [5]. Consequently, with the miscellany of cloud services, an important challenge for cloud consumers is to search for "ideal" services that meet their needs. There is often arbitration between the diverse quantitative and qualitative requirements satisfied by diverse cloud service providers [6]. This makes it hard to calculate service levels of diverse cloud service providers in an objective way such that mandatory attributes such as quality, availability, reliability, and privacy of an application can be guaranteed in the Cloud. Henceforth, finding the services over the cloud is not sufficient, it is similarly imperative to estimate which service is the utmost appropriate cloud service.

Besides, the process of selection of cloud service is well thought-out by many authors as an MCDM problem. Generally, the MCDM technique is utilized where several choices exist, and a conclusion must be made to select one of them concerning certain criteria. But instead of considering many of these schemes, there is a lack of an extensive survey paper to review its various facets. However, the author of [1] investigated MCDM-based strategies, but they only considered schemes published up to 2014. Furthermore, it does not distinguish clearly between any of the MCDM decision methods. This prompted us to conduct a study that focuses exclusively on MCDM-based schemes and compares each decision-making method's advantages and disadvantages. This article focuses on all the MCDM-based techniques which are applied to select the best cloud services. It first provides a literature review and the basic concept about the various MCDM methods used for cloud services selection and then compares these techniques and features their substantial contribution,

characteristics, and drawbacks. Finally, in the conclusion, future challenges in the context of MCDM-based cloud services selection are stated.

2 Literature Review

Multiple criteria decision-making is a field of operational research that deals with discovering ideal solutions in difficult scenarios, together with various inconsistent objectives and attributes [2]. This technology is gaining widespread attention in the field of the selection of cloud services because it allows decision-makers to make decisions while simultaneously considering all criteria and objectives. In simple, MCDM is a process that allows conclusions to be drawn on the occurrence of several conflicting criteria. It explicitly evaluates several conflicting criteria in decision-making. The difficulties of MCDM can be broadly divided into two classes:

- MADM: MADM stands for Multi-Attribute Decision-Making. This includes the choice of the best option from detailed substitution defined in terms of several features.
- MODM: MODM stands for Multi-Objective Decision-Making. This encompasses the scheme of substitutions that enhance many of the decision-making goals.

In the literature, MADM techniques are preferred by more research scholars over MODM techniques. MADM approaches are related to problems that have limited and finite substitutes. In divergence, MODM methods are for problems that have no pre-defined alternatives. Multi-Criteria Decision-Making Method is alienated into different groups according to the same characteristics. Table 1 shows the outline of the MCDM method. The method for classifying the items that best meet the aim, requirements, and constraints of the decision-maker is called Decision-making [3]. MCDM problems are decision problems with finite criteria [1].

Decision-maker prerequisite to make numerous comparisons of services grounded on multiple criteria at the time of services selection process. Therefore, services selection is classified as an MCDM problem [15]. Widely used MCDM techniques are Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) [16], Analytical Hierarchy Process (AHP) [7], Analytical Network Process (ANP) [8], Simple Additive Weighting (SAW) [17], Vls' ekriterijumsko KOmpromisno Rangiranje (VIKOR) [18], ELimination Et Choice Translating REality (ELECTRE) [3], Preference Ranking Organization METHod for Enrichment Evaluation (PROMETHEE) [4], Gray Relational Analysis (GRA) [19], Data Envelopment Analysis (DEA) [13], Fuzzy [20]. Table 2 shows the comparison of MCDM techniques along with their advantages and challenges.

In [16], author proposes a novel framework that allows cloud consumers to compare services built on Quality of Service (QoS) criteria. The framework utilizes the hybrid MCDM approach. The Best Worst Method (BWM) is utilized to rank and order the QoS attributes, and TOPSIS is employed to rank the cloud services. In [15], author proposed a framework "Methodology for Optimal Services Selection

Table 1 Multi-criteria decision-making model description

Groups	Definition	MCDM methods	References
Scoring method	The simplest MCDM methods are called Scoring Methods. These methods involve making use of simple arithmetic operators for evaluating substitutes	SAW COPRAS	[4] [5] [6]
Distance-based method	The general idea of the distance-based method is finding the distance between each alternative and a specific point. There are two divergent ideologies within this group. To find alternatives that satisfy a set of goals is the aim of Goal Programming (GP). Contrary to that, to obtain a substitute close to the imaginary optimal option is the goal of the Compromise Programming (CP) method aims	Goal programming Compromise programming TOPSIS VIKOR DEA	[7] [4] [8, 9] [10]
Outranking method	Generating a preference relation on a set of substitutes that specifies the degree of dominance among them. The outranking method deals with imprecise and partial information and their application outcomes in the limited preference ranking of substitutes, in place of a cardinal degree of their preference relation	PROMETHEE ELECTRE	[11] [12] [13]
Pairwise method	These methods are useful to get the weight of distinct attributes and equate options regarding a subjective attribute. The drawback of pairwise methods is that they depend only on the knowledge of decision-makers. Also, it is conceivable that diverse decision-makers have dissimilar tactics to the same problem	AHP ANP	[6] [8] [14]

(continued)

Table 1 (continued)

Groups	Definition	MCDM methods	References
Utility/Value	These methods define terminologies that determine the degree of satisfaction of the attributes. According to the method (MAUT or MAVT), these functions convert the ratings that state the nature of the choices concerning the attributes into their degree of satisfaction. The appearance of the function can model dissimilar figures to re-count the scores and the degree of satisfaction	Multi-Attribute Value Theory (MAVT), Multi-Attribute Utility Theory (MAUT)	[14] [7]

(MOSS)". In this paper, the author uses the BWM to compute the weights of QoS and Quality of Experience (QoE) attributes. Then, use an MCDM-based approach entailing well-known existing MCDM methods to get QoS and QoE based levels in the ranking phase. Author of [19], presents a framework for the optimal selection of cloud services. In this paper, to calculate the weights of the attributes and rank the services, GRA method is used. In [21], author uses AHP to alter consumer's qualitative and partial-qualitative personalized preferences into quantifiable numeric weights. A case study in the medical cloud platform is used to authenticate the viability and competence of the model. In [14], author introduced a methodology for defining the suitable cloud services by mixing the AHP weighting method with TOPSIS method.

AHP technique is used to estimate the criteria weights and TOPSIS is utilized to attain the final ranking. The author of [22], uses Fuzzy – AHP procedure to proficiently tackle quantifiable and non-quantifiable decision features elaborated in the selection of cloud services. This paper publicized that fuzzy AHP is proficient in competently handling the fuzziness of data involved in multi-criteria decision-making. In [23], a model grounded on merging weights and GRA is proposed. Firstly, direct trust, reputation trust, and recommendation trust create a complete trust resulting in more precise overall trust. Secondly, rough set theory and AHP methods are used for direct trust. In [15], decision-maker is permitted to choose ideal cloud services by considering together QoS and QoE criteria. Then the MCDM techniques were used to rank the services according to the user's requirement. The best worst method is used to obtain the consistency ratio. Recently, [1] conducted a comprehensive literature review on the state-of-the-art techniques for the selection of cloud services and concluded that the maximum of the existing research either lacks proof or is inadequate.

Table 2 Comparison of MCDM-based cloud services selection techniques

Name	References	Objective	Approach	Advantage	Challenges
Analytical Hierarchy Process (AHP)	[7]	Assist decision-maker to discover the optimal solution that satisfied users' requirements Also, aid to understand the goal or problem	A beneficial technique for the hierarchical relationship of attributes. Done pairwise comparison of attributes organizes in a hierarchical relationship	An elastic, spontaneous demand to the decision-makers and capability to check discrepancies	The total pairwise comparison to being done may be converted to very high (n(n-1)/2), and thus developed to a lengthy task
Analytical Network Process (ANP)	[8]	A general depiction of interrelationships or inter-dependence among decision levels and criteria	Model static and irreversible relationships with trust and response other than of hierarchy	Divide the problem into smaller parts, which are easy to understand and solve. A real-world representation of the problem by making use of clusters	Identifying attributes requires extensive brainstorming sessions ANP entails more calculations in contrast to the AHP process
Technique for Order of Preferences by Similarity to Ideal Solution (TOPSIS)	[16]	TOPSIS is a multi-criteria decision analysis method that comes under a distance-based group. It selects the alternatives and ranks them	The optimal result is the one that has the smallest distance from the best solution and utmost from the worst solution	The benefits of this method are affluence of application, ubiquity, contemplation of distances to an ideal solution	Maximum subjectivity in the uncertainty conditions for finding the weights of the criteria
Simple Additive Weighting (SAW)	[17]	SAW method is built on the weighted average calculation of the criteria. In this method, each of the criteria is given a certain weight and each substitute is determined concerning the corresponding attribute	Calculating numerous alternatives in terms of several deciding factors	The calculation algorithm of this method is not complicated Normalized values of the evaluation help visually calculate the differences between the alternatives	Only, when all the data are represented in the same unit SAW is applicable. The SAW method's estimates yielded do not always reflect the real status

(continued)

Table 2 (continued)

Name	References	Objective	Approach	Advantage	Challenges
Elimination and Choice Expressing Reality (ELECTRE)	[13]	Based on a concord and cacophony index that is evaluated with mined data from a conclusion table ranking of alternatives is done	Check only if one alternative is best or worse compared to the other available alternatives	Even if there is not a clear preference for one of those substitutes. Still, the comparison of the alternatives can be achieved More Reliable	It is a somewhat tough decision-making method and needs a lot of key data
Preference Ranking Organization Method of Enrichment Evaluation (PROMETHEE)	[24]	It comes under the outranking group. PROMETHEE method does not eradicate any substitute in pairwise evaluation according to criteria and decision-maker preference as an alternative it sets the alternatives in the ranking order	Instead of indicating a "right" decision, this method helps decision-makers find an alternative that is best suited to goals and their understanding of the problem	PROMETHEE can concurrently deal with qualitative and quantitative criteria The benefits of this method are easiness, lucidity, and stability. PROMETHEE requires fewer inputs than ELECTRE	Rank Reversal problem when a new alternative is presented. It doesn't provide a chance to structure a decision problem
Decision-Making Trial and Evaluation Laboratory (DEMATEL)	[25]	For the recognition of cause-effect chain components of a complex system, this is an efficient method Through a visual structural model, it estimates the inter-reliant relationships among criteria	The power of influence representing the numerical contextual relations among the elements	It efficiently examines the common effects (both direct and indirect effects) amid dissimilar aspects and recognizes the complex cause and effect relationships in the decision-making problem	Based on interdependent relationships among alternatives it determines the ranking Other independent criteria are not incorporated into the decision-making problem The relative weights of professionals are not included

(continued)

Table 2 (continued)

Name	References	Objective	Approach	Advantage	Challenges
Grey Relation Analysis (GRA)	[19]	The general belief of the GRA technique is that the selected substitute should have the "major degree of grey relation" from the positive ideal solution and vice versa	Based on grey system theory	The outcomes are built on the original data. The calculations are modest to comprehend. To make decisions in the business environment, GRA helps management	Cannot give an optimal solution
VIKOR	[18]	Conciliation ranking order and the compromise results attained with the starting weights are obtained by the VIKOR method. In the presence of contradictory conditions, this technique concentrates on selection and ranking from a set of substitutes	Catalogs obtained from a degree of "closeness" to the "optimal result" utilized to present compromise ranking	The benefit of this method is that it offers a procedure for ranking positive criteria and negative criteria, as soon as it is utilized and studied in decision support	The greatest disadvantage of the method is searching for the concession ranking order, i.e. compromise between uncertain and predictable solutions
Fuzzy	[13]	Calculation of the importance of weights of attributes and fitness of substitutes can be estimated using the linguistic values represented by fuzzy numbers	Fuzzy terms are defined by linguistic variables that are then mapped to arithmetic variables	Inadequate information and the progress of accessible knowledge is taken into account by fuzzy logic It permits vague input	Sometimes, fuzzy systems are difficult to develop. In many cases, they need abundant simulations before being used in the real world

(continued)

Table 2 (continued)

Name	References	Objective	Approach	Advantage	Challenges
Data Envelopment Analysis (DEA)	[13]	DEA presents a model for weight calculation and exploiting the competence of the decision-making units. The prime concentration of the DEA model is to relate decision-making units (substitutes) in terms of their competence in altering inputs into outputs	The relative competence of decision-making units is estimated using the linear programming technique	Can handle multiple inputs and outputs. The relationship between input–output is not necessary. Evaluations are straight in distinction to peers. Dissimilar units are allocated to Inputs and outputs	Major problems occur because of measurement error Absolute competence cannot be determined Large problems can be challenging Arithmetic tests cannot be applied

3 Challenges in Cloud Services Selection

Services which are delivered on-demand via the internet and can be accessed or managed from the provider's server instead of the organization's on-premise server are called cloud services [26]. From ordinary consumer's perception, cloud services are almost similar to web services. For example, software as a service is treated as web services. Also, IaaS and PaaS services can be accessed through a web-services interface. Additionally, few web services searching techniques are applied to discover SaaS services. However, it is clear from the definition of cloud services that we can't site both web services and cloud services in the same category. When we extend web services selection techniques to the selection of cloud services, new challenges and problems concurred. These challenges are:

3.1 More Types of Services

In the cloud infrastructure, everything can be considered as a service, e.g., IaaS, PaaS, and SaaS. Depending on the existing cloud infrastructure, any individual user or organization can create its cloud services. Hence, there is a pool of numerous cloud services with dissimilar cost and functional or non-functional performance attributes. Cloud consumers' necessities are fairly diverse for distinct types of cloud services.

3.2 More Performance Criteria

Only a few significant service performance parameters are considered in traditional web service selection, e.g., service throughput, availability, service response time, etc. Contrary to that, cloud-delivered services with composite functional attributes or criteria, e.g., IaaS or PaaS cloud services. Therefore, numerous inimitable cloud performance characteristics, e.g., service scalability, service elasticity, service reliability, service payment policies, service privacy, etc., should be considered while performing the selection of cloud service. Additionally, persistent storage, service networking, etc., specific performance parameters are needed to be considered according to different requirements. Therefore, recognizing critical performance parameters is the primus for the selection of cloud services.

3.3 More Dynamic Performance

Cloud services are frequently used simultaneously by an enormous number of users. Scalability and elasticity are cloud computing features that ensure each user can practice consistent service performance at the same time. Cloud services have the competence to allocate the essential computing resources to new cloud users close to real time, even at the peak-load time. Although, in a real-life situation, cloud services may have wobbly performance due to incessantly changing the number of users. Hence, corporate customers are the primary purchasers of cloud infrastructure and cloud platform services. In addition, unlike web services, cloud application services target both business and individual customers.

3.4 Target Different User Groups

Unlike web services, cloud services don't aim for a single user group. Different cloud services target different user groups. IaaS service providers target IT, administrators, PaaS providers targets software developers, and SaaS targets end-users. Hence, cloud infrastructure services and cloud platform services are mostly acquired by business users, whereas cloud application services contrasting to web services target both enterprise and individual users.

4 Identified Issues

Under this section, open issues on the recent approaches for cloud services selection based on MCDM methods, are identified from the literature and discussed:

- In the literature, most of the MCDM -based approaches used for the selection of cloud services depend on a fixed value of QoS attributes. Though, in the future variable values of QoS criteria should be considered by taking into account the dynamic nature of cloud computing and fluctuating service provider performance.
- Another significant parameter that must be considered in the forthcoming researches is "security". More security-related parameters need to be examined to select secure cloud services that preferably support Privacy, Integrity, and authentication.
- In the literature, while studying MCDM-based services selection approaches very few QoS criteria are considered. However, many other important QoS criteria need to be considered while making the selection of cloud services.
- In the utmost of the literature, to calculate the relative significance of the parameters weight is subjectively assigned by users. Which either be unfair or based on a judgmental scale and thus is not well-organized for decision-making. A method

for evaluating subjective opinion is needed to upsurge the subjective services selection. Which also record user's preference automatically and proficiently.

- Multi-tenancy technology forms the basis of cloud resource distribution. When choosing cloud services, the ability of cloud service providers to maintain performance for multi-tenants and fulfill their QoS constraints is a significant decision-making component for service users. In this circumstance, the service evaluation characteristics and their estimation values are disparate from those functional to the performance of a single-tenant job in one period. From the literature, it is clear that most of the research focuses on single-tenant services selection. Services selection for Multi-tenants is still an open problem, which requires further attention.

5 Conclusion

Cloud computing offers many benefits to cloud users such as elasticity, availability, pay-as-you-use policy, scalability, anytime and anywhere access, etc. However, as every coin has two aspects, cloud computing too has its own portion of issues and challenges. Amid all these issues and challenges, that cloud computing confronts "Selection of cloud services" is supreme. Realizing the importance of cloud services selection, the contemporary literature on MCDM-based cloud services selection approaches is extensively analyzed. The MCDM-based approaches are grouped and compared based on the cloud services selection process. After conducting a thorough analytic discussion of the literature, we pin down the open problems or issues in MCDM-based approaches for cloud services selection and put a foundation for future research.

References

1. L. Sun, H. Dong, F.K. Hussain, O.K. Hussain, E. Chang, Cloud service selection: state-of-the-art and future research directions. J. Netw. Comput. Appl. **45**, 134–150 (2014). https://doi.org/10.1016/j.jnca.2014.07.019
2. A. Kumar et al., A review of multi criteria decision making (MCDM) towards sustainable renewable energy development. Renew. Sustain. Energy Rev. **69**, October 2016, 596–609 (2017). https://doi.org/10.1016/j.rser.2016.11.191
3. E. Triantaphyllou, *MCDM methods: a comparative study* (Springer, Boston, MA, 2000), pp. 5–21
4. Z.U. Rehman, O.K. Hussain, F.K. Hussain, Iaas cloud selection using MCDM methods. in *Proceedings - 9th IEEE International Conference on E-Business Engineering, ICEBE 2012* (2012), pp. 246–251. https://doi.org/10.1109/ICEBE.2012.47
5. G. Büyüközkan, F. Göçer, O. Feyzioğlu, Cloud computing technology selection based on interval valued intuitionistic fuzzy COPRAS. Adv. Fuzzy Logic Technol. pp. 318–329 (2017). https://doi.org/10.1007/978-3-319-66830-7

6. M. Hosseinzadeh, H.K. Hama, M.Y. Ghafour, M. Masdari, O.H. Ahmed, H. Khezri, Service selection using multi-criteria decision making: a comprehensive overview. J. Netw. Syst. Manage. **28**(4), 1639–1693 (2020). https://doi.org/10.1007/s10922-020-09553-w
7. A.E. Youssef, An integrated MCDM approach for cloud service selection based on TOPSIS and BWM. IEEE Access **8**, 71851–71865 (2020). https://doi.org/10.1109/ACCESS.2020.2987111
8. A. Jaiswal, R.B. Mishra, Cloud service selection using TOPSIS and fuzzy TOPSIS with AHP and ANP. in *Proceedings of the 2017 International Conference on Machine Learning and Soft Computing 2017* (2017), pp. 136–142. https://doi.org/10.1145/3036290.3036312
9. R.K. Tiwari, R. Kumar, *G-TOPSIS: A Cloud Service Selection Framework using Gaussian TOPSIS for Rank Reversal Problem*, vol. 77, 1st edn. (Springer US, 2021)
10. C. Jatoth, G.R. Gangadharan, U. Fiore, Evaluating the efficiency of cloud services using modified data envelopment analysis and modified super-efficiency data envelopment analysis. Soft. Comput. **21**(23), 7221–7234 (2017). https://doi.org/10.1007/s00500-016-2267-y
11. R.K. Gavade, Multi-criteria decision making: an overview of different selection problems and methods. Int. J. Comput. Sci. Inf. Technol. **5**(4), 5643–5646 (2014)
12. H. Ma, H. Zhu, Z. Hu, K. Li, W. Tang, Time-aware trustworthiness ranking prediction for cloud services using interval neutrosophic set and ELECTRE. Knowl.-Based Syst. **138**, 27–45 (2017). https://doi.org/10.1016/j.knosys.2017.09.027
13. M. Liu, Y. Shao, C. Yu, J. Yu, A heterogeneous QoS-based cloud service selection approach using entropy weight and GRA-ELECTRE III. Math. Probl. Eng. (2020). https://doi.org/10.1155/2020/1536872
14. R.R. Kumar, C. Kumar, A multi criteria decision making method for cloud service selection and ranking. Int. J. Ambient Comput. Intell. **9**(3), 1–14 (2018). https://doi.org/10.4018/IJACI.2018070101
15. A. Hussain, J. Chun, M. Khan, A novel customer-centric methodology for optimal service selection (MOSS) in a cloud environment. Futur. Gener. Comput. Syst. **105**, 562–580 (2020). https://doi.org/10.1016/j.future.2019.12.024
16. R.R. Kumar, B. Kumari, C. Kumar, CCS-OSSR: a framework based on Hybrid MCDM for optimal service selection and ranking of cloud computing services. Clust. Comput. 1–17 (2020). https://doi.org/10.1007/s10586-020-03166-3
17. M. Eisa, M. Younas, K. Basu, I. Awan, Modelling and simulation of QoS-aware service selection in cloud computing. Simul. Modell. Pract. Theory **103**, 102108 (2020). https://doi.org/10.1016/j.simpat.2020.102108
18. J. Rezaei, Best-worst multi-criteria decision-making method. Int. J. Manage. Sci. **53**, 49–57 (2015). https://doi.org/10.1016/j.omega.2014.11.009
19. G. Obulaporam, N. Somu, G.R. ManiIyer Ramani, A.K. Boopathy, S.S. Vathula Sankaran, *GCRITICPA: A Critic and grey relational analysis based service ranking approach for cloud service selection* (Springer Singapore, 2018)
20. A. Singh, K. Dutta, A genetic algorithm based task scheduling for cloud computing with fuzzy logic. IEEK Trans. Smart Process. Comput. **2**(6), 367–372 (2013)
21. M. Sun, T. Zang, X. Xu, R. Wang, Consumer-centered cloud services selection using AHP. In *Proceedings of International Conference on Service Science, ICSS* (2013), pp. 1–6. https://doi.org/10.1109/ICSS.2013.26
22. S. Le, H. Dong, F.K. Hussain, O.K. Hussain, J. Ma, Y. Zhang, A hybrid fuzzy framework for cloud service selection. in *Proceedings - 2014 IEEE International Conference on Web Services, ICWS 2014* (2014), pp. 313–320. https://doi.org/10.1109/ICWS.2014.53
23. Y. Wang, J. Wen, X. Wang, B. Tao, W. Zhou, A cloud service trust evaluation model based on combining weights and gray correlation analysis. Secur. Commun. Netw. (2019). https://doi.org/10.1155/2019/2437062
24. O. Gireesha, N. Somu, M.R.G. Raman, M.S. Reddy, K. Kirthivasan, V.S. Sriram, *WNN-EDAS: A Wavelet Neural Network Based Multi-criteria Decision-Making Approach for Cloud Service Selection* (Springer Singapore, 2019)

25. V. Ghafori, R. Manouchehri Sarhadi, Best cloud provider selection using integrated ANP-DEMATEL and prioritizing SMI attributes. Int. J. Comput. Appl. **71**(16), 18–25 (2013). https://doi.org/10.5120/12441-9133
26. K. Stanoevska-Slabeva, T. Wozniak, Cloud basics–an introduction to cloud computing. Grid Cloud Comput. A Business Perspect. Technol. Appl. (2010). https://doi.org/10.1007/978-3-642-05193-7

OULAD Learners' Withdrawal Prediction Framework

Moohanad Jawthari and Veronika Stoffa

Abstract Nowadays, web-based courses are popular as students can learn at their convenience and according to their free time. One advantage is the availability of study materials that can be used in a blended learning program. However, it suffers from a high withdrawal problem. This article contributes to the research by proposing a withdrawal prediction framework based on the Open University, a large distance-learning institution. The study contribution is a dropout prediction framework. The prediction process includes handling missing data using the MissForest algorithm, tackling imbalanced data issues using SOMTEFUNA, reducing dimensions using PCA, and training different classifiers. The experiments show Decision Tree classifier as the best predictive model with an F1-score of 0.99. The proposed framework outperforms other researches by 12% when compared to the previous research work.

Keywords OULAD · Withdrawal prediction · Imbalanced data

1 Introduction

Internet-based learning has become familiar recently in education. It can be of many forms, from massive open online courses (MOOCs) to virtual learning environment (VLE) and learning management system (LMS). MOOCs is a preferred option for students as they can study anytime and anywhere. VLE is also adopted by top-ranked academic institutions to offer open, and sometimes, free education. The common platforms are Coursera, Edx, and Harvard. Although VLE is striving to achieve high-quality education, it suffers from learners' dropout problem. According to [1], the completion rate is frequently less than 7%. Also, Other studies like [2] reveal the dropout rate in Coursera was between 91 and 93%, and [3] indicates the dropout

M. Jawthari (✉) · V. Stoffa
Eötvös Loránd University, Budapest 1053, Hungary
e-mail: mjawthari@inf.elte.hu

V. Stoffa
Trnava University in Trnava, Hornopotočná 23, 918 43 Trnava, Slovakia

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_52

rate in Open University (OU) of UK reached 78%. Such rates are higher compared to traditional learning rates [1].

Such statistics cast considerable doubt on the VLEs trustworthiness and challenge the effectiveness of this learning technology. As a result, considerable effort is expended in developing advanced approaches to increase the learner's dedication to the course. Several solutions introduce easy, sometimes interactive, tools to monitor the progress of learners. To provide interesting statistics, those tools usually use log files. However, they are not reliable in predicting learners' withdrawal from a course. On other hand, data mining (DM) techniques are better alternatives used in many disciplines to uncover hidden patterns in data aspects.

In [4], publicly available datasets in the field of education were reviewed. Those datasets included many learners' features such as demographic data, forum/discussion data, logs data, etc. However, there is no complete dataset. A dataset would contain one information side of learner's features and ignore the other sides due to privacy issues. For instance, researchers have investigated in the discussion forums for analyzing the cognitive process in Coursera platform, and others have dealt with learners' logs data in videos to predict their future behaviors [5]; it is not open-access for scientists because of privacy issues. Luckily, OULAD [6], includes demographic and behavioral data. Besides, it is anonymized, free, and open.

This paper employs DM methods to tackle the learners' dropout problem in VLE. Hence, a framework is proposed to preprocess and prepare the final dataset. Then encoding of categorical features and missing value problems are tackled. We cope with imbalanced data issues. To remove redundant features and reduce training time, feature selection or reduction is tackled. Ultimately, we consider machine learning (ML) methods to obtain an efficient predictive model.

2 Dataset and Related Work

2.1 Dataset

In 2017, Open university learning analytics (OULA) dataset was developed to support learning analytics and educational data mining (EDM) research fields. The database is freely available and comprises seven CSV files as shown in Fig. 1. What makes the dataset distinctive is the truth that it includes demographic data, engagement data from virtual learning environment, and assessments scores. The Open University (OU) is one of the largest distance education institutes worldwide with 170,000 students who were registered in different programs. In OU, VLE is mainly used to deliver teaching materials and other content, and students' interactions with the educational materials are logged and stored in a data warehouse. OULA dataset details about 22 courses that were delivered over the period of 2013–2014. The courses belong to two subjects: "Social science" and "Science, Technology, Engineering, and Mathematics", and are called modules. Modules can be offered many

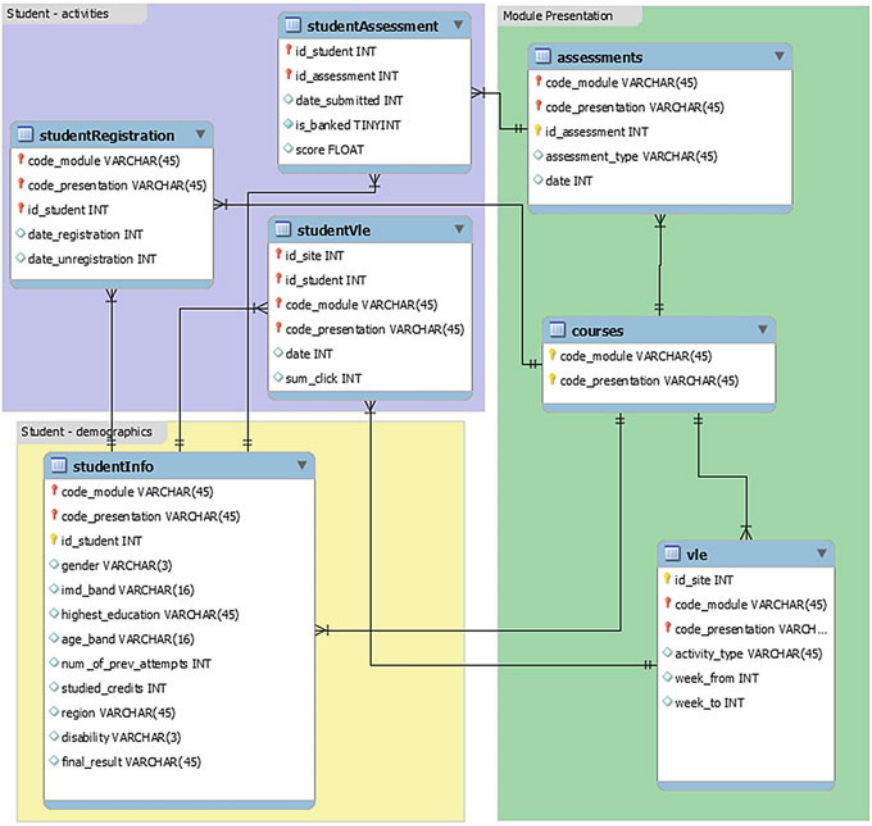


Fig. 1 Database schema [6]

times during the year. To distinguish among different presentations of a module, each presentation is named by the year and the starting month. For example, "2013 J" indicates that the presentation started in October 2013. Figure 1 shows the dataset's structure. The final dataset in this study is obtained by joining all seven tables.

The course table contains all available modules and their presentations. The registration table includes learner record timestamps and course enrollment dates of a course. The assessments table contains the assessment type and the submission day of that assessment for each module presentation. Assessments are classified into three types: tutor marked assessment, computer marked assessment, and the final exam. The student assessment table refers to the assessment results for each student.

The students' interaction data concerning the different learning activities is stored in the StudentVLE and VLE tables. The VLE interaction data presents the clickstream within the VLE, aggregated as daily summaries of learner clicks made while studying the course. Each course activity is identified with a label (e.g., forum, url, wiki, and homepage). The following table details the fields of tables.

2.2 Dropout Prediction and OULAD Literature Review

Substantial research has been conducted to investigate the reason for students' withdrawals; researchers have used different ML and DM techniques to predict whether the learner would dropout or not using OULAD. It is a difficult task due to the used features; each feature can be used for a slightly different problem design. Some researchers dealt with learners' dropouts as a binary classification problem and utilized one or more classification algorithms [7, 7, 7]. For example, Alshabandar et. al. applied the probabilistic model and Gaussian Mixture to predict learners' retention based on assignments deadlines and the VLE interactions [10]. They used a variety of VLE types for each student in some time interval and combined them into a single value. Also, they used features such as dynamic behavioral features, demographic features, and different assignments feature. Other researchers transformed the learner's clickstreams data of all courses to 45 time series datasets and considered the disengagement prediction problem as a time series classification problem. Their goal was to reveal which period contributed to the learning progress of learners.

Furthermore, other researchers considered some of the dataset features in their studies. The authors of [11] investigated the dynamic impact of six demographic features on academic outcomes in different online learning disciplines. Nevertheless, they ignored the effect of variables tracked within the VLE. Another study investigated only the learning activities, namely: "forum", "OUtcontent", and "resources" to analyze the learners' behaviors [12]. At the same contest, only "OUcontent", "forumng", "subpage", and "homepage" were used by [8] to predict low-engagement learners in OU dataset. Unlike those studies, [9] combined all the learning types of activities into one metric, but they neglected the semantic of those activity types. Learners tend to have different behaviors to achieve their goals. For instance, some learners dedicate time to explore the course contents, but others choose collaborative tasks to obtain knowledge [13]. Although a recent study used demographic and all the activity types of clickstreams data as features, they combined all the assessments types into a single metric [14].

However, most of the previous work lacked a significant aspect of data understandability.

In the current study, we use demographic data, VLE activity-type data, and assessments types of data. We did not consider features like course model, model presentation, and course duration because they have the same values in all samples. Also, we aggregate learners' clicks of each activity type and use it as a feature. Unlike previous research, we transform each learner's assessment type to a feature based on the score and weight of the assessment.

3 Methodology

3.1 Data Preparation

In this study, data from "DDD" module and "2013B" presentation is explored. The course belongs to "Science, Technology, Engineering, and Mathematics" category, and started in February of 2013. The module has 1303 enrollments, 536,837 VLE interactions, 4,903 available learning activities, 173,912 submitted activities, 14 available assessments. The course duration is 240 days.

Regarding the withdrawal issue, 77 students dropped out before the course started, i.e. before day 0, and the total number of withdrawals during the course is 355. An interesting observation is that the more learners engaged, they unlikely to dropout the did not. For example, the dropout rate of learners who have engaged with the course until 74% of the course duration is 9.99% vs. 28.11% until 15% of the course length.

OULAD data could not be used directly in the ML models. We preprocessed the raw data (OU tables) using Python to obtain one CSV file acceptable by ML models. Algorithm 1 shows the feature extraction steps to obtain final data. Each data row will be represented as [Demographic, Sum_Clicks_ $S_i_A_j$,$Score_S_i_A_{TMA}$,$Score_S_i_A_{CMA}$,$Score_S_i_A_{Exam}$] after applying the algorithm. The student id in the original data tables is converted as row index in the transformed data.

Algorithm 1 relational to tabular transformation

1) **For** each Student $S_i, i = 1 \cdots m$, do
2) **For** each module activity type $A_j, j = 1 \cdots n$, do
3) **For** each Site W per activity Type $W_{j,k}, k = 1 \cdots s$,
4) Sum_Clicks_$S_i_A_j = \sum_{k=1}^{s} sum_click(W)$
5) **End**
6) **End**
7) **For** each assessment type in assessments $A_p, p = 1 \cdots 3$,
8) $Score_S_i_A_{p=} \sum score_p * weight_p/100$
9) **End**
10)
11) **End**

In addition to demographics, the final dataset file has a column for each activity type in the course; the column represents the aggregated number of sites' clicks of the student. The selected model of the study has 11 learning activities types, namely: externalquiz, externalquiz, glossary, homepage, oucontent, ouelluminate, ouwiki, page,resource, subpage,url. For the assessments, as can be seen from the algorithm, we sum up all assessments that belong to a type (i.e., TMA, Exam, etc.) after multiplying it by its weight and dividing by 100.

The final_result field in studentInfo table is the target variable. As shown in Table 1, it takes 4 values. Since this study deals with the dropout issue as a binary classification task, the class variable is transformed into two values: 1 for withdrawal and 0 for other cases.

Table 1 Database features descriptions

Table	Column	Description
studentInfo	id_student	Student identification number
	gender	Sudent's gender
	region	The geographic region the student belongs to
	highest_education	the student's highest education level on entry to the course
	imd_band	Index of multiple deprivations of the place where the student lived during the course
	age_band	Students' ages range (0–35, 35–55, 55 <)
	num_of_prev_attempts	A student's previous attempts number to pass the course
	studied_credits	Total credits for the modules the student is currently studying
	disability	disability status (yes or not)
	final_result	Student's final result in the module
Course	code_module	the model identifier
	code_presentation	Presentation's code name
	length	Number of days from beginning to end of the module
studentRegistration	date_registration	Number represents the day when the student registered in the module presentation
	date_unregistration	A number to represent a day when the student canceled registration. Its value is empty for students who finished the course, and withdrawal for students who unregistered
Assessments	id_assessment	The assessment identification number
	assessment_type	The assessment's type. Its values are Tutor Marked Assessment (TMA), Computer Marked Assessment (CMA), and Final Exam (Exam)
	date	The cut-off day of the assessment
	weight	The. Exams weight 100%, all other assessments weight 100%
studentAssessment	date_submitted	The last day to submit the assessment
	is_banked	A flag to indicate whether the result has been transferred from a prior module presentation or not
	score	The assessment score
vle	id_site	The identification number of the material
	activity_type	A module material role that reflects number of clicks a student performed during the term
	week_from	The week to start using the material

(continued)

Table 1 (continued)

Table	Column	Description
	week_to	The week to start using the material
studentVle	date	The day of student's interaction with the material
	sum_click	The number of times the learner clicked the material

3.2 Missing Values

There were 4 rows with missing values in date_registeration and date_unregistration columns. We dropped those rows. The imd_band column had 61 missing values. To overcome this issue, MissForest method was used. MissForest is a non-parametric missing value imputation technique that works with mixed-type data and is based on random forests [15]. Some features created based on VLE activities in the previous step had missing values. For example, if a student did not click on the homepage and did not take the exam, we substituted zero values in the corresponding attributes. Date_unregistration is empty for students who finished the course, we substituted 240 values, the course duration.

3.3 Unbalanced Data

Class imbalance arises in classification problems in which the distribution of classes samples is not equal. In the binary case, samples of the majority class outperform the other minority class. Using unbalanced data to train a standard classifier causes predictive biases that cause poor performance on the class(es) with lower prior probabilities. In [16], an algorithm, SMOTEFUNA, generates a new synthetic example by employing the furthest neighbor of a candidate example. Its main advantage is a non-parametric algorithm; it does not need tuning like K in SMOTE algorithm because it is a parameter-less technique. Consequently, it is easier to be employed by practical applications. The authors showed it as an efficient alternative to SMOTE by comparing the results. Researchers have employed SOMTE in the VLE domain [17], but this is the first time to use SMOTEFUNA in this domain.

Figure 2 shows the dataset has 33% withdrawn samples. The figure on the right shows the distribution of binary classes: 1 refers to withdrawn and 0 refers to non-withdrawn.

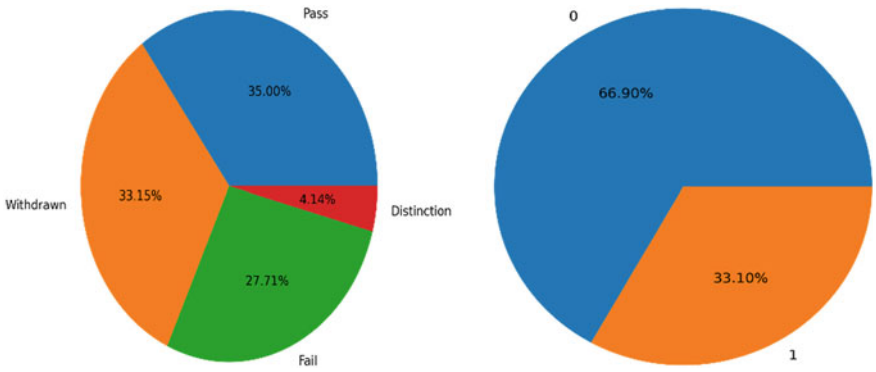


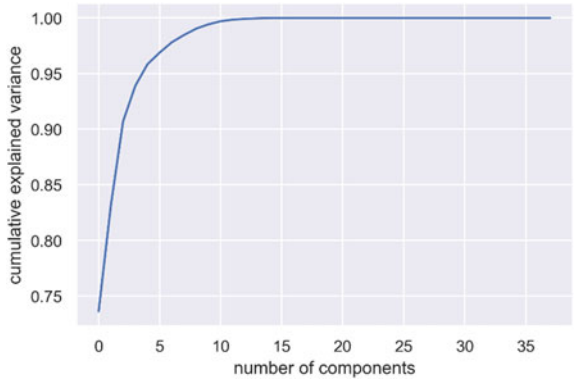
Fig. 2 Class values distribution

3.4 Principal Component Analysis (PCA)

PCA is a method used to reduce the large dimensionality of datasets; it decreases information loss but increases interpretability at the same time [18]. It creates new uncorrelated variables that minimize variance continuously; newfound variables, the principal components, reduce to solving eigenvalue/eigenvector problem. An important PCA's goal is to extract the most important information from the dataset.

After applying one_hot encoding on categorical features, the dataset ended up with 38 features. According to Fig. 3, we need about 5 components to retain about 95% of the variance.

Fig. 3 PCA components versus variance



3.5 Machine Learning Techniques and Evaluation

This study uses five classifiers from Scikit-learn library [19], namely: Logistic Regression (LR), Linear Discriminant Analysis (LDA), Decision Tree Classifier (CART), Gaussian Nave Base (NB), and Support Vector Machines Classifier (SVC). All classifiers had default values—no hyperparameter tuning.

After training the classifiers, a fivefold cross validation approach was used to evaluate the performance of the models. To obtain the prediction results, trained models are tested using test data and a count of the number of true positive (TP), true negative (TN), false positive (FP), and false negatives (FN) was used for performance evaluation. Accuracy, Precision, Recall, and F1 Score were used in this study to measure performances. These measures are defined according to the following formulas:

$$\text{Accuracy} = \text{TP} + \text{TN}/\text{TP} + \text{FP} + \text{FN} + \text{TN} \tag{1}$$

$$\text{Precision} = \text{TP}/\text{TP} + \text{FP} \tag{2}$$

$$\text{Recall} = \text{TP}/\text{TP} + \text{FN} \tag{3}$$

$$\text{F1 Score} = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision}) \tag{4}$$

where:

TP = correctly predicted positive values number.

TN = correctly predicted negative values number.

FP = incorrectly predicted as positive values number (i.e., withdrawn in our case, but the true label is non-withdrawn).

FN = incorrectly predicted negative values number (opposite of the above).

4 Experiments and Results

This utilizes various ML algorithms to predict whether a student will dropout of the course or not and investigates their performances. After the tables mapping step, the categorical features were converted to one-hot representation to be more expressive. Then, the missing values inimd_band column were handled by MissForest algorithm. The third step was transforming the class variable to binary case (1 for withdrawn and 0 for non-withdrawn). Fourthly, the resulting data had approximately 33% withdrawn samples so that SMOTEFUNA, a new Synthetic Minority Oversampling Technique, was utilized.

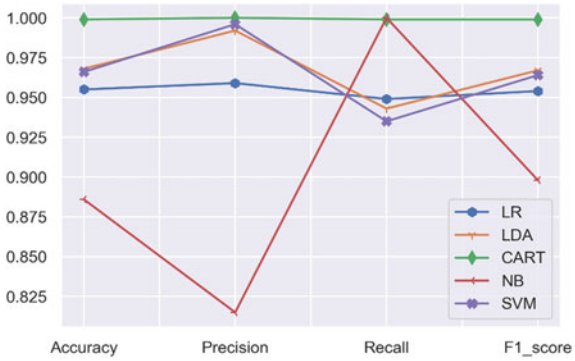


Fig. 4. Performance metrics of the models using all the features

Figure 4 shows the performance results of the classification models. As can be seen in the figure, CART model is best in accuracy measure with a score of 0.999. Both SVM and LDA scored 0.967. NB is the lowest accuracy model with 0.892.

In terms of precision, the best model was CRAT with 1. The lowest model was NB with a score of 0.822. Other models scored high. NB model was the best in Recall; it scored 1.

Regarding the F1-score, CART was the winning model with 0.999. Current study results outperformed [14, 20] by more than 12% and 14%. The best F1-score obtained in [14] was 0.874. While the minimum F1-score in this study is 0.902. Another study conducted by the same authors had the best F1-score equal to 0.853 [20] (Table 2).

To assess the effect of features reduction using PCA, the reduced dimensional dataset is used after applying PCA and setting it to 5 components. Table 3 shows the

Table 2 Performance measures without PCA

Model	Accuracy	Precision	Recall	f1 score
LR	0.955	0.963	0.945	0.954
LDA	0.967	0.993	0.941	0.966
CART	0.999	1.000	0.999	0.999
NB	0.892	0.822	1.000	0.902
SVM	0.967	0.998	0.935	0.965

Table 3 Performance measures after using PCA (5 components)

Model	Accuracy	Precision	Recall	f1 score
LR	0.971	0.986	0.956	0.970
LDA	0.907	0.981	0.829	0.898
CART	0.921	0.924	0.919	0.921
NB	0.923	0.995	0.850	0.916
SVM	0.956	0.992	0.917	0.953

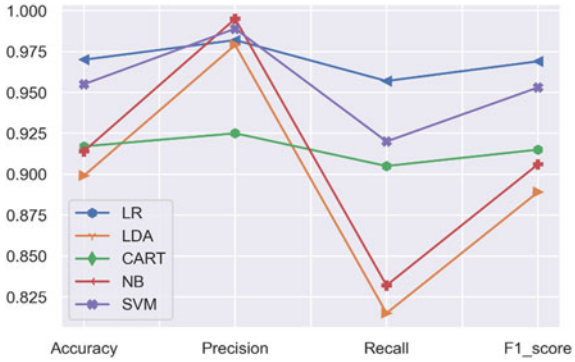


Fig. 5 Performance metrics of the models using 5 PCA components.

results of the considered models. Some models' F1-scores are improved like LR and NB, but others got lower scores. The best model hence is LR, and the worst model is NB (Table 3).

The below figure shows the performances of the models using a reduced dimensional dataset. As it is shown, LA, LDA, CART, and SVM models precisions' values are close to each other, but they have different recall values. Hence, they have different f1 score values. LR scored 97% as the best model. The effect of using few features is noticeable in LDA and CART models as their accuracy results decreased 6% and 7%, respectively (Fig. 5).

5 Conclusion

Predicting whether a learner will dropout or not is a challenging task in Virtual Learning Environments as it makes stakeholders understand learner's motivation and the desire to finish the course. This research contributes to solving dropout prediction by building a prediction framework utilizing data mining techniques. The framework includes preprocessing steps, missing values imputation, feature reduction, building predictive models, and evaluating the models' performances. The results showed that Decision Tree classifier outperformed other models in terms of f1 score measure, of up to 0.999. Compared with previous research, our results are better by 12%. Being able to predict dropout based on learner's data, institutions will take preventative measures on time. Although this study' results are interesting and promising, future work directions would include predicting at-risk students at the early stages of learning. Developing a real-time learning analytics dashboard is another direction by which instructors can identify struggling learners and help them on time. Moreover, attributes like assignments submission date, deadlines date, and student's weekly clickstreams in VLE can be used to build a sequential model using time series which can result in monitoring the students in weekly frequency [21].

References

1. S. Jiang, D. Kotzias, Assessing the use of social media in massive open online courses. arXiv: 1608.05668 (2016)
2. W. Li, M. Gao, H. Li, Q. Xiong, J. Wen, Z. Wu, Dropout prediction in MOOCs using behavior features and multi-view semi-supervised learning. in *2016 international joint conference on neural networks (IJCNN)* (IEEE, 2016), pp. 3130–3137
3. M. Tan, P. Shao, Prediction of student dropout in e-Learning program through the use of machine learning method. Int. J. Emerg. Technol. Learn. **10**(1) (2015)
4. M.C. Mihaescu, P.S. Popescu, Review on publicly available datasets for educational data mining. Wiley Interdiscip. Rev. Data Mining Knowl. Discov. e1403 (2021)
5. M. Shridharan, A. Willingham, J. Spencer, T.-Y. Yang, C. Brinton, Predictive learning analytics for video-watching behavior in MOOCs. in *2018 52nd Annual Conference on Information Sciences and Systems (CISS)* (IEEE, 2018), pp. 1–6
6. J. Kuzilek, M. Hlosta, Z. Zdrahal, Open university learning analytics dataset. Sci. data **4**(1), 1–8 (2017)
7. J Kuzilek, J. Vaclavek, V. Fuglik, Z. Zdrahal, Student drop-out modelling using virtual learning environment behaviour data. in *European Conference on Technology Enhanced Learning* (Springer, Cham, 2018), pp. 166–171
8. M. Hussain, W. Zhu, W. Zhang, S.M.R. Abidi, Student engagement predictions in an e-learning system and their impact on student course assessment scores. Comput Intell Neurosci (2018)
9. H. Heuer, A. Breiter, Student success prediction and the trade-off between big data and data minimization. DeLFI 2018-Die 16. E-Learning Fachtagung Informatik, 219–230 (2018)
10. R. Alshabandar, A. Hussain, R. Keight, A. Laws, T. Baker, The application of Gaussian mixture models for the identification of at-risk learners in massive open online courses. in *2018 IEEE Congress on Evolutionary Computation (CEC)* (IEEE, 2018), pp. 1–8
11. S. Rizvi, B. Rienties, S.A. Khoja, The role of demographics in online learning; a decision tree based approach. Comput. Edu. **137**, 32–47 (2019)
12. L. Haiyang, Z. Wang, P. Benachour, P. Tubman, A time series classification method for behaviour-based dropout prediction. in *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)* (IEEE, 2018), pp. 191–195
13. A. Ramesh, S.H. Kumar, J. Foulds, L. Getoor, Weakly supervised models of aspect-sentiment for online course discussion forums. in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (vol. 1: Long Papers, 2015), pp. 74–83
14. F. Hlioui, N. Aloui, F. Gargouri, Withdrawal prediction framework in virtual learning environment. Int. J. Serv. Sci. Manag. Eng. Technol. (IJSSMET) **11**(3), 47–64 (2020)
15. D.J. Stekhoven, P. Bühlmann, MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics **28**(1), 112–118 (2012)
16. A.S. Tarawneh, Smotefuna: synthetic minority over-sampling technique based on furthest neighbour algorithm. IEEE Access **8**, 59069–59082 (2020)
17. G. Dimic, D. Rancic, N. Macek, P. Spalevic, V. Drasute, Improving the prediction accuracy in blended learning environment using synthetic minority oversampling technique. Inf. Discov. Deliv. (2019)
18. S. Wold, K. Esbensen, P. Geladi, Principal component analysis. Chemom. Intell. Lab. Syst. **2**(1–3), 37–52 (1987)
19. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel et al, Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
20. F. Hlioui, N. Aloui, F. Gargouri, A Withdrawal prediction model of at-risk learners based on behavioural indicators. Int. J. Web-Based Learn. Teach. Technol. (IJWLTT) **16**(2), 32–53 (2021)
21. K. Palani, P. Stynes, P. Pathak, Clustering techniques to identify low-engagement student levels. In CSEDU (2), 248–257(2021)

Cloud Computing in Healthcare Industries: Opportunities and Challenges

Vinayak Rai, Karan Bagoria, Kapil Mehta, Vandana Mohindru Sood, Kartik Gupta, Lakshya Sharma, and Manav Chauhan

Abstract In the present age, Cloud Computing is acting as a most flexible and collaborative approach that can be applied in different public and private domains. Apart from playing a crucial role in terms of employment and revenue, the healthcare sector also plays a significant role in saving the lives of numerous people. Directly or indirectly, the lives of patients depend on their previous records, new advancements in the health sector, and interconnection among various health staff. All these factors were not feasible by using traditional ways, but this has become feasible since the introduction of cloud computing in the health sector. Cloud Computing has eased the access of applications and data from any worldwide location from any device with an internet connection. In this paper, the role of cloud computing in healthcare services and the opportunities associated are presented. This paper elaborates on the challenges faced by the industries with current and future innovations of cloud computing in healthcare. Security concerns along with the benefits, risks, stakeholders, challenges associated with cloud computing in health sectors have also been discussed.

Keywords Cloud computing · Healthcare · Information technology · Public cloud · Private cloud · Hybrid cloud · Healthcare services · Cloud service providers

1 Introduction

The term 'Cloud Computing' is derived from two words, Cloud and Computing. Here, the word Cloud means the servers associated with the Internet, while the word computing means the use of operations. So, the term 'Cloud Computing' means the data centers available to the users over the Internet. Cloud computing is the

V. Rai · K. Bagoria · K. Mehta (✉) · K. Gupta · L. Sharma · M. Chauhan
Department of Computer Science and Engineering, Chandigarh Group of Colleges, Mohali, Punjab, India

V. M. Sood
Department of Computer Science & Engineering, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_53

computer system resource, especially to store data and computing power without the direct interaction of users. IT plays a big role in achieving reduced costs and improving every outcome for patients. Many hospitals in America already started to include electronic health records (EHRs), a digital medium of storing medical data or records which was edited by the Health Information Technology for Economic and Clinical Health Act (HITECH) and enforced by The American Recovery and Reinvestment Act in 2009 (ARAR). According to the recent Gartner report about healthcare prediction in 2019, artificial intelligence is moving into healthcare operations as more cloud platforms join AI and ML into their services [1]. Cloud computing not only allows the users to access data, but also includes automation backups and data recovery options in case the data ever is lost because of any reason [2]. Instead of owning and maintaining data centers physically, one can access services by cloud. There are specific types of cloud, we have to use according to our needs and these are:

Public cloud: The cloud which lets users share resources while preserving the privacy of each user's data is known as Public Cloud. Public cloud is owned and operated by third-party cloud providers who deliver their services over the internet like Microsoft Azure.

Private cloud: The cloud which is built for the usage of a single business or origination, i.e., the resources used by it cannot be used by any other business or organization is known as Private Cloud.

Hybrid cloud: The cloud which is the combination of both public and private clouds is known as a hybrid cloud. It allows data and applications to be shared between them [3]. Figure 1 depicts the various types of clouds and interconnection among these clouds.

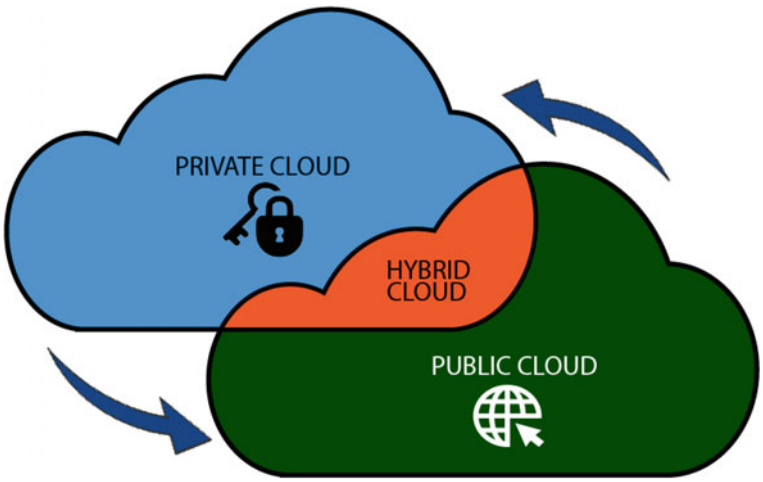


Fig. 1 Types of cloud interconnections

An electronic health record (EHR) is a digital version of a patient's health records or data. EHR shall make the record visible or instantly available to authorized users [4]. CSCC (Cloud Standards Customer Council) authors said cloud service provider data centers are highly secure and well protected [5].

1.1 Cloud Computing-Based Healthcare Services

The integration of cloud computing with healthcare plays a vital role in the improvement of various healthcare-related functionalities include chatting with the health experts, video appointments, telemedicine, care plans, or any other virtual medication adherence. The healthcare services related to cloud computing areas are listed below:

- **On-demand service**: Resources can be used without any delay when in demand.
- **Broad network access**: Services can be availed at any time having no bar of location.
- **Resource pooling**: No limit of users in a single pool, many can access the services simultaneously.
- **Elasticity**: According to the usage or requirements, resources can be provisioned.
- **Measured service**: Pay according to the user's access.

These services can be used through Mobile Apps or websites. There are numerous sectors in which healthcare can be used as shown in Fig. 2.

Google's healthcare API facilitates the sharing of data between healthcare applications and builds on the Google cloud. If this is adopted worldwide, then it shall be easier for patients and doctors to access their healthcare information by using third-party applications. AWS in the Covid-19 pandemic is collaborating with various health organizations around the world to support the world dealing with the effects of Covid-19 [6]. The two components of Cross-cloud architecture are the VMware Cloud Foundation platform and Cross-cloud services [7].

1.2 Organization of Paper

This paper is organized as follows. Section 2 represents the background of the literature. Section 3 presents the postulates of cloud computing in the healthcare industry. Section 4 emphasizes the opportunities associated with cloud computing in healthcare concerns associated with Cloud Computing in Healthcare. Section 5 elaborates on the implementation of Cloud Computing in healthcare. Section 6 presents the challenges of cloud computing in healthcare. Section 7 concludes the paper.

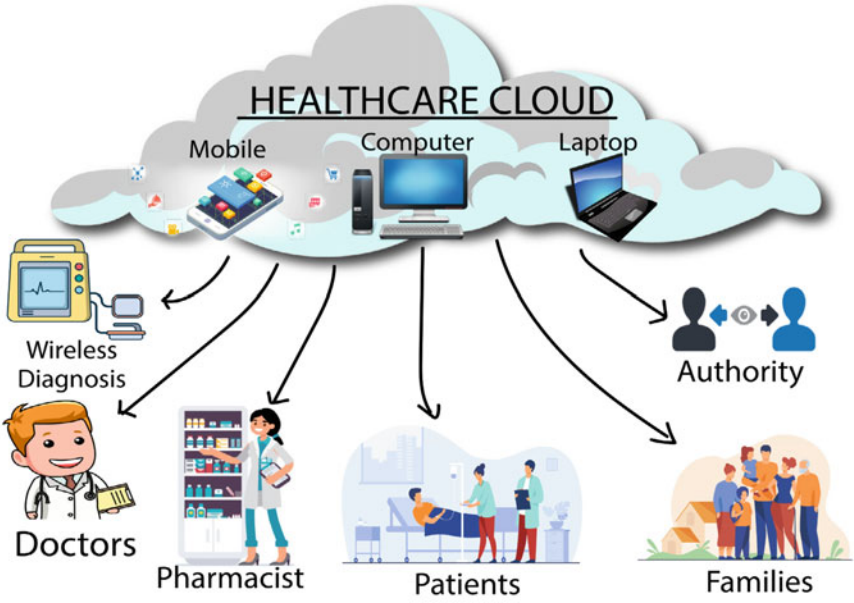


Fig. 2 Numerous sectors of healthcare cloud

2 Literature Review

As per the reports of Synoptek, $3.73 billion in healthcare was spent on cloud services in 2019, which is expected to increase this number by nearly $ 9.48 billion in 2020. As the demand is on the rise for secure access to information that should be cost-efficient, it is projected to triple the cloud-based computing in the healthcare market. According to the recent HIMSS survey, over 83% of healthcare organizations are using cloud services [8]. According to the Population Reference Bureau (PRB), 2004, there are approximately one billion people in the world who do not have access to regular healthcare, and vanishing of patient's records or test reports such problems are very common in healthcare organizations [9]. Bamiah et al. stated the numerous reasons for using cloud computing in healthcare that optimize resource management and proceed toward the maintenance of e-reports instead of maintaining the paper-wise record [10]. Mutlag et al. presented the fundamental factors for resource management in cloud-based healthcare systems. The limitations of recent methods, systems, and frameworks were also illustrated [11]. Kumari et al. addressed the fundamental opportunities and challenges for the application of cloud strategies in the healthcare industry. A three-tier architecture was introduced to use for real-world applications [12]. De la Torre-Díez et al. suggested the migration of healthcare toward cloud computing that introduces risk factors related to privacy and security [13].

Table 1 Review of literature

References	Cloud	Contribution
Nermeen et al. [14]	√	Examined the challenges for cloud adoption in eHealth in developing countries. Demonstrated the legal factors were considered as critical to governing the new system and demonstrating the dimension
Sivan et al. [15]	√	Highlighted a comprehensive study of existing cloud-based secure eHealth solutions for protecting the privacy and security of digital data
Khan et al. [16]	X	Proposed a wearable heart disease prediction system run by IoT using the MDCNN classifier
Yang et al. [17]	X	Addressed the development of eHealth system run by IoT on hardware and software
Chenthara et al. [18]	√	Security mechanisms for electronic health records maintain the confidentiality and protection of patients' information
Kim et al. [19]	√	Presented a novel Health System that supports multilevel privacy-preserving data sharing
Zhang et al. [20]	√	Proposed a method suitable for remoting health monitoring systems
Yeh et al. [21]	√	Discussed the secure and efficient scheme for cloud-assisted health system
Avila et al. [22]	√	Reviewed the applications developed in the health industry which are focused on patient care from home and implemented a service-oriented (SOA) design in architecture

Table 1 describes multiple statements from the various authors and researchers for the integration of cloud computing with healthcare.

3 Cloud Computing in the Healthcare Industry

In the past, the underutilized sector was Healthcare that needs improvement in inpatient care. Traditionally, systems were operated manually, used paperwork to maintain medical records, etc. Maintaining healthcare with the latest technologies such as the cloud is the need of the hour. While comparing, healthcare sector with other industries, there are various key differences which areas listed below:

a. The Healthcare sector strictly works under the regulations of government laws to serve patients more care and comfort.
b. Higher cost and major chances of risk involved in healthcare as compared to other industries.
c. Requires more staffing due to the numerous units such as administration, labs, patients, etc.

There might be numerous stakeholders involved in providing cloud services to medical sectors but as per NIST cloud computing reference architecture, we can define it in major five categories. These categories are described below:

1. Cloud Service providers: The associations or organizations that deliver cloud services to the consumers with the various components of cloud computing are known as Cloud Service providers. The major categories of services provided by Cloud Service Providers are Infrastructure as a service (IaaS), Software as a service (SaaS), and Platform as a service (PaaS) [23].
2. Cloud Carriers: The mediators that offer connectivity and transport of cloud services being within service providers and consumers are termed as cloud carriers.
3. Cloud Brokers: The association responsible for the management of performance, use, and delivery of cloud services is known as Cloud Brokers. Major services provided by the broker are Service intermediation, service aggregation, service arbitrage [24].
4. Cloud Auditors: In general, cloud auditors are the people who keep an eye on all the processes being implemented. There are three types of cloud auditors, i.e., security audits, privacy impact audits, performance audits.
5. Cloud Consumers: Cloud Consumers are the end-users who can browse and use the services offered by service providers. For specifying the technical performance requirements to be accomplished by the service provider, consumers use Service- level agreements (SLAs) [25].

4 Opportunities Associated with Cloud Computing in Healthcare

Despite the numerous merits of cloud computing, there exist severe concerns about cloud computing which can affect healthcare. The providers of Healthcare need their data to be available at a specific time, so any delay on the cloud platform side regarding data will hurt productivity. The Cloud Security Alliance 2010, has published a list of best practices that should be executed in the cloud network to achieve high security in the platform as services (PaaS) and services as a platform (SaaS) [26]. Insecure APIs and loose interface users can be easy targets. The disaster recovery policies and backup services appear to be cloud providers and more robust than the local data centers that organizations maintain [27]. There is an immense scope of cloud computing in the upcoming future. It provides a whole new world of jobs, platforms, services, and applications. Bart McDonough, CEO of Agio, believes the recent rapid adoption of the cloud is mainly due to the understanding of the "ease of use and scalability" of the technology [28]. Cloud computing is the widespread adoption that incurs various opportunities as listed in Table 2.

Table 2 Cloud computing opportunities

Aspect	Opportunities
Management/Administration	i. Low-cost infrastructure ii. On-demand resources iii. Payment while the use
Technological Advancements	i. IT Maintenance burden ii. Scalable resources iii. Green computing
Security	i. Data protection ii. Replication of data iii. Authenticated Information
Legacy	i. Commitment to secure user's data ii. Maintain privacy iii. Fostering government laws and regulations

4.1 Cloud Computing Factors Affecting Healthcare

Nowadays, cloud computing is used everywhere around the world. As each system has too much data to handle, which can't be stored on their PCs, thus the need for the cloud has emerged since the last few decades. Thus, in the initiation, it was mostly used by the diverse healthcare industries which have branches across distinct parts of the world [29]. But, if some add-ons are done on it to make it prominent of the characteristics, including tenacious data provided by data centers, adaptable infrastructure, security models, swift access to information, etc. [30]. Figure 3 shows the cloud computing factors which affect the health sector.

1. **Infrastructure and Dynamic Scalability**: The world's population is inflating day by day. With the advancement in population, the number of patients is also inflating. Not only does cloud usage eliminates the cost of the user, but also it will diminish the tension of the user to upgrade or manage the data as it will all be done by the vendor [31].
2. **Information Sharing**: Consider a hospital, which has different branches all over the world. Thus, this transmission can't be possible through LAN or WAN due to the tremendous amount of data and the security concerns [32]. For example, the device provided by Microsoft, HealthVault, can store enormous health information, including medical images, doctor's fax records, digital formats turning, etc.
3. **Availability**: While using the computing capacity and IT resources, if anything in the system gets flawed, then the doctor is not able to access the data in such circumstances. The uploading of data or usage of the information on the cloud is not too easy, instead, it requires extensive knowledge [33]. For example, let us imagine a small hospital in which the patient's data is stored in files rather than the storing of data on the cloud.

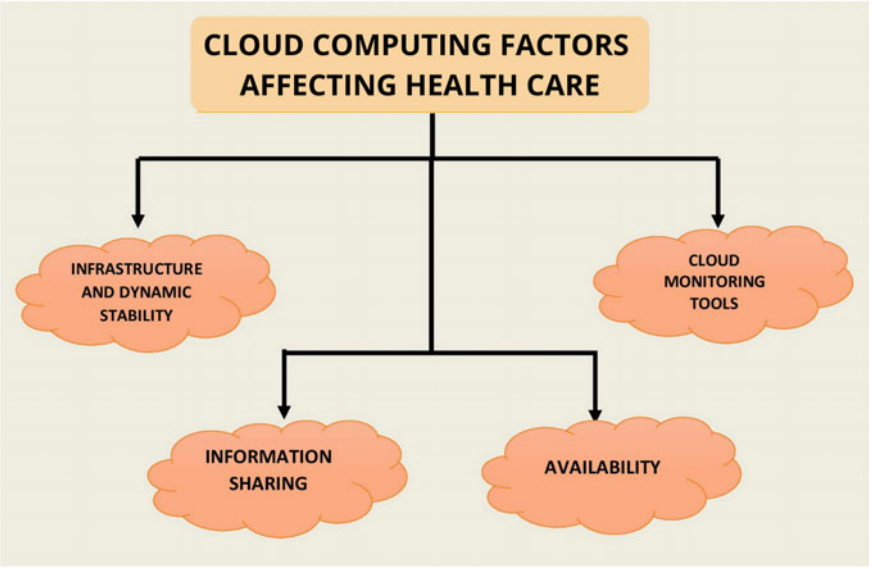


Fig. 3 Cloud computing factors affecting healthcare

4. **Cloud Monitoring Tools**: The data stored in the cloud is of no use until the proper management and usage of data are not done. The only platform which has its monitoring tool is Amazon Web Services. For example, while working in a team, there is always a team leader to check if the functioning of each member of the group is done appropriately [34].

5 Implementation of Cloud Computing in Healthcare

Cloud computing is defined as the storing of tremendous volumes of data in a secure environment that is sustained by IT professionals. Since the EMR mandate has been introduced, the healthcare organization has adopted cloud computing for preserving and accumulating patient EMRs. The asset of this collaborative approach enables healthcare services between them to offer an adequate and swift response [35].

It helps to advance the quality of patient service by allocating information across healthcare organizations. Therefore clinics, hospitals, pharmacies, imaging centers, and insurance companies can share medical records of patients. This information can be produced anytime and anywhere by certain entities [36]. Figure 4 depicts the interconnection among storage devices, access points, and numerous clients. These service providers resort to different safeguard mechanisms which is the most reliable and safest option for data-related needs [37].

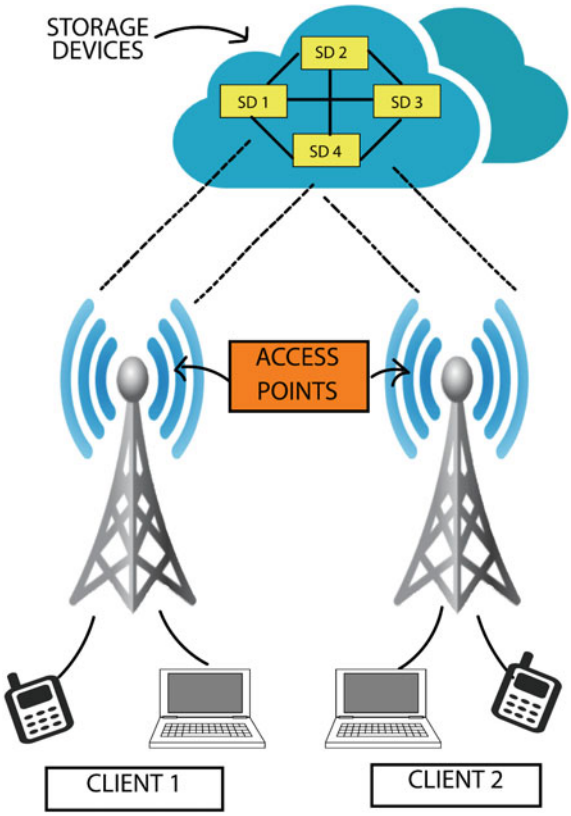


Fig. 4 Client–server connection in cloud

6 Challenges of Cloud Computing in Healthcare

Despite numerous progress in the healthcare industry, the enormous concern which has remained till now is the patient's security issue. Hence, knowing the various challenges related to cloud computing, the user can try to overcome those disadvantages. The following are the various challenges of cloud computing in healthcare.

1. **Privacy and security challenges**

   Patient's complete personal, confidential, or private information is stored in the cloud, which is not to share with anybody else [38]. Despite having a foolproof cloud storage system provided by Google, Microsoft, Amazon, etc., there remains a doubt in a person's mind regarding his data. Global concerns related to data jurisdiction, the privacy of data, security, and compliance are having a huge impact on the adoption of these cloud technologies by healthcare organizations [39].

2. **Interoperability And Portability**

   Interoperability refers to exchanging information and services—'Saas' and 'Iaas' between two or more cloud systems but it is not an easy task because it requires a large cost and due to a lack of standards and management tools users are stuck in using the same vendor's services [40]. Thus, creating an issue for the user who is not satisfied with the vendor's service of that particular cloud. This situation is known as 'vendor lock-in'. Moving data and applications onto other clouds is a big challenge [41]. Gartner's Skorupa, who is a Research analyst, explains that moving large data sets to clouds even after resolving interoperability is still an issue because of latency and time problems [42].
3. **Reliability and Performance**

   As far as challenges have been discussed, the most challenging part is reliability. Reliability means the capability of the system to work in a given amount of time in given conditions or that a system is more efficient and reliable which is capable of tolerating faults and failures, it does not affect the working of the system [43]. Numerous failures affect our cloud such as— overflow, timeout, data resource missing, network failure, database failure, hardware, and software failure. These are categorized into 2 categories— Request stage failures and execution stage failures [44].
4. **Computing Technologies**

   The evolution of Serverless Computing technology enables developers and users to shift operational data and services to AWS. It gives an environment to users and developers to deploy code, and run our applications without thinking of servers needed [45]. Another important factor of using serverless computing is scalability as it has a broad range of services and users and providers need not be worried about servers. Due to this, their main focus will be on services only and hence good quality services will be given by the provider which will improve the reliability [46].

7 Conclusion and Future Scope

Cloud Computing, one of the burgeoning technologies in healthcare deals with storing and retrieving Electronic Health Records of patients over the cloud server. In this paper, we concluded that integration of the healthcare cloud among the doctors, medical staff, and patients could be fruitful but the integration of ambulances with the patients and doctors could prove to be more worthy and save the lives of tremendous people. The availability of resources and information on the cloud makes further processing easy in bad circumstances. As cloud computing is the mere future of the

healthcare sector, organizations providing cloud services should look into the resolution of it by any of the means including Artificial Intelligence, Machine Learning technologies, or Serverless Computing. This paper addresses the opportunities and challenges which are aimed to identify the state of research and determine the potential areas of future research in the domain.

References

1. P. Srivastava, R. Khan, A review paper on cloud computing. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **8**(6), 17–20 (2018)
2. A. Shahidinejad, G.A. Mostafa, M. Masdari M, Resource provisioning using workload clustering in cloud computing environment: a hybrid approach. Clust. Comput. J. Netw. Softw. Tools Appl. 319–342 (2021)
3. M. Anuradha, T. Jayasankar, N.B. Prakash, M.Y. Sikkandar, G.R. Hemalakshmi, C. Bharatiraja, A.S. Britto, IoT enabled cancer prediction system to enhance the authentication and security using cloud computing. Microprocessors and Microsystems (2021)
4. M. Sadeeq, M. Nasiba, R.M. Subhi, D.M. Ahmed, R. Zebari, IoT and cloud computing issues, challenges and opportunities: a review. Qubahan Acad. J. (2021)
5. S.E. Shukri, R. Al-Sayyed, A. Hudaib, S. Mirjalili, S, Enhanced multi-verse optimizer for task scheduling in cloud computing environments. Expert Syst. Appl. **168**, 114230 (2021)
6. I. Mohiuddin, A. Almogren, Security challenges and strategies for the IoT in cloud computing. in *11th International Conference on Information and Communication Systems* (2020), pp. 367–372
7. F. Wang, G. Xu, G. Xu, Y. Wang, J. Peng, A robust IoT-based three-factor authentication scheme for cloud computing resistant to session key exposure. Wirel. Commun. Mob. Comput. (2020)
8. F.E. Samann, S.R. Zeebaree, S. Askar, IoT provisioning QoS based on cloud and fog computing. J. Appl. Sci. Technol. Trends 29–40 (2021)
9. A. Shakarami, M. Ghobaei-Arani, M. Masdari, M. Hosseinzadeh, A survey on the computation offloading approaches in mobile edge/cloud computing environment: a stochastic-based perspective. J. Grid Comput. 1–33 (2020)
10. M. Bamiah, S. Brohi, S. Chuprat, J.L Ab Manan, A study on the significance of adopting cloud computing paradigm in the healthcare sector. in *Cloud Computing Technologies, Applications and Managemen. International Conference on IEEE* (2012), pp. 65–68
11. A.A. Mutlag, M.K.A. Ghani, N. Arunkumar, M.A. Mohamed, O. Mohd, Enabling technologies for fog computing in healthcare IoT systems. Future Gener. Comput. Syst. 62–78 (2019)
12. A. Kumari, S. Tanwar, S. Tyagi, N. Kumar, Fog computing for healthcare 4.0 environment: Opportunities and challenges. Comput. Electr. Eng 1–13 (2018)
13. I. De la Torre-Díez, F.J. Díaz-Pernas, G. Fernández, M. Antón-Rodríguez, M. Martínez-Zarzuela, D. González-Ortega, D. Boto-Giralda, (2012)
14. N. Mekawie, K. Yehia, Challenges of deploying cloud computing in eHealth. Proced. Comput. Sci. **181**, 1049–1057 (2021)
15. R. Sivan, Z.A. Zukarnain, Security and privacy in cloud-based E-Health system. Symmetry (2021)
16. M.A. Khan, An IoT framework for heart disease prediction based on MDCNN classifier. IEEE Access 34717–34727 (2020)
17. M. Yang, Y. Hara-Azumi, Implementation of lightweight eHealth applications on a low-power embedded processor. IEEE Access 121724–121732 (2020)
18. S. Chenthara, K. Ahmed, H. Wang, F. Whittaker, Security and privacy-preserving challenges of e-health solutions in cloud computing. IEEE Access 74361–74382 (2019)

19. J.W. Kim, K. Edemacu, B. Jang, Multilevel privacy-preserving data sharing in a collaborative eHealth System. IEEE Access 109910–109923 (2019)
20. Y. Zhang, C. Xu, H. Li, K. Yang, J. Zhou, X. Lin, HealthDep.: an efficient and secure deduplication scheme for cloud-assisted ehealth systems. IEEE Trans. Ind. Inform. 4101–4112 (2018)
21. L. Yeh, P. Chiang, Y. Tsai, J. Huang, Cloud-based fine-grained health information access control framework for lightweightiot devices with dynamic auditing and attribute revocation. IEEE Trans. Cloud Comput. 532–544 (2018)
22. K. Avila, P. Sanmartin, D. Jabba, M. Jimeno, Applications based on service-oriented architecture (SOA) in the field of home healthcare. Sensors (2017)
23. A. Ahmed, E. Mohamed, S. Ahmed, A.M. Riad, A machine learning model for improving healthcare services on cloud computing environment Measurement (2018)
24. K. Silambarasan, P. Kumar, Healthcare service in cloud and internet of things using cuckoo search and PSO optimization techniques. J. Comput. Theor. Nanosci. 3571–3575 (2018)
25. T. Saheb, L. Izadi, *Paradigm of IoT big data analytics in the healthcare industry* (A review of scientific literature and mapping of research trends, Telematics and Informatics, 2019), pp. 70–85
26. C. Kotas, T. Naughton, T. Imam, A comparison of Amazon web services and microsoft azure cloud platforms for high-performance computing. in *IEEE International Conference on Consumer Electronics* (2018)
27. E. Jonas, J.S. Smith, V. Sreekanti, C. Tsai, A. Khandelwal, Q. Pu, V. Shankar, J. Carreira, K. Krauth, N. Yadwadkar, J.E. Gonzalez, R.A. Popa, I. Stoica, D.A. Patterson, Cloud Programming Simplified: A Berkeley View on Serverless Computing. Cornell University (2019)
28. A. Omar, A. Shrestha, J. Soar, S.F. Wamba, Cloud computing-enabled healthcare opportunities, issues, and applications: a systematic review. Int. J. Inf. Manag. (2018)
29. X. Xiaolong, L. Qingxiang, Y. Luo, K. Peng, X. Zhang, S. Meng, L. Qi, A computation offloading method over big data for IoT-enabled cloud-edge computing. Future Gener. Comput. Syst. 522–533 (2019)
30. A. Darwish, A. Hassanien, A. Elhoseny, A.K. Sangaiah, M. Khan M, The impact of the hybrid platform of internet of things and cloud computing on healthcare systems: opportunities, challenges, and open problems, J. Ambient Intell. Hum. Comput. (2017)
31. J. Ren, G. Yu, H. Yinghui, Y.L. Geoffrey, Collaborative cloud and edge computing for latency minimization. IEEE Trans. Vehicul. Technol. (2019)
32. L.M. Dang, M.J. Piran, D. Han, K. Min, H. Moon, *A Survey on Internet of Things and Cloud Computing for Healthcare* (Sejong University, Department of Computer Science and Engineering, 2019)
33. A. Giuseppe, P. Valerio, P. Antonio, Industry 4.0 and Health: internet of things, big data, and cloud computing for healthcare 4.0. J. Indust. Inf. Integr. (2020)
34. P. Dong, N. Zhaolong, S.O. Mohammad, X. Jiang, Y. Guo, H. Xiping, H. Bin, S. Balqies S, Edge Computing Based Healthcare Systems: Enabling Decentralized Health Monitoring on Internet of Medical Things (2020)
35. M.S. Hossain, G. Muhammad, Cloud-assisted Industrial Internet of Things (IIoT)- Enabled Framework for health monitoring (2016)
36. C.L. Chen, P.T. Huang, Y.Y. Deng, H.C. Chen, Y.C. Wang, A secure electronic medical record authorization system for smart device application in cloud computing environments. Human-Centric Comput. Inf. Sci. (2020)
37. F. Workneh, A. Adem, R. Pradhan, Understanding cloud-based health care services with its benefits. in *2018 Second International Conference on Inventive Communication and Computational Technologies* (2018)
38. Elhoseny M., Abdelaziz A., Salama A.S., Riad A.M., Khan M., Sangaiah A.K.: A hybrid model of Internet of Things and cloud computing to manage big data in health services applications, Future Generation Computer Systems, pp. 1383–1394 (2018).
39. E.R. Muhammad, K. Micheal, M. Jayabalan, Privacy and security challenges towards cloud-based access control in electronic health records. Asian J. Inf. Technol. (2017)

40. C. Kyungyong, C.P. Roy, Chatbot-based healthcare service with a knowledge base for cloud computing. Clust. Comput. 1925–1937 (2019)
41. T. Chandu, S. Revathi, M. Gunasekaran, R. Varatharajan, M.K. Priyan, Centralized fog computing security platform for IoT and cloud in healthcare system. Fog Comput. Breakthr. Res. Pract. (2018)
42. P.K. Senyo, E. Addae, R. Boateng, Cloud computing research: a review of research themes, frameworks, methods, and future research directions. Int. J. Inf. Manag. 128–139 (2018)
43. J. Moura, D. Hutchison, Review and analysis of networking challenges in cloud computing. J. Netw. Comput. Appl. (2016)
44. K. Mehta, Y. Kumar, Implementation of efficient clock synchronization using elastic timer technique in IoT. Adv. Math. Sci. J. 4025–4030 (2020)
45. V. Mohindru, Y. Singh, R. Bhatt, Securing wireless sensor networks from node clone attack: a lightweight message authentication algorithm. Int. J. Inf. Comput. Secur. **12**, 217-233 (2020)
46. E. Jonas, J. Schleier-Smith, V. Sreekanti, C.-C. Tsai, A. Khandelwal, Q. Pu, V. Shankar, et al., Cloud programming simplified: A berkeley view on serverless computing. arXiv preprint arXiv: 1902.03383 (2019)

Security and Privacy

The Latest Trends in Collaborative Security System

Monika Arora and Sonia

Abstract A collaborative system is an extended Information system. It facilitates the sharing and caring of documents, data, files and information among the employees and team members of an organization. An amalgam of Artificial Intelligence and blockchain is becoming popular in business day by day. Blockchain usage with AI has an advantage that it mitigates the risk factors associated within the business and assists a company in getting a secure collaborative system. This paper aims to use artificial intelligence for business in contemplation with the traits of blockchain (BC) and deep learning (DL) components. Overall presented research explores the relationship between BC, DL and AL with the benefits and challenges of using enhanced Artificial Intelligence, influenced by blockchain and deep learning, in business. This paper concluded with the future recommendations of Artificial Intelligence used in business in collaboration of blockchain, deep learning and Internet of Things that will definitely provide a future edge to a business.

Keywords Artificial intelligence · Deep learning · Blockchain · Information systems · Security

1 Introduction

Every business needs to use all its resources and gain the maximum profitability. There are many disruptive technologies that are used in almost every business, some of them are artificial intelligence, blockchain and deep learning. Every business is using these technologies directly or indirectly to provide support in the ecosystem of business [1]. The business environment uses all these important technologies in one or the other way. Artificial intelligence (AI) is considered to be a discipline of computer science which uses and imitates the intelligence of humans [2]. The application of AI can be in perception in visuals, recognition of speech that helps in decision-making.

M. Arora (✉)
Apeejay School of Management, Dwarka, New Delhi, India

Sonia
Yogananda School of AI, Computer & Data Science Shoolini University, Solan, H.P, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_54

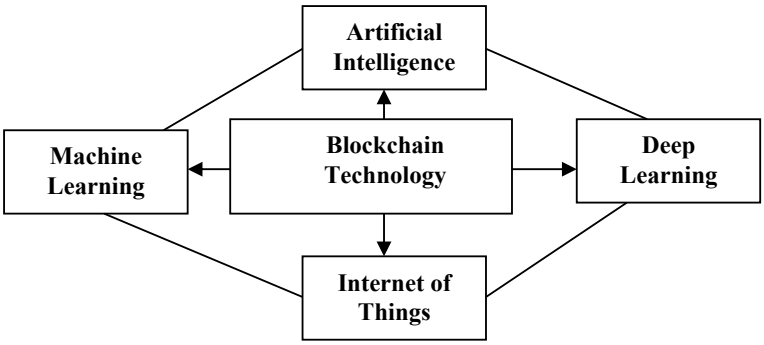


Fig. 1 Contribution of latest technologies framework

It also helps in translation and creation of machines that use repetitive tasks called machine intelligence. It will take care of all such human acts used to build a system. The tasks which are done by humans efficiently, Artificial Intelligence intervened those tasks with the use of the computers. Computers use AI and perform the task better than humans. Human beings create computer-based systems using data and information that happened in the past, it also enables computers to study the learning pattern from the data stockpiled to use them. The data stores and structures evolve procedures to the correct use of the data and information. Artificial Intelligence (AI) can also be used in many emerging areas such as marketing where the data used in AI applications to create a machine learning system will help the business in anticipating the customer's action and also help the business to gain the profit [3]. The progression of AI to the next step toward the future will take it to provide big data solutions using analytical tools and techniques. The interrelation between the four collaborative systems works as under shown in Fig. 1.

These solutions will change the entire business perspectives and their working. The data used in the creating machine learning pattern is based not only on the target audience but also on the application and best use of AI into the business. AI considers the history data in developments of its application, it uses the blockchain infrastructure that will support the implementation of AI in the business and system. The support of AI and blockchain and vice versa will do wonders in the field of business.

As blockchain is a growing area in the technical field of business and it is a disruptive technology, take the business to the next generation. In this era of businesses, the new internet will be grabbing the whole scenario and the market needs and requirements. Many developed countries have already started using the concept of blockchain in their businesses. They have also started exploring the use and possibility of blockchain in their organization. The primary objective of blockchain is that it uses distributed databases and the ledger entries will be updated in all the local databases and they don't have centralized database servers. The real-time update will help the system to minimize the frauds in the businesses. There will be a great shift from e-commerce companies to use blockchain technology in future. Once the

companies will move to the blockchain in totality, there will be a huge paradigm shift in businesses. The blockchain can be easily fitted to any business scenario and even filtered to any industry. If the businesses have not changed themselves, they will be in loss. The organizations have to implement the blockchain for their own benefits.

A big fancy word "Blockchain technology" records every transaction in the database. The blockchain is a chain of blocks where any change in any two entities will be recorded in the database. Whenever there is any change, it must be recorded in the database. Also, it contains a confirmable recorded history. The blockchain offers improved transparency and reliability. All the transactions are available for public view over the network [4]. If something happens between two parties and they don't update the other party's record, that will not match and it can be considered fraud. The data has to be updated at every moment of time. The blockchain also maintains privacy and transparency. The Internet of things (IOT) contributes in talking to each other and helps in machine learning techniques to evolve further. The renowned concept of deep learning is defined as deep learning into deep, where the slicing and dicing of data will be studied and will be helping to create the pattern for machine learning. ML is associated with deep learning which helps in creation of neural networks and helps then these patterns can be used in decision-making. These networks will help to create the AI patterns and finally in machine learning algorithms for decision-making. Deep learning will study in depth and in detail to anything that is in extending in layers [5]. The learning in depth means the gaining of acquaintance and skills through research study and experience that are used in making and building the machine learning patterns. Machine learning (ML) is a subset of AI and contributed to the blockchain technology. The objectives of the research study are based on the learning can be in different forms; it can be unsupervised, partially supervised and supervised. The objectives covered in this paper are as follows:

(a) The Machine Learning (ML) is a subset of Artificial Intelligence (AI) that is based on defined patterns with labels.
(b) The incorporation of blockchain in ML and AI has been studied.
(c) The proposed model was prepared for using all the three technologies for the business and useful for taking the decision.

In this paper, Sect. 1 provides an in-depth understanding of Artificial Intelligence (AI), blockchain (BC) and deep learning (DL). It also discusses the relationship between the three. In Sect. 2, background and motivation covered that is the need of this topic to study, Sect. 3 highlights the work related to the topic by other authors, and it also throws light on relationships of Data, Information, Artificial, deep learning (DL), blockchain (BC) using tool for creating visual analytics. Section 4 discussed the data analysis and interpretation and Sect. 5 has deliberated the proposed model with all three techniques. In Sect. 6, conclusion and future of this research study has been made and used in business.

2 Background and Motivation

In 2008, the blockchain technology was introduced by the name of Bitcoin, by Satoshi Nakamoto. This technology helps to track transactions in a public ledger between different users. These days it is used commercially also. In various applications viz tracking ownership documents, digital assets or voting rights are some of the applications of blockchain. A blockchain system used for storing digital information. This information is in encrypted form and also it is in distributed form. This will create an extremely secured database. In blockchain the information is stored in a structured way and also available to the public. Human imagination is facilitated by AI and there are many applications where AI plays an important role in Business. The dimension of field Computer science has developed in the era of intelligent computers and further it is merging to a virtual reality. The computers use AI concepts. Projects that use work with humans are called human intelligence. This technology uses artificial neural networks. The learning by computers is machine learning and deep learning wherever possible. Deep learning or deep structured learning or hierarchical learning uses machine learning-based methods and are depending on representations of learning data. Learning can be of three types: supervised, semi-supervised or unsupervised. AI uses hand holding by blockchain and provides the solution using the agreement among experts using these technologies for business uses and implications. The joint use of AI and blockchain uses a paradigm shift that can take the business toward development.

The two progressive technologies blockchain has changed the perception of users and allowed the transactions to be more transparent. On the other hand AI has opened new horizons in business and helps us to optimize routine work. It switches the attention of human workers involved in tasks which are more important and complex.

Blockchain uses the technology for the infrastructure setup in the implementation and use of it. It also uses the strong support from Artificial intelligence for the trust. As AI uses the history data and helps in getting a pattern that helps people to work better in decision-making, human reactions help the system to create the machine pattern and that helps in implementation of AI in system. Humans help in creating the neural networks to form an artificial network called artificial neural network. This is based on experiences of the human and also helps them to form a machine learning algorithm. The developers of the algorithm are not comfortable in forecasting but it describes its way of thinking for decision-making. The AI systems operate on the basis of using complex patterns and complex decision trees will provide the results through black box and match it with human decision. Now, the computer analyzes the huge amount of data with the help of technology, or capability of the computer. The memory used in the machine has no comparison with the brains of the most intelligent people in the world. The data is increasing day by day due to many means of collection of data from various sources and also in various structures. The algorithms are developed based on the instructions provided. AI uses the history data and it will help to give the input to the blockchain. AI and blockchain together work for the better decision-maker and do wonders in society. The more the data

used in the AI the better are the results. Blockchain assures the security of data and information which is stored in the form of data and information.

The use of AI will help in efficiency of data/information of the blockchain technology. All the transactions are verified/authenticated by the miners. AI also trained the machine algorithm that helps efficient ways for taking decisions. Using blockchain technology every part of information is secure and transparent and recorded over the network. As the data storage and the number of blocks will be increasing every moment, the chain becomes heavier. The blockchain with the use of machine learning algorithms will optimize the data storing methods.

The integration of various technologies where AI is decentralized, parallel computing is blockchain. Optimized data is machine learning that uses all the computation power to the complete degree. It also analyzes the huge data quickly.

The splitting of the data to the smaller units will analyze the data in much detail and then integrate and unite it into an integral database. This database is available worldwide and the information will be available to all the members of the network. This advancement helps to use efficient data everywhere and anytime. All the decisions are saved in the database, the decision is more powerful and correct as it is comprehensible and transparent.

Data analysis works on the progressive data in government institutions and businesses. It deals with the summarization, classification and analysis of large volumes of text data. Almost every sector the analysis of big data is happening for decision-making. The indicative sector is as follows:

- Financial sector analyzes the market of currency, estate and securities. It creates the forecasts, structuring the data and also creates various charts and reports)
- Telecommunication and media analyze the latest news and trends and also used in generating the content
- Consulting, customer service, tourism, medicine, etc., use the training chatbots for improving the user experience and is widely used these days.

3 Related Work—The Relationships

There are major developments and deficiencies of the most common information System and its concepts and approaches [6]. An information system has proposed that this study can be considered into any supply chain systems that account supply chain constraints regarding the availability of a product while nudging the customers to purchase it [7]. Various researchers have given deep insight toward information security systems. This has contributed by suggesting a framework for the secure software systems [8, 9]. Many researchers have also evaluated the performance of various security intrusion detection methods and its performance analysis using machine and deep learning techniques [10, 11]. The research study provides evidence of classification of the various techniques and about inclinations in systems and future direction on systems [12].

The relationships play an important role in understanding these technologies. Based on the research paper studied using this technology it was found that AI has a relationship between deep learning, blockchain [13]. The open source Voyant tool is used for visual analytics for the purpose. The network diagram was created for the analysis purposes. It was also found that the data and information bifurcate the use of technology [14]. Data or data set/data structure/databases are more aligned for AI, deep learning, as AI uses the history data and also deep into the learning of the data leads to data learning [15]. The data is used for AI and deep Learning for analysis and reporting purposes [16]. Also the data layer is more for the storage purposes.

The information layer AI and blockchain is used in collaboration of various tools. Blockchain makes the availability of information to all the blocks and used as information to complete the analyzing in various businesses [17].

3.1 Deep Learning

Deep learning (DL) a subset of machine learning and AI is a new era in the area of machine learning which applies to an idea of AI [18]. The emphasis on the neural networks where it takes its own decision. Deep works in the concept of layers which emphasize neural networks. It is based on gaining knowledge based on the experience and experiment [19]. It is also known as hierarchical learning or deep structural learning as shown in the network diagram as shown in Fig. 2.

The relationship between all three has become stronger. This will provide help in data security and information security [20]. It is based on how you approach it. The use of AI and deep learning and blockchain has become popular these days in business in global scenarios [21]. The three are connected with each other through data and information.

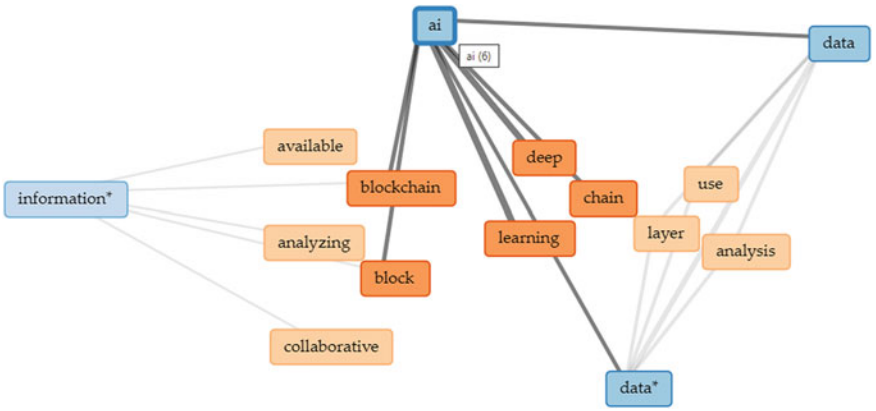


Fig. 2 Deep learning with blockchain and AI

The collaborations AI, deep learning and blockchain uses in the Content-based Filtering [22]. It matches knowledge of a single user with the knowledge of available items. The companies started using the user data in blockchain technology. A blockchain is an immutable, open, append-only transaction log replicated among a network of nodes [23]. One node is connected to the other and assuring through proof-of-work as it gets acknowledged by other participants.

3.2 Flow of Data Between Three Technologies

The above discussed relationships between blockchain, deep learning and AI used for the secured system [22]. The data input enters into the AI system. AI uses the first layer as a transparent layer where the genuinity of the user was checked. All the non-genuine users failed to enter into the system. The filtered data will be the input for deep learning [7] flow. This works on hierarchical study, which further confirms the genuine user. The data transaction verifies and acknowledges the same. The blockchain starts using the chains for the connection and authorization [21].

4 Data Analysis and Interpretation

The world cloud depicts the use of bigger words as compared to small words. The most occurrence of the words discussed in Fig. 3, where data, chain, deep learning, blockchain, AI, information, etc. based on the popularity of different words in the

Fig. 3 Word cloud

document. The use of the words and the network diagram will help us to create a flow chart.

This proposed model will finally help in the implementation of the project as shown in figure defining in Figs. 4 and 5. Figure 4 discusses the involvement of information in AI, blockchain and deep learning. Figure 5 discusses the involvement of data in AI, blockchain and deep learning.

Figure 4 discusses that the information is dependent on blockchain directly and connected to AI indirectly through blockchain and further AI is linked to deep learning and is directly linked to the data. Figure 5 discusses the data connection which is directly connected to AI and deep learning. Further the AI is connected to blockchain and finally linked to the information.

The use of these technologies can be interpreted in Table 1.

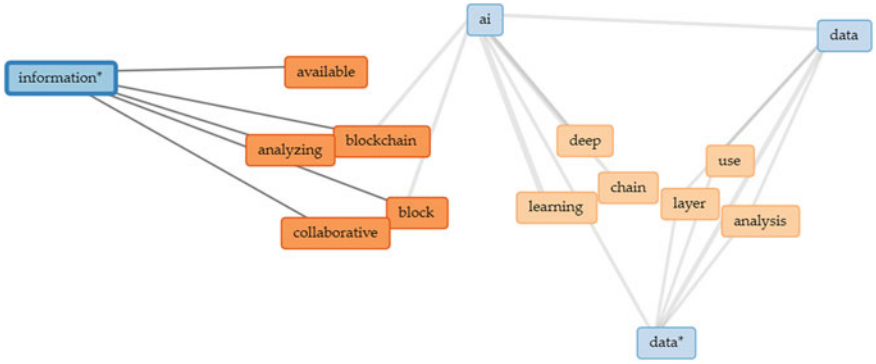


Fig. 4 Information based connections

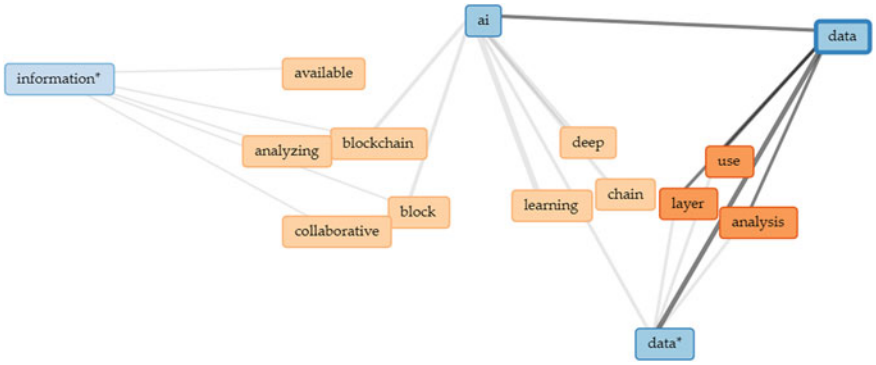


Fig. 5 Data based connections

Table 1 Literature review

S.No	Details	Description	References
1	AI in security	Captcha is used for implementation of AI. Testing in security and use of AI for this	[24, 25]
2	AI and deep learning in security	The implementation discussed the security of the models. The data privacy under the notion of secure and private AI model was discussed	[6, 26, 27]
3	AI and blockchain in security	Both will work together to enhance the security. This can be used as an operational parameter between **security**, decentralization, performance and many others	[26, 28]
4	Blockchain and deep learning in security	Patient are used as a block and they are linked to their wallet. The details are entered captured by the patients only	[29]
5	AI, blockchain and deep learning in security	The combination of it was discussed	[21]

Source Prepared by author

5 Proposed Model

In the normal working with AI, blockchain and deep learning in security will not only enhance the security but also automate the process to a great extent. This can not only be applied as model but also used to automate the process and further used as a machine learning system where the extensive parameters may be incorporated. Based on the detailed study by various researchers the follow of process can be drawn in Fig. 6 and as follows:

The model is proposed using the data and information layer. The information layer will be using the summarization, analyzing and collaboration of the data using blockchain and AI. The data layer enables deep learning and AI using analysis and its use. The approach using data and information in different layers is much secure. This mitigates the risk and increases the trust and security in the system. The privacy/personal data threats have to be taken care of. The model discussed in this paper is the major concern for the protection and the storage of the data. The collaborative system designed will be based on the three futuristic techniques, i.e., AI, blockchain and deep learning.

The model has been divided and studied in six stages—Stage 1: AI; Stage 2: Deep learning; Stage 3: Machine Learning; Stage 4: Blockchain; Stage 5: Clustering; and Stage 6: Mainframe.

Stages 1–3 works under the Data layer, wherein stages 4–6 works for the information layer.

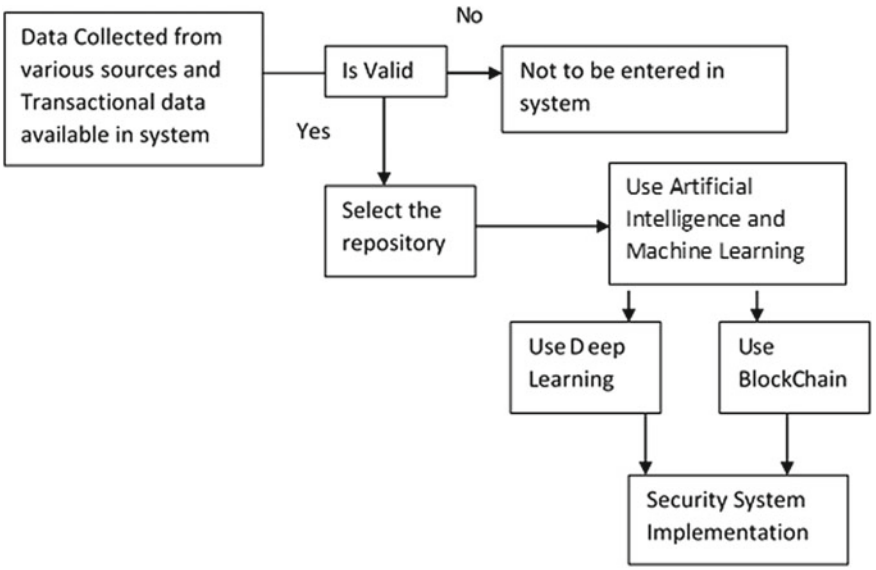


Fig. 6 Flowchart of security system

Stage 1, Stage 2 and Stage 3 use the data. It uses and analyzes the data and works confidently in AI and deep learning and also approaches the chain. The input is taken for the data and from different sources. This data is in the form of structured and unstructured data and used for the item of interest. Which can be used for AI in future for creating patterns and will be used satisfactorily. The data flows into the data layer. AI approach is used to check for any robotic interference, as discussed in our previous work. The filters are considered for making the pattern. The stage AI gathers the data and passes on to deep learning, which considers the data and tries to get different filters for the consideration of categories. The genuine data will be passed on for deep learning. This stage will filter the pattern as per the requirement and based on the filter the pattern has been decided. The AI further connects to the blockchain where the information is required while doing the analysis, summarizing and collaborating the information. The left data is passed to stage 4 for blockchain formation. As like in Bitcoin, blocks of users are created. This helps to accommodate all data without revealing an individual's identity as shown in Fig. 7. In this whole model the security will take place and this is used as an information layer.

6 Conclusion and Future

This paper presents AI, machine learning and deep learning in detail for the implementation of blockchain technology. This platform will use the blockchain technology for a secured purpose. Initially, the paper started with the use of AI, ML and

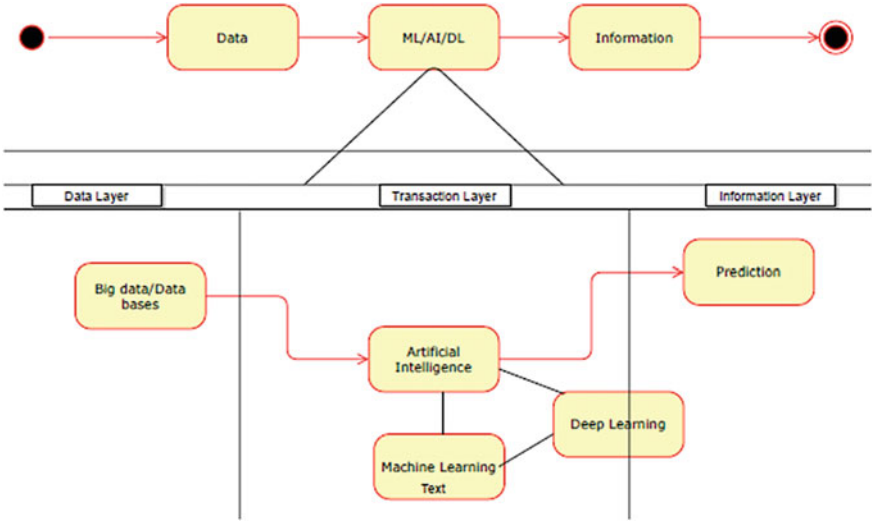


Fig. 7 Proposed model

DL in the application of effective use of technology in businesses. The use of data and its application using various technologies gives the advantage in future implications of technology in businesses. Data science implements the effective use of data and also connects its related components. The relationship between AI, ML and DL will provide a new edge to the business. The data and its validation is the basis of this paper by justifying the relationship between AI, DL and ML by network model. The visual analytics will give you an idea how the things are working by using a network diagram. All these latest technologies/approaches used to design and used to recommend a model to develop a secured collaborative system. The study will further discuss while conducting the exploratory study in future. The challenge for the proposed model is the trust on the dataset selected or used. The authenticity and the reliability of data used in the system. The AI application will churn only the relevant and genuine data and pass it to the next level, that will be used in deep learning and blockchain implementation. AI and blockchain together can be applied to the system where the authenticated data can be further used. In respect to industries where privacy of data is important, this technology will come in handy. With both the technologies currently trending in the industry, the combination of these two technologies promises to provide greater benefits in the coming future.

With all the emerging technologies being used in combination will provide a different edge to the business. These technologies always help the companies to achieve greater benefits in the coming future. This has also redefined the issue of the security and provided the solution for them for the future solution.

References

1. M. Swan, Blockchain for business: next-generation enterprise artificial intelligence systems, in *Advances in Computers*, vol. 111 (Elsevier, 2018), pp. 121–162
2. S.K. Singh, S. Rathore, J.H. Park, Blockiotintelligence: a blockchain-enabled intelligent IoT architecture with artificial intelligence. Futur. Gener. Comput. Syst. **110**, 721–743 (2020)
3. M. Arora, A. Datta, Artificial intelligence and consumer behaviour in marketing. Trends Market. Contemp. Issues 42–47 (2018)
4. M. Arora, A. Arora, Digital information tracking framework using BlockChain. J. Supply Chain Manag. Syst. (UGC Approved) **7(2)**, 2277–1387 (2018)
5. W. Lotter, G. Kreiman, D. Cox, Deep predictive coding networks for video prediction and unsupervised learning (2016). arXiv:1605.08104
6. H. Bae, J. Jang, D. Jung, H. Jang, H. Ha, S. Yoon, Security and privacy issues in deep learning (2018). arXiv:1807.11655
7. Y. Zhang, M. Jiang, X. Chen, Y. Yu, Q. Zhou, Modeling of the filtered drag force in gas–solid flows via a deep learning approach. Chem. Eng. Sci. **225**, 115835 (2020)
8. F. Ahmadi, Sonia, G. Gupta, S.R. Zahra, P. Baglat, P. Thakur, Multi-factor biometric authentication approach for fog computing to ensure security perspective, in *8th International Conference on Computing for Sustainable Global Development (INDIACom)* (2021), pp. 172–176
9. A. Singhal, Development of agile security framework using a hybrid technique for requirements elicitation, in *International Conference on Advances in Computing, Communication and Control* (Springer, Berlin, 2011), pp. 178–188
10. Sonia, A. Alsharef, P. Jain, M. Arora, S.R. Zahra, G. Gupta, Cache memory: an analysis on performance issues, in *8th International Conference on Computing for Sustainable Global Development (INDIACom)* (2021), pp. 184–188
11. P. Tahiri, S. Sonia, P. Jain, G. Gupta, W. Salehi, S. Tajjour, An estimation of machine learning approaches for intrusion detection system, in *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* (IEEE, 2021), pp. 343–348
12. H.D. Park, K.H. Kim, Y.II. Choi, K.J. Kim, A literature review and classification of systems research. Expert Syst. Appl. **39**(1), 10059–10072 (2012)
13. S. Saberi, M. Kouhizadeh, J. Sarkis, L. Shen, Blockchain technology and its relationships to sustainable supply chain management. Int. J. Prod. Res. **57**(7), 2117–2135 (2019)
14. K. Salah, M.H.U. Rehman, N. Nizamuddin, A. Al-Fuqaha, Blockchain for AI: review and open research challenges. IEEE Access **7**, 10127–10149 (2019)
15. M. AlQuraishi, ProteinNet: a standardized data set for machine learning of protein structure. BMC Bioinform. **20**(1), 1–10 (2019)
16. M. Niranjanamurthy, B.N. Nithya, S. Jagannatha, Analysis of blockchain technology: pros, cons and SWOT. Clust. Comput. **22**(6), 14743–14757 (2019)
17. M. Nagendran, Y. Chen, C.A. Lovejoy, A. C. Gordon, M. Komorowski, H. Harvey, E.J. Topol, J.P. Ioannidis, G.S. Collins, M. Maruthappu, Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. Bmj, **368** (2020)
18. L. Zhang, J. Tan, D. Han, H. Zhu, From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug Discovery Today **22**(11), 1680–1685 (2017)
19. D. Wu, Supplier selection: a hybrid model using DEA, decision tree and neural network. Expert Syst. Appl. **36**(5), 9105–9112 (2009)
20. K. Parsons, A. McCormac, M. Butavicius, M. Pattinson, C. Jerram, Determining employee awareness using the human aspects of information security questionnaire (HAIS-Q). Comput. Secur. **42**, 165–176 (2014)
21. M. Arora, A.B. Chopra, V.S. Dixit, An approach to secure collaborative recommender system using artificial intelligence, deep learning, and blockchain, in *Intelligent Communication, Control and Devices. Advances in Intelligent Systems and Computing,* ed by S. Choudhury, R. Mishra, R. Mishra, A. Kumar, vol. 989 (Springer, Singapore, 2020). https://doi.org/10.1007/978-981-13-8618-3_51

22. K. Chung, H. Yoo, D. Choe, H. Jung, Blockchain network based topic mining process for cognitive manufacturing. Wireless Pers. Commun. **105**(2), 583–597 (2019)
23. R. Frey, D. Worner, A. Ilic, Collaborative filtering on the blockchain: a secure system for e-commerce, in *Twenty-Second Americas Conference on Information Systems*, San Diego. https://www.researchgate.net/publication/306119856_Collaborative_Filtering_on_the_Blockchain_A_Secure_System_for_e-Commerce
24. L. Von Ahn, M. Blum, N.J. Hopper, J. Langford, CAPTCHA: using hard AI problems for security, in *International Conference on the Theory and Applications of Cryptographic Techniques* (Springer, Berlin, 2003), pp. 294–311
25. J. Bozic, F. Wotawa, Planning the attack! Or how to use ai in security testing? in *IWAISe: First International Workshop on Artificial Intelligence in Security*, vol. 50 (2017)
26. T.N. Dinh, M.T. Thai, AI and blockchain: a disruptive integration. Computer **51**(9), 48–53 (2018)
27. P.J. Taylor, T. Dargahi, A. Dehghantanha, R.M. Parizi, K.K.R. Choo, A systematic literature review of blockchain cyber security. Digital Commun. Netw. **6**(2), 147–156 (2020)
28. T. Marwala, B. Xing, Blockchain and artificial intelligence (2018). arXiv:1802.04451
29. P. Bhattacharya, S. Tanwar, U. Bodke, S. Tyagi, N. Kumar, Bindaas: blockchain-based deep-learning as-a-service in healthcare 4.0 applications. IEEE Trans. Netw. Sci. Eng. (2019)

Security Analysis and Deployment Measurement of Transport Layer Security Protocol

Marwah Yaseen, Mohammed B. M. Kamel, and Peter Ligeti

Abstract The Transport Layer Security (TLS) protocol is intended to provide secure communication over insecure computer networks, such as the Internet [3]. Many studies covered the verification of security protocols. The motivation comes from the fact that many security issues may not be detected during the testing phases, as some of the issues happened in the presence of a malicious adversary only [4]. This paper focuses on verifying the security of Transport Layer Security (TLS) protocol using the ProVerif verification tool, a fully automatic tool for formal analysis of security protocols using Horn clauses. We intensively studied the TLS protocol and outlined the security issues on its various versions and how these issues have been addressed in the later releases. This analysis will give the ability to propose an abstract model of it to formalize the achieved understanding of TLS. Then, the ProVerif tool will be used to analyze and verify the security and privacy properties of the TLS protocol. Finally, based on the analysis some modifications to the TLS protocol description will be presented and analyzed, and then compared with the original protocol. In addition to that, we monitored the use of the different versions of TLS by scanning a large range of popular domains and analyzing the adopted TLS versions. We showed that TLS 1.3 is deployed speedily and with no security concerns, compared to TLS 1.2 where the adoption took more than five years. Our analysis shows that 52% of the scanned domains support TLS 1.3. Furthermore, we categorized the data according to

This research has been partially supported by Application Domain Specific Highly Reliable IT Solutions project which has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the Thematic Excellence Programme TKP2020-NKA-06 (National Challenges Subprogramme) funding scheme. We would like to thank Tempus Public Foundation and the Ministry of Higher Education in Iraq for the financial support.

M. Yaseen (✉) · M. B. M. Kamel · P. Ligeti
Eötvös Loránd University, Budapest, Hungary
e-mail: mkamel@inf.elte.hu

P. Ligeti
e-mail: hs-furtwangen.deturul@cs.elte.hu

M. B. M. Kamel
Hochschule Furtwangen University, Furtwangen, Germany

University of Kufa, Najaf, Iraq

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_55

the main services categories, such as the educational domains, informative domains, general domains and so on, to analyze the distribution of TLS versions adoption over these categories. We showed that the ".org" domains have a high TLS 1.3 deployment rate and the ".edu" domains have a low TLS 1.3 deployment rate among the scanned domains.

Keywords TLS 1.3 · TLS security · Forward secrecy · ProVerif · TLS 1.3 deployment

1 Introduction

Transport Layer Security (TLS) aims to provide secure communications on the World Wide Web. It is the most commonly used authentication secure channel protocol that protects the communications of billions of users [6]. It verifies the certificates for authentication and encrypts data with keys for confidentiality. Furthermore, it ensures integrity by digesting the data. It was released by Netscape Communications in 1995 as Secure Sockets Layer (SSL). Transport layer in the Open System Interconnection (OSI) model layers is responsible for enabling the end-to-end communication over a network [13]. TLS protocol works on the transport layer, as illustrated in Fig. 1. Many studies covered the verification of security protocols. The motivation comes from the fact that many security issues may not be detected during the testing phases, as some of the issues happen in the presence of a malicious adversary only [4].

Netscape Communications developed the Secure Sockets Layer (SSL) to allow secure access of a browser to a web server. The first SSL version (SSL 1.0) has some problems in the protection of credit card transactions on the web. The next versions of SSL went through several development stages and then it was converted to TLS protocol in order to make it more secure. Figure 2 illustrates the time line of evolution of TLS. TLS protocol has been exposed to many attempts to hack, so it developed over the time with several upgrades and each version adds new features

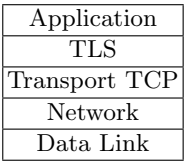


Fig. 1 TCP TLS Temp

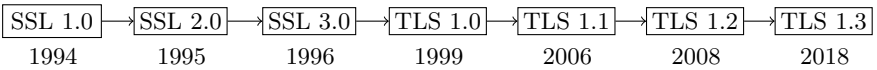


Fig. 2 TLS Development Time Line

and continues to be supported for backwards compatibility. The last one was named TLS 1.3, in 2018.

TLS 1.0 was first defined in 1999 as an upgrade of SSL Version 3.0 but does not have the capability of interoperability. The protocol name changed from SSL to TLS because TLS works not just with sockets but over any bidirectional stream of bytes [15]. In 2006, TLS 1.1 was defined and provided a few major changes to TLS 1.0, primarily to prevent specific types of attacks on the Cipher Block Chaining (CBC) mode [16]. TLS 1.2 was defined in 2008 and it introduced many cryptographic enhancements, especially enabling the use of SHA-2 hash function for computing hash, MAC, and Pseudo Random Function (PRF). It also enables the Authenticated Encryption with Associated Data (AEAD) cipher suites. In addition to that, it allows the client and server to choose hashes and signature algorithms, and many other enhancements. TLS 1.3 is the last introduced version. It was defined in 2015 as a working draft and has not yet completed its development with many changes on TLS 1.1 and 1.2.

The organization of this paper is as follows: Section 2 explains briefly the TLS protocol and Sect. 3 gives some TLS 1.3 properties. TLS 1.2 and TLS 1.3 verification using ProVerif is provided in Sect. 6. In Sect. 8, tracking the TLS 1.3 deployment in a list of scanned domains from the Alexa Top1M domains list is studied. Finally, conclusions and future directions are provided in Sect. 9.

2 Transport Layer Security Protocol

In this section, we briefly explain the latest two versions of TLS: TLS 1.2 and TLS 1.3. Handshake protocol and the record protocol are the two most important protocols that construct the TLS protocol [7]. The TLS handshake protocol is responsible for the authentication between a client and a server and also for setting a key. The TLS record protocol is responsible for providing confidentiality and verifying the integrity of the application data.

The handshake in TLS 1.3 has one round-trip, while previous versions of TLS/SSL required two round-trips. The TLS has six stages. The first two stages are the key exchange deriving the internally used handshake traffic secrets. The other phases are for the authentication and the client will authenticate the server. The following steps show how the TLS protocol works:

1. **Client:** It starts by sending the *clientHello* message consisting of cryptographic information such as supported protocols and supported cipher suites.
2. **Server:** The server will send *HelloServer* containing the selected algorithms offered by the client. The server should answer with a *HelloRetryRequest* message if the server selects a protocol and the client did not provide a supported version in the initial ClientHello.
3. **Server:** Any extensions that are not required to establish the server handshake traffic key are sent and encrypted in the encrypted extensions. Then the server

requests a public-key-based client authentication by sending a *CertificateRequest* message carrying its own public key. After that it will compute the authentication value by signing the session hash with the private key, then sending it in a *ServerCertVfy* message.

4. **Client:** The client upon receiving the messages verifies the SCV signature and the ServerFinished signature and now the client is able to compute client/server application traffic secret and Exporter Master Secret.
5. **Server:** If the server has requested client authentication, the client will begin by sending its digital certificate (with its public key) in the *ClientCertificate* message.
6. Here the client derives the final key and signs the session hash and sends it with the *ClientFinished* message.

These steps are illustrated in Fig. 3, where the client sends the ClientHello to the server and the server selects the cipher suite and key agreement algorithm. Since it already has the client key share, then it is ready to generate the key. The server sends the ServerHello, its key share, the encrypted certificate, and the server Finished message to the client. After receiving all the transmitted information by the client, it generates the keys using the key share, checks the certificate, and the client Finished message. Then it is ready to send the encrypted HTTP request, by only one round-trip, while in TLS 1.2, by two round-trips.

3 TLS 1.3 Properties

Figure 3 shows one of the main overcomes of TLS 1.3 over TLS 1.2, that is, its handshake which is shorter in TLS 1.3. One of the basic forms of TLS 1.3 uses Diffie-Hellman ephemeral (DHE) for handshake usually using the elliptic curve (ECDHE), which is a key agreement protocol, and the application data encryption is done by using one of the AEAD schemes such as AES-GCM.

Handshake.

From the early drafts of TLS 1.3, the 1-RTT essential structure has remained stable [3]. TLS 1.3 handshake has two main differences from the TLS 1.2 handshake:

1. The execution of the key exchange and negotiation protocol is done at the same time (in parallel), then the client can send the application data immediately in its second message, i.e., after one round-trip, which makes it 1-RTT. On the other side, the client in TLS 1.2 has to wait for two messages from the server.
2. Some of the problematic properties in TLS 1.2 were removed in TLS 1.3, like removing the RSA during key exchange and the weak encryption schemes such as RC4, TripleDES, and AES-CBC. Furthermore, it illuminates the renegotiation, where TLS 1.3 requires strong standardized Diffie-Hellman group negotiation. And to prevent some attacks like Triple Handshake, TLS 1.3 binds session keys to the handshake log.

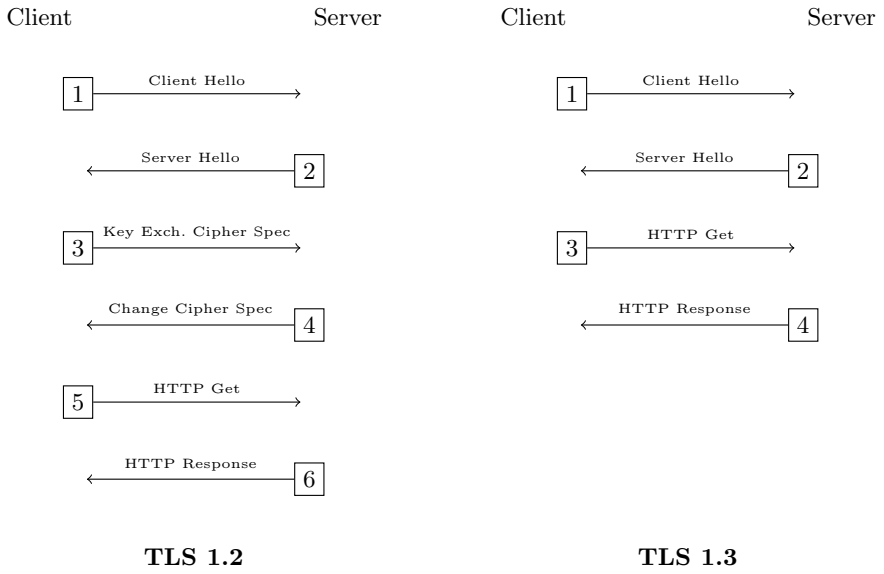


Fig. 3 TLS 1.2 and TLS 1.3 handshake protocols

3.1 Finished Message and Pre-shared Key Extension

The Finished message is the final message in the authentication phase. It is used as an essential part to provide authentication of the handshake and to compute keys. The recipients of Finished messages must verify that the contents are correct and if incorrect must terminate the connection with a *decrypt-error* alert. When receiving the Finished messages, it is necessary to verify whether their contents are correct or not. If it is not, the connection must be terminated with a decrypt-error alert. When the client has sent its Finished message and has received Finished message from the server, it is possible to begin sending and receiving application data over this connection [17]. When the client is sending 0-RTT data, it may be permitted to send application data before receiving the server's Finished message.

The pre-shared key extension is used by the client and the server to agree on the pre-shared key to be used in a given handshake in association with the PSK key establishment.

4 Key Authentication

Key authentication is a property that guarantees only the identified parties can know the agreed secret key. After completing the handshake, the client and the server can update their sending cryptographic keys using *KeyUpdate* handshake message.

KeyUpdate handshake message indicates that the client or the server is updating its sending cryptographic keys. The resulting traffic keys are called client application traffic secret ats_c and server application traffic secret ats_s that have been generated from the previous sending cryptographic keys. If the client terminates when in an honest session, the server accepts the honest client's session and they share the same parameters, including the keys ats_c, ats_s, and *exporter master secret ems*, as well as all messages sent in the protocol before the server Finished message. We can't make any claims about the client's Finished message because it hasn't been received by the server yet. Also, we can't make any claims about PSK because it depends on the client's Finished message.

5 Forward Secrecy

Secrecy is achieved if a message is sent between an honest client and an honest server over a session. Then the message is kept confidential from the attackers who cannot break the cryptographic constructions used in the session. Forward secrecy is achieved if the secrecy holds even if the keys (sk_C, pk_C, PSK) are given to the adversary after the session has been completed and the session keys k_c, k_s are deleted by the client and the server. Following is a forward secrecy scenario between a client and a server:

1. The client as a sender and server as a receiver generate a pair of public and private keys and the symmetric pre-shared keys (PSK).
2. The client uses Diffie-Hellman key exchange to securely agree on an ephemeral session key and authenticate one another using the keys from step 1.
3. The client sends an encrypted message to the server and the server decrypts the message and they both use the keys agreed in step 2.
4. For each new message sent between them (switching client's/server's roles as sender/receiver as required), the steps are repeated starting from step 2 as step 1 is executed only once.

TLS 1.3 generates a one-time key to be used in the current session using the ephemeral Diffie-Hellman (DHE) key exchange protocol. When the session is terminated, the key is discarded. Attackers cannot crack the ciphertext later on in a reasonable time, as it is a very difficult process that could require a high computing power with too long process duration [19].

External PSK and resumption PSK bootstrap a long-term shared secret into a specific collection of short-term session keys for each link. It is possible that this secret was established during a previous handshake. These session keys would be forward secret if PSK with DHE key establishment is used. The resumption PSK was created in such a way that the resumption master secret computed by connection N and used to form connection $N + 1$ is kept separate from the traffic keys used by connection N, ensuring forward secrecy between the connections. Furthermore, multiple tickets are associated with different keys if they are established on the same

connection, so a compromise of the PSK associated with one ticket does not result in the compromise of connections established with PSKs associated with other tickets. If tickets are kept in a database (and therefore can be deleted), rather than being self-encrypted.

5.1 Forward Secrecy Applications

Many large internet information providers consider forward secrecy as an important security feature for their services.

- Google provides by default the TLS forward secrecy to users of its Gmail service, Google Docs service, and encrypted search services since late 2011 [11].
- Twitter officially states the providing of TLS forward secrecy to its users since late 2013.
- In an investigation into email encryption, Facebook reported in Facebook Notes that as of May 2014, 74% of hosts that support StartTLS (a protocol command used by the email client to inform the email server that the client wants to upgrade from an insecure connection to a secure one by using TLS or SSL) also provide forward secrecy.
- TLS 1.3 provides support only for ciphers with forward secrecy. Qualys's global dashboard (for monitoring the quality of SSL/TLS supports over time across 150,000 SSL- and TLS-enabled websites) surveyed that 99% of web servers support some form of forward secrecy, and 66.3% servers use forward secrecy with most browsers.
- At Apple Worldwide Developers Conference (WWDC) 2016, Apple announced that App Transport Security (ATS) would be needed for all iOS applications, where ATS is a feature that requires using only encryption ciphers that provide forward secrecy. In January 2017, ATS became mandatory for iOS applications.

1-RTT Connection

A typical connection in TLS 1.3 is described above for TLS 1.3 1-RTT Draft 18. The protocol flow goes as follows:

1. The *ClientHello* message containing a nonce n_C and an offer C that lists the supported versions, groups, and algorithms is sent to the server. The client may include a Diffie-Hellman key share g^x for each supported group.
2. The server chooses a group G that the client gave a public value, then it sends the *ServerHello* to the client that contains a nonce n_S, mode S, and key share g^y. The *HelloRetryRequest* is applied with a key share $g^{x'}$ for the server's preferred group G'. Therefore, the handshake requires two round-trips in this case. And the negotiation is completed. After completion of the negotiation, the handshake encryption keys are derived from $g^{x'y}$ to encrypt the handshake messages. Also, the client and the server generate the MAC keys k_c^m and k_s^m to be used in the Finished messages.

3. Then, the server sends up to 5 encrypted messages:
   - Protocol extensions that were not sent in the ServerHello.
   - An optional client certificate request
   - A certificate of the server's public key
   - *CertVerify* contains a signature with server's private key sk_S
   - The Finished message that contains a Message MAC.
4. The server generates k_c and k_s (the 1-RTT traffic keys) and can use k_s for encrypting the application data to the client.
5. After that, the client start verifying the received certificate, the signature, and the MAC. Then it sends a second flight that contains a Certificate, signature CertVerify, and a mandatory Finished message with a MAC. After this, it can send its own application data encrypted by k_c.
6. After the handshake is completed, all the session parameters are placed in both client and server local session databases (C, S).

The 1-RTT handshake generates an *ems* to bind authentication credentials to the TLS channel. In addition, it generates a *resumption master secret* psk' as a pre-shared key for the future TLS connections between the client and the server.

6 TLS 1.2 and TLS 1.3 Security Analysis

Considering the wide deployment of TLS, there are many successful attacks against it. There were many practical attacks on different versions of TLS that received significant attention [7]. These attacks include:

- Weaknesses of the used cryptography primitives, for example, RC4 weaknesses.
- TLS protocol design errors, such as Lucky 13 Attack, Browser Exploit Against SSL/TLS (BEAST) Attack, Triple Handshake Attack, Padding Oracle On Downgraded Legacy Encryption (POODLE) Attack, and Logjam Attack
- Faulty implementation. Some examples include Heartbleed Exploit and State Machine Attacks (SMACK)

TLS 1.2 has many improvements over the previous versions such as the use of more secure hash algorithms. But as new vulnerabilities raised, TLS 1.2 is under many security risks. One of the vulnerabilities in TLS 1.2 is continuing the support for CBC mode of encryption. In CBC, man-in-the-middle attacks (MITM) can take place on users' encrypted web and VPN sessions. Although TLS 1.3 does not support CBC mode of encryption, Lee et al. [12] concluded that the TLS 1.3 sessions can be downgraded by exploiting the downgrade attacks which then allows the attacker to perform the MITM attack.

Many studies [2, 8, 10, 14] went through analyzing the security properties of the different versions of TLS and modified some versions of the protocol. Because keys have no strict separation in TLS 1.2 and earlier versions, session keys were also used

to encrypt messages within the handshake protocol. TLS 1.3 is designed mainly to overcome the security problems of TLS 1.2 with a completely new functional security design. So, while there were many vulnerabilities discovered in TLS 1.2, TLS 1.3 was designed to support only state-of-the-art algorithms with no known vulnerabilities to solve those issues.

TLS 1.3 offers options of both highly secure, and also fast but with security compromises. These options show the main compromise between security and efficiency [1]. The main features that have been removed from TLS 1.3 are Static RSA handshake, CBC MtE modes, RC4, SHA1, and MD5. The main changes that are made on TLS 1.2 to produce TLS 1.3 are as follows:

- In some handshake messages the encryption is done with an intermediate session key for providing confidentiality to some sensitive data.
- Signing the entire handshake transcript for authentication.
- Including the hashes of handshake messages in many calculations of keys.
- Using different keys to encrypt handshake messages and application data.
- The algorithms of encryption and authentication are combined by using AEAD encryption.
- Reducing latency by providing fewer message flows for handshakes.

7 TLS Security Verification

In this section, we verify the security specifications of TLS 1.2 and 1.3 using ProVerif and analyze the two versions of the protocol in the security model.

7.1 ProVerif Verifier

ProVerif [5] is a tool used to automatically analyze and verify the security of cryptographic protocols in the formal model (the so-called Dolev-Yao model). It has been developed by Bruno Blanchet and supports a wide range of cryptographic primitives protocols including symmetric and asymmetric encryption; digital signatures; hash functions; bit-commitment; and non-interactive zero-knowledge proofs.

The protocol needs to be represented in an abstract way using Horn clauses to be verified. Horn clauses are set of facts $F_1, \ldots, F_n, F$ in the form of $F_1 \wedge \ldots \wedge F_n \Rightarrow F$. This representation introduces abstractions, where pi calculus is a mathematical formalism used to describe and analyze the properties of a computation.

- **Secrecy query:**
  query event(ClientAEKeyLeaked(TLS12,cr,sr,psk,p)) ⟹
  event(ServerChoosesKEX(cr,sr,p,TLS12,DHE(WeakDH))) ||
  event(ServerChoosesKEX(cr',sr',p,TLS12,RSA(WeakRSADecryption)))||
  event(ServerChoosesHash(cr,sr,p,TLS12,WeakHash))
  False: Proverif finds a counterexample.
- **Authentication query:**
  query event(ClientFinished(TLS12,cr,sr,NoPSK,p,o,m,ck,sk,cb,ms)) ⟹
  event(ServerFinished(TLS12,cr,sr,NoPSK,p,o,m,ck,sk,cb,ms)) ||
  event(ServerChoosesKEX(cr,sr,p,TLS12,DHE(WeakDH))) ||
  event(ServerChoosesKEX(cr',sr',p,TLS12,RSA(WeakRSADecryption)))||
  event(ServerChoosesHash(cr',sr',p,TLS12,WeakHash)).
  False Proverif finds a counterexample.
- **appData query:**
  query attacker(m_c(TLS12,cr,sr,p,psk)) ⟹
  event(WeakOrCompromisedKey(p)) ||
  (event(PostSessionCompromisedKey(p)) &&
  event(ServerChoosesKEX(cr,sr,p,TLS12,RSA(ra)))) ||
  event(ServerChoosesAE(cr,sr,p,TLS12,WeakAE)) ||
  event(ServerChoosesKEX(cr,sr,p,TLS12,DHE(WeakDH))) ||
  event(ServerChoosesKEX(cr',sr',p,TLS12,RSA(WeakRSADecryption)))||
  event(ServerChoosesHash(cr',sr',p,TLS12,WeakHash))
  True: Proverif finds no counterexample.

Fig. 4 TLS 1.2 ProVerif verification results

7.2 TLS 1.2 Verification

The implementation of TLS 1.2 in ProVerif provides several queries with events to be tested. The verification summary of the execution shows all the queries executions, with true and false responses. In some queries, weak algorithms are given such that the primitives are completely broke [3]. Figure 4 shows the secrecy, authentication, application data queries, and their corresponding output.

7.3 TLS 1.3 Verification

The ProVerif implementation of TLS 1.3 models the client and server processes (shown in Fig. 3). We are going to verify TLS 1.3 Draft-18 1-RTT protocol. The ProVerif model treats each message, for example, the *ClientHello* message, as a symbolic constructor and a pattern match so that it is assumed that the message serialization and the parsing are correct. That means it is not possible to find an attack that uses parsing ambiguities or bugs.

- **Secrecy Query:**
  Query event(ClientAEKeyLeaked(TLS13,cr_20,sr_14,psk_18,p_9))
  $\Longrightarrow$ (event(WeakOrCompromisedKey(p_9)) && (psk_18 = NoPSK
  || event(CompromisedPreSharedKey(psk_18)))) || event(ServerChoosesKEX
  (cr_20,sr_14,p_9,TLS13,DHE_13(WeakDH,e_1))) **is false**.
- **Authentication Query:**
  Query inj-event(ClientFinished(TLS13,cr_20,sr_14,psk_18,p_9,o_16,m_13,ck_12,sk_9,
  cb_5,ms_11)) $\Longrightarrow$ inj-event(PreServerFinished(TLS13,cr_20,sr_14,psk_18,p_9,o_16,
  m_13,ck_12,sk_9,cb_5)) || (event(WeakOrCompromisedKey(p_9)) &&
  (psk_18 = NoPSK || event(CompromisedPreSharedKey(psk_18)))) ||
  event(ServerChoosesKEX(cr'_5,sr'_3,p_9,TLS12,RSA(WeakRSADecryption)))
  is false.
- **appData Secrecy query:**
  Query attacker_p1(m_c(TLS13,cr_20,sr_14,p_9,psk_18)) $\Longrightarrow$
  (event(WeakOrCompromisedKey(p_9)) && (psk_18 = NoPSK ||
  event(CompromisedPreSharedKey(psk_18)))) ||
  event(ServerChoosesAE(cr_20,sr_14,p_9,TLS13,WeakAE)) ||
  event(ServerChoosesKEX(cr_20,sr_14,p_9,TLS13,DHE_13(WeakDH,e_1))) ||
  event(ServerChoosesKEX(cr'_5,sr'_3,p_9,TLS12,RSA(WeakRSADecryption))) ||
  event(ServerChoosesHash(cr'_5,sr'_3,p_9,TLS13,WeakHash)) **is false**.
- **appData Authentication query:**
  Query event(ServerReceives(TLS13,cr_20,sr_14,psk_18,p_9,n_7,ad_7,m_13)) $\Longrightarrow$
  event(ClientSends(TLS13,cr_20,sr_14,psk_18,p_9,n_7,ad_7,m_13)) ||
  psk_18 = NoPSK || event(CompromisedPreSharedKey(psk_18)) ||
  event(ServerChoosesAE(cr_20,sr_14, p_9,TLS13,WeakAE)) ||
  event(ServerChoosesKEX(cr_20,sr_14,p_9,TLS13,DHE_13(WeakDH,e_1))) ||
  event(ServerChoosesHash(cr'_5,sr'_3,p_9,TLS13,WeakHash)) **is true**.

Fig. 5 TLS 1.3 ProVerif verification results

By verifying TLS 1.3 using ProVerif, it generates a large amount of output. Figure 5 shows the results of the queries, the queries with "true" results are related to the security properties such as forward secrecy and authenticity. The queries with "false" results are related to queries that are expected to fail so that these queries showed that the verification is tight.

From the verification results shown in Fig. 5, ProVerif can prove that the forward secrecy and authentication queries for the application data messages m.

- For **forward secrecy** property, it proves that any message between C and S are secret in the session. This is due to the fact that the private keys of C and S are kept secret until the end of the session and the server chose a strong DH group, hash function, and authenticated encryption algorithm.
- For **authentication** property, it proves for any accepted message by the server over a session with an honest client that this message is a unique message sent by the client over a matching session.

As shown in Fig. 5, the secrecy query failed and ProVerif found a counterexample, where the client encryption key is leaked to the attacker if it can compromise the private key (WeakOrCompromisedKey(p9)). The authentication query failed and the

counterexample is when the attacker can compromise the PSK (CompromisedPre-SharedKey(psk18)). The appData Secrecy and appData Authentication queries gave "false" and "true" responses, respectively.

The ProVerif implementation of TLS 1.3 has TLS 1.2 combined, assuming that the client and the server support both versions. The execution of this model shows that these two versions of TLS may cause some problematic properties to the system. These problematic properties are shown when the server supports the RSA encryption and decryption supported in TLS 1.2. Then ProVerif shows that the RSA decryption can be used by an attacker to break secrecy and authentication properties of TLS 1.3 by forging TLS 1.3 server signatures. So it is not enough to only remove the RSA key transport from TLS 1.3, but it is necessary to disable the use of RSA mode in TLS 1.2 for any server whose certificate may be accepted by a TLS 1.3 client. In addition to that, when the client or the server supports TLS 1.2 weak hash function for signatures, then ProVerif shows that the attacker forges TLS 1.3 signatures using this weakness. Therefore, it is necessary to disable weak hash functions not only in TLS 1.3 but also in all supported versions. Disabling the weak algorithms of TLS 1.2 proves the expected security properties and even when it is combined with TLS 1.3.

8 Tracking the TLS 1.3 Deployment

In this section, we monitor the implementation of the different versions of TLS by scanning a large range of popular domains and analyzing the adopted TLS versions. Furthermore, we categorize the data according to the main services categories, such as the educational domains (.edu), informative domains (.info), and general domains (.com and .net), to analyze the distribution of TLS versions adoption over these categories.

We used for this the public list of domains (Alexa Top1M).[1] Alexa Top1M list is a solid choice of functional websites frequently visited by users [18]. Ralph et al. [9] show that the deployment of TLS 1.3 is increasing. They showed that as of May 2019, 21% of the domains on Alexa Top1M-listed domains deployed TLS 1.3, and it increased to 31% by November 2019. Qualys's global dashboard showed in their monthly scan that as of July 2020, 32% of the 41350 scanned sites deployed TLS 1.3. And as of November 2020, they showed that 41% of the 55957 scanned sites deployed TLS 1.3.

We prepared a sample from the dataset based on hardware availability and removed all the inactive domains. We scanned the data to get the supported version of TLS for each domain. We used the SSL/TLS library in Python[2] for the scan.

Our experiment is done on the updated dataset (as of March 2021) and we found that 52% of the domains in the list support TLS 1.3 as shown in Fig. 6, which is a

[1] https://domains-index.com/downloads/alexas-top-1000000/.

[2] https://docs.python.org/3/library/ssl.html.

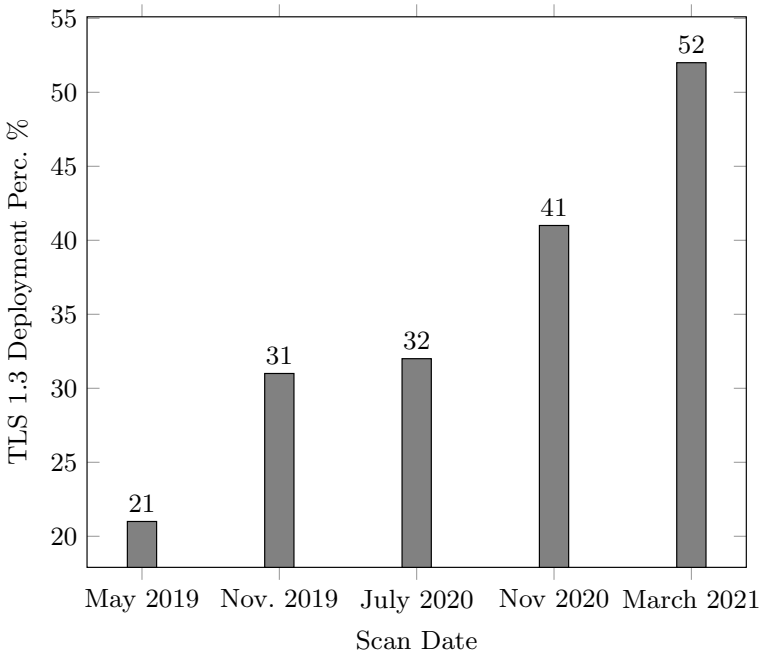


Fig. 6 The deployment of TLS 1.3 based on scanning Alexa Top1M domain list

very vast deployment rate. TLS 1.3 is deployed faster than TLS 1.2, where TLS 1.2 adoption took more than five years and faced many attacks on the earlier versions. This strong deployment proves the stability of TLS 1.3.

Furthermore, we categorized the data into three main categories based on the general services provided to show the distribution of TLS 1.3 deployment for different online services. The categorized groups are as follows: the educational domains with ".edu" domain 1.06%, informative domains with ".info" domain 1.05%, organizational domains with ".org" domain 5.3%, governmental domains with ".gov" domains 0.6%, general domains with ".com", ".net," and all the other domains 92%. Figure 7 shows the distribution of TLS 1.3 deployment among these domain categories.

Our scan analysis result shows that the organizational and informative domains made a high TLS 1.3 deployment rate. Although online education takes an important role due to current pandemic since 2020 all around the world, we observed that the educational domains are still in a low deployment rate of the TLS 1.3.

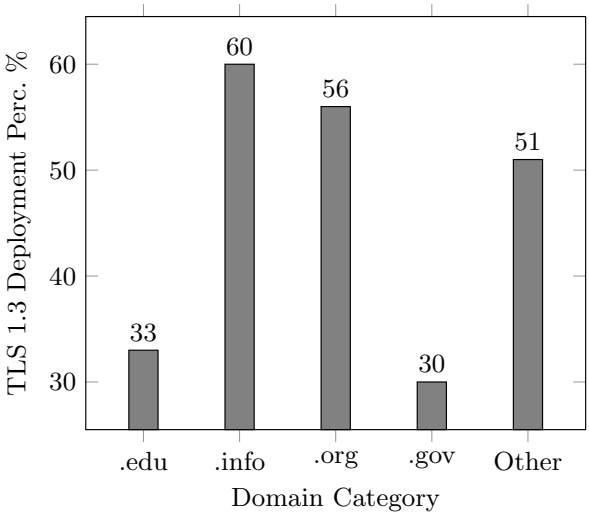


Fig. 7 The deployment percentage of TLS 1.3 for each domain category in Alexa Top1M domain list

9 Conclusions and Future Directions

TLS provides secure communication over insecure computer networks, like the Internet. Many studies covered security issues that may not be detected during the testing phases, where some security issues happen only in the presence of a malicious adversary.

We intensively studied the TLS protocol and outlined the security issues in its recent versions. We analyzed and verified the security and privacy properties of TLS protocol. We used the ProVerif verification tool to verify the security of TLS 1.2 and 1.3 protocols. The verification of both the versions is analyzed and compared. We showed that it is necessary to disable the use of RSA mode in TLS 1.2 for any server whose certificate may be accepted by a TLS 1.3 client. Furthermore, we showed that disabling the weak algorithms of TLS 1.2 proves the expected security properties, even when it is combined with TLS 1.3.

Based on the scanning of a large range of popular domains, we showed that TLS 1.3 is deployed speedily and with no security concerns, it is compared to TLS 1.2 where the adoption took more than five years. Our analysis shows that 52% of the scanned domains supports TLS 1.3. We also categorized the data according to the main services categories and analyzed the distribution of TLS versions adoption over these categories. We found that the ".org" domains have a high TLS 1.3 deployment rate while the ".edu" domains showed a low TLS 1.3 deployment rate among the scanned domains. One future research direction is verifying the recent drafts of TLS 1.3 using different verification tools, such as CryptoVerif and Tamarin prover and finding out the differences of the verification results. Another direction is tracking the TLS 1.3 deployment over domains such as online payment domains that require very high security properties.

References

1. S.B. Alemu, The transformation of TLS from version 1.2 to 1.3 (2020)
2. M. Bellare, P. Rogaway, Entity authentication and key distribution, in *Annual International Cryptology Conference* (Springer, 1993), pp. 232–249
3. K. Bhargavan, B. Blanchet, N. Kobeissi, Verified models and reference implementations for the TLS 1.3 standard candidate, in *IEEE Symposium on Security and Privacy (SP)*, vol. 2017 (IEEE, 2017), pp. 483–502
4. B. Blanchet, Automatic verification of security protocols in the symbolic model: the verifier proverif, in *Foundations of Security Analysis and Design VII* (Springer, 2013), pp. 54–87
5. B. Blanchet et al. ProVerif 2.00: automatic cryptographic protocol verifier, user manual and tutorial, in Version from 2018, 05-16 (2018)
6. C. Brzuska et al., Key-schedule security for the TLS 1.3 standard, in *IACR Cryptol. ePrint Arch. 2021* (2021), p. 467
7. B. Dowling et al., A cryptographic analysis of the TLS 1.3 hand- shake protocol. J. Cryptol. (2020)
8. S. Gajek, A universally composable framework for the analysis of browser-based security protocols, in *International Conference on Provable Security* (Springer, 2008), pp. 283–297
9. R. Holz et al., Tracking the deployment of TLS 1.3 on the web: a story of experimentation and centralization. ACM SIGCOMM Comput. Commun. Rev. **50**(3), 3–15 (2020)
10. J. Jonsson, B.S. Kaliski, On the security of RSA encryption in TLS, in *Annual International Cryptology Conference* (Springer, 2002), pp. 127–142
11. A. Langley, Protecting data for the long term with forward secrecy, in *Google Online Security Blog (November 2011)* (2011), http://googleonlinesecurity.blogspot.co.il/2011/11/protecting-data-for-long-term-with.html
12. S. Lee, Y. Shin, J. Hur, Return of version down- grade attack in the era of TLS 1.3, in *Proceedings of the 16th International Conference on Emerging Networking Experiments and Technologies* (2020), pp. 157–168
13. T. van der Merwe, An analysis of the transport layer security protocol. Ph.D thesis (Royal Holloway, University of London, 2018)
14. P. Morrissey, N.P. Smart, B. Warinschi, A modular security analysis of the TLS handshake protocol, in *International Conference on the Theory and Application of Cryptology and Information Security* (Springer. 2008), pp. 55–73
15. D. Ozden, Analysis of recent attacks on SSL/TLS protocols. MA thesis (Middle East Technical University, 2016)
16. T. Polk, K. McKay, S. Chokhani, Guidelines for the selection, configuration, and use of transport layer security (TLS) implementations. NIST Spec. Public. **800**(52), 32 (2014)
17. E. Rescorla, T. Dierks, The transport layer security (TLS) protocol version 1.3 (2018)
18. Q. Scheitle et al., A long way to the top: Significance, structure, and stability of internet top lists. Proceedings of the Internet Measurement Conference **2018**, 478–493 (2018)
19. Y. Sheffer, R. Holz, P. Saint-Andre, Recommendations for secure use of transport layer security (tls) and datagram transport layer security (dtls), in *RFC 7525* (RFC, 2015)

A Web Application Vulnerability Testing System

Roseline Oluwaseun Ogundokun, Sanjay Misra, Tobe Segun-Owolabi, Abhiram Anand Gulanikar, Akshat Agrawal, and Robertas Damasevicius

Abstract Online apps, as opposed to standard applications, are complex and collaborative. Standard testing methods and tools, however, are not appropriate for testing web applications. This study focuses on developing a web-based web application vulnerability scanner that hybridizes the black-box and white-box pen testing methodologies, thereby reducing scanning costs and improving efficiency while still proving a very low false-positive rate. The system can detect cross-site scripting (XSS), flash injection mixed content, cleartext passwords, and usage of insecure JavaScript libraries. The system had a 99% accuracy for the XSS, mixed content, and cleartext passwords. The system satisfies the vast majority of OWSAP top-ten criteria and provides a detailed actionable scan result that is easily understandable and easy to debug when compared to similar existing systems and other research works.

Keywords Web application · Vulnerability · Cross-site scripting · Black-box pen testing · White-box pen testing

R. O. Ogundokun · T. Segun-Owolabi
Landmark University Omu Aran, Omu-Aran, Kwara, Nigeria
e-mail: ogundokun.roseline@lmu.edu.ng

T. Segun-Owolabi
e-mail: segun-owolabi.tobe@lmu.edu.ng

S. Misra (✉)
Ostfold University College, Halden, Viken, Norway
e-mail: sanjay.misra@hiof.no

A. A. Gulanikar
Birla Institute of Technology, Pilani, Hyderabad, India
e-mail: f20150105h@alumni.bits-pilani.ac.in

A. Agrawal
Amity University, Gurgaon, Hariyana, India

R. Damasevicius
Kaunas University of Technology, Kaunas, Lithuania
e-mail: robertas.damasevicius@ktu.lt

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_56

1 Introduction

A web application has a lot of benefits. Some common benefits of web applications are that it enables access by multiple users to the same application version; web application does not need installation; and web application can also be reached via multiple browsers, therefore promoting versatility [1]. With the advent of web applications in 1999, they have become a very important tool in business, entertainment, and politics, which has enabled the dispersal of information, the simplification of organizational operations, and easier access to information. A 2019 statistic by digital trend stated that every day since January 2018, over one million people have come online for the first time and about 4.39 billion internet users were reported in 2019, an increase of 366 million (9%) [2]. In order to ensure web application security assurance, a web application has to be tested continuously. This is because hackers are becoming smatter as years go by and new vulnerability is being discovered [3].

Testing is one of the most important activities for any application [4, 5]. In order to ensure web application security among small-scale industries that have vital information about customers and clients on the web, this study proposes an automatic web application security testing system that will remove the abstraction of employing an expensive penetration tester and also prevent high-cost continuous application testing and time-wasting in carrying out tests. The system will have the ability to perform both black-box and white-box penetration testing scans which will major in identifying vulnerabilities like cross-site scripting SQL injection, the use of outdated JavaScript libraries, and flash injection, while still able to deliver a very low false-positive rate.

The remaining part of the paper is organized as follows: Sect. 2 discusses the literature review on web applications and related works were also reviewed. Section 3 discusses the materials and methods used in the study. Section 4 discusses the results gotten from the implementation and testing of the system. The paper was concluded in Sect. 5 with a suggestion of future works.

2 Literature Review

In information technology (IT), security is the safeguarding of software or data and IT resources from internal and external threats, both pernicious and incidental. This defense incorporates discovery, counteractive action, and reaction to dangers using security policies, software tools, and IT services. Over the years research has been able to show us the most exploited web vulnerabilities [6]. Figure 1 shows us the most popular vulnerabilities from 2005 to 2009. In the year 2017, for the five most frequent attacks, the relative frequency modifications were small [7]. Cross-site scripting was detected in one-third of all attacks that involved users' accounts ("Web Application Attacks Statistics: 2017 Outcomes and Expectations for 2018," 2017). Other well-known assaults included the capacity to get to the information or SQL injection,

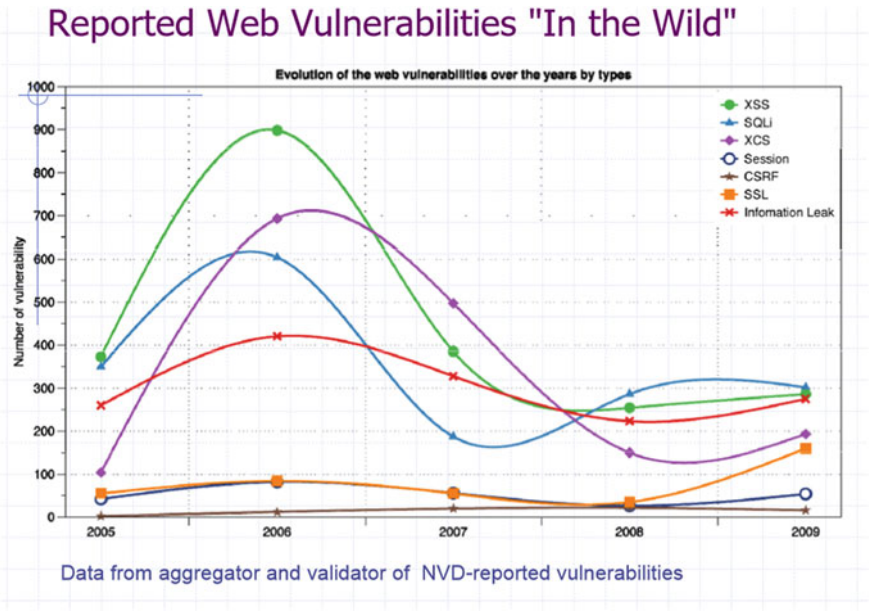


Fig. 1 Vulnerability report between 2005 and 2009 [6]

path traversal, local file inclusion, remote code execution, and OS command execute commands on the server [7].

According to this report higher exploit trends are on security misconfigurations which can render web applications open to a variety of attacks ranging from cross-site scripting (XSS) to SQL and other threats [8] (Fig. 2).

In the course of this study, reviews have been carried out on significant journals related to this study in order to detect and prevent web application vulnerabilities.

CLUSIF [6] came up with a study on web application security. In this paper the focus was majorly on SQL injection, CSRF, and OWASP XSS attacks, and how to identify the causes and preventive mechanisms. The limitation of this paper was that the study was only related to a limited amount of web application vulnerabilities.

Durić [10] proposed the web application penetration testing tool (WAPTT), which scans web applications based on popular SQL injection (SQLI), cross-site scripting (XSS), and buffer overflow (BOF) weaknesses, and have modularity capabilities that enable the end-users to easily extend the tool to suit their requirement in order to improve the detection of SQLI vulnerabilities. Although the system has a modular capability, it lacks detection abilities of other modern critical issues which can make SQLI a step more achievable like mixed (HTTP in HTTPS) material and obsolete/insecure repositories.

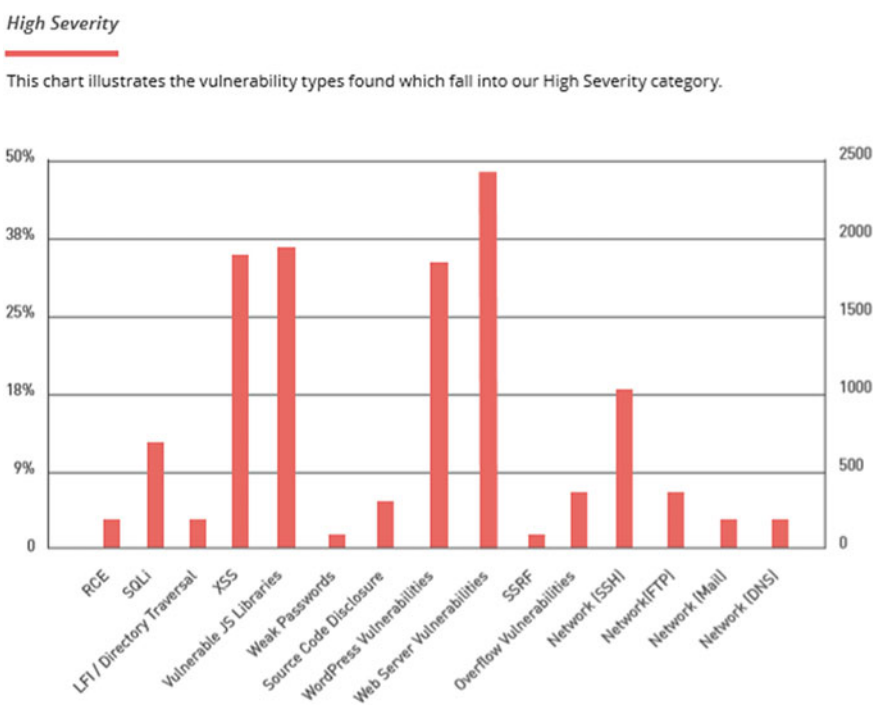


Fig. 2 Vulnerability report for the year 2019 [9]

Kah Seng et al. [11] came up with a study in order to automate penetration testing within the ambiguous testing environment. In this paper, the state-of-the-art black-box web server scanner is tested to explore the method of detecting web application vulnerabilities in an ambiguous testing environment.

Lei et al. [12] designed and implemented an automatic scanning tool to detect SQL injection vulnerability based on a web crawler. This study focused on creating a method that automatically identifies SQL injection vulnerabilities based on web crawlers. Additionally, functions such as URL (uniform resource locator) optimization and similarity determination are added to the characteristics of each module so that the validator can conduct accurate and rapid testing.

Manasa and Reddy [13] web server architecture and vulnerability identification using the Vega vulnerability scanner. This study focuses on designing a web app scanner with an open-source scanner and testing platform called Vega. The Vega architecture already has some structure in place, making it easy to design a good pen testing system in order to detect SQL injection vulnerabilities, such as cross-site scripting.

Mohammadi et al. [14] came up study named "Automated Cross-Site Scripting Vulnerability Detection and Repair by Unit Testing" In this study, to detect XSS

vulnerabilities caused by unsuitable data encoding, a security unit testing technique is used and a vulnerability repair technique was also proposed in order to automatically fix detected vulnerabilities.

Patil et al. [15] proposed an approach to design an efficient web vulnerability scanner. In this study, the clustering method was adopted in order to track injections in cross-site scripting threats, SQL, and XPath. The study was able to grow the vulnerability scanner's identification potential while retaining low false-positive and false-negative rates. This study was only limited to the detection of only SQL and XPath injection.

Parvez et al. [16] proposed a study to analyze the black-box web server scanner's success in detecting stored SQL injection and stored XSS vulnerabilities in this paper. The detection capability of black-box web application security scanners against stored cross-site scripting and stored SQL injection was analyzed and it was concluded that black-box scanners still require enhancements in the detection of stored SQL injection due to the analyzed results of these studies: SQL injections and cross-site scripting.

3 Materials and Methods

This system performs web application security scans to detect the web application security vulnerabilities that attacks can use as exploits to attack web applications which can result in information theft and change of important information. The system will be able to carry out black-box penetration testing (uninformed) and white-box penetration testing (informed) by combining methodologies that are proven to deliver a more accurate vulnerability scan.

The system collects the URL to the web application which the user wants to scan, also known as the target web application, and performs a scan using the hybridized methodology.

Figure 3 shows the flowchart for the proposed system, and the step-by-step processes are highlighted as follows:

- The first step is to check if an account exists for a user.
- If yes move to user login; if no register a user and store user details in the database.
- The next step is to input the URL for the web application which is to be scanned.
- Perform a white-box scan and generate a report.
- Perform a black-box scan and generate a report.

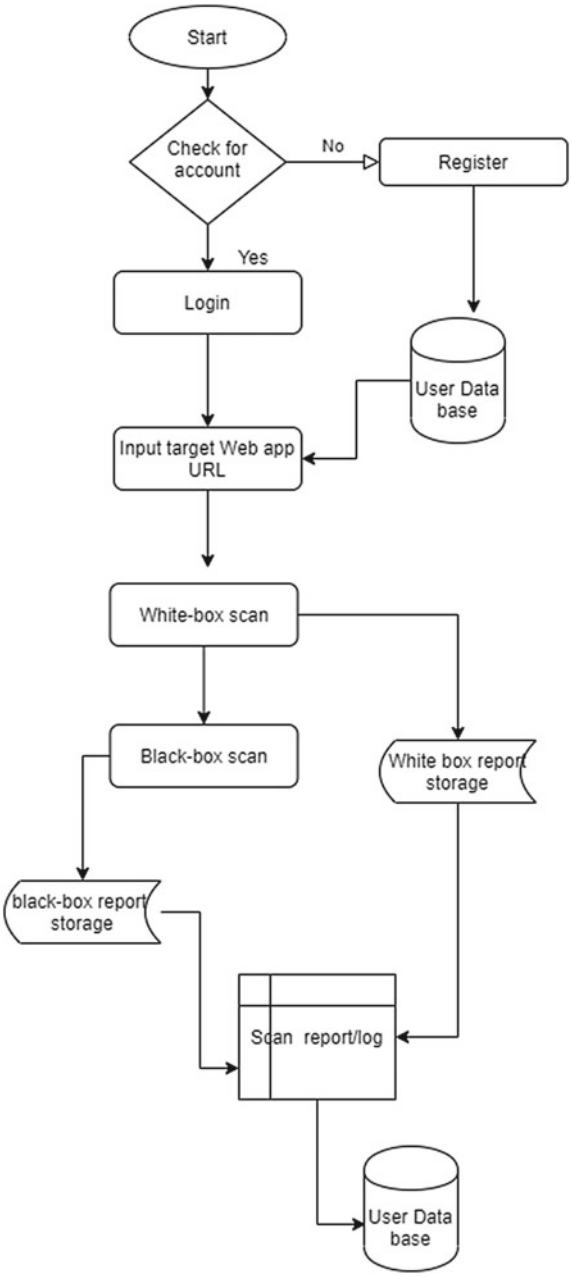


Fig. 3 The proposed system flowchart

3.1 Datasets

The following are links below that contain the data origins used in the implementation of this project.

https://cloud.google.com/security-command-center
https://dev.to/nielsenjared/big-o-logarithmic-time-complexity-gng

4 Results and Discussion

This section contains the step-by-step processes involved in the development of the web security scanner, the tools used, and the screenshot of the output interface. The proposed system consists of an input page that requires the input starting URL for the application target. The target web application must be hosted on either a Google App Engine Standard or Flexible, Compute Engine, or GKE environments. You can also provide IP addresses mapped to starting URLs, but these must be explicitly reserved as static for the current project (e.g., http://172.217.3.206). The scan result output will contain a list of vulnerabilities detected on the target web application. The web security scanner automatically crawls your URL within the scope of the starting URL and tries to exercise as many user inputs and event handlers if possible. The purpose of this work is to develop a system that uses both black-box and white-box methodologies in order to carry out penetration testing.

A vulnerable fragile web application was created and hosted on the Google application engine during the execution and testing period. It served as a case study and allowed one to see the device and its complete functionality. The system screenshots are arranged in a way that gives a proper understanding of the system and how the system executes vulnerability identification scans by providing each page's graphical user interface and the purpose of each screen shown.

Register Page

On this page, the user fills out all field which is required in order to register a user to use this system. Although the system can work without a registered user, this process is necessary if the user wants the result after each scan to be sent to the user's registered email address, as shown in Fig. 4.

Web Application page

The diagram below shows the page the user is redirected after the user login, and it is also the landing page. This page prompts the user to input the service account private key in which the target web application was hosted with owner or editor's rights, which can be either on Google App Engine Standard or Flexible, Compute Engine, or GKE environments that this field required. The next field is the starting URL. For the target web application this field is also required. Then the user hits the search icon bottom for the search to begin as shown in Fig. 5.

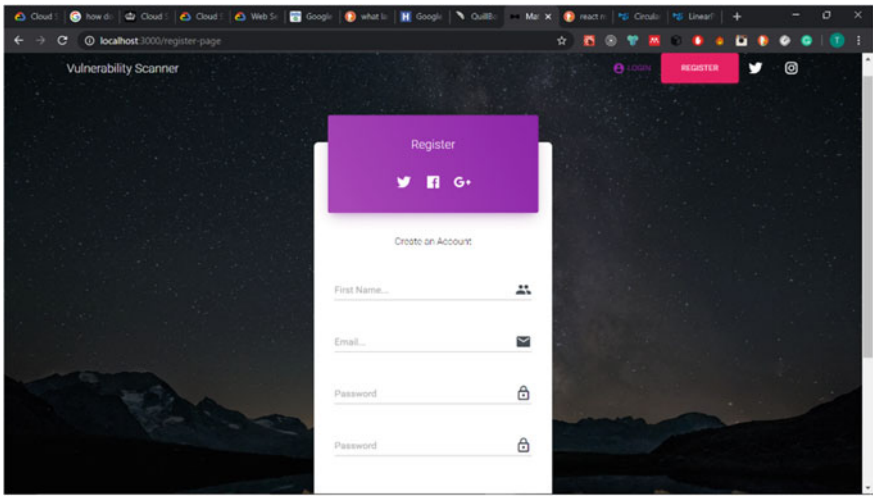


Fig. 4 Register page

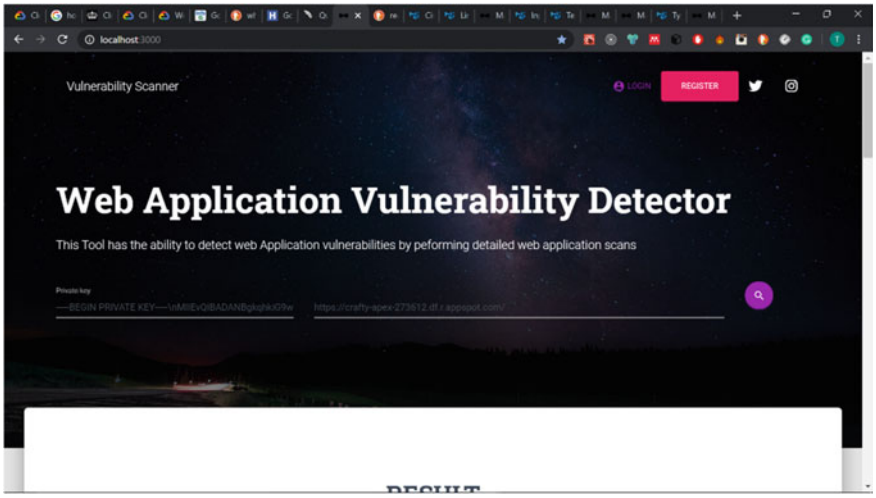


Fig. 5 Configure page

Vulnerabilities displayed page

Figure 6 shows the result from the scan which was completed. The san result has three tabs: the vulnerabilities tab, the URL crawled tab, and the scan summary tab.

Comparative analysis

A comparative analysis of the existing works and security tools was performed with the developed system. The system developed worked on the limitations discovered

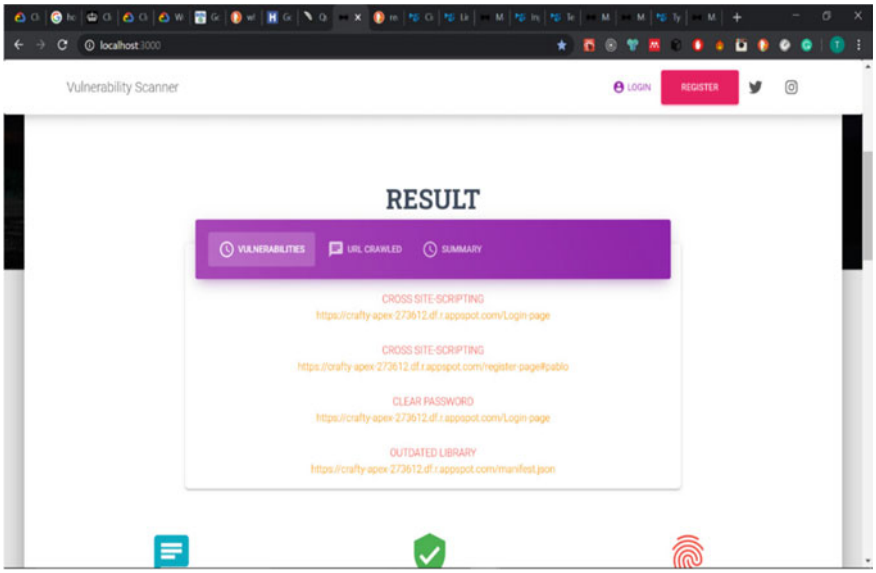


Fig. 6 Detected vulnerabilities

in existing tools and research. Table 1 shows a list of existing research work, the limitations which the developed was able to solve, and expansions that the developed system improved on.

5 Conclusion

The designed system focuses on detecting web application common vulnerabilities easily, with little to no technical knowledge needed by operators which will be normally achievable by penetration testing professionals. The systems can be used easily by developed, and also web administrators. The system hybridizes the two known methodologies in pen testing which are black-box (uninformed) and white-box (Informed) in order to provide a very accurate vulnerability detection with a low false-positive rate the system can detect vulnerabilities. In future works, the system can be improved in order to detect more vulnerabilities which the system cannot detect currently.

Table 1 Comparative analysis with previous researches

S. no.	Authors and years	Methods	Title	Accuracy
	Proposed system	Hybridization of black-box and white-box penetration testing	A web application vulnerability testing system	XSS: 99% Mixed-content: 99% Clear text passwords: 99%
	Lei et al. [11]	Automatic scanning tool for SQL injection vulnerability based on web crawler	Design and implementation of an automatic scanning tool of SQL injection vulnerability based on web crawler	SQL Injection: 74%
	ĐURIĆ [9]	The system detects web application vulnerabilities and uses a powerful page-similarity detection algorithm to make SQLI vulnerability detection modular	Web application penetration testing tool	The system was not evaluated

References

1. S. Kemp, Digital trends 2019: Every single stat you need to know about the internet. Tnw. (2019). https://thenextweb.com/contributors/2019/01/30/digital-trends-2019-every-single-stat-you-need-to-know-about-the-internet/
2. D. Esposito, M. Rennhard, L. Ruf, A. Wagner, Exploiting the potential of web application vulnerability scanning. in *ICIMP 2018 - The Thirteenth International Conference on Internet Monitoring and Protection*, (c), 22–29 (2018)
3. N. Moustafa, G. Misra, J. Slay, Generalized outlier gaussian mixture technique based on automated association features for simulating and detecting web application attacks. IEEE Transactions on Sustainable Computing (2018)
4. S. Misra, A. Adewumi, R. Maskeliūnas, R. Damaševičius, F. Cafer, Unit Testing in global software development environment. Communications in Computer and Information Science (CCIS, vol. 799 2017), pp. 309–317
5. J.K. Alhassan, S. Misra, A. Umar, R. Maskeliūnas, R. Damaševičius, A. Adewumi, *A Fuzzy Classifier-Based Penetration Testing for Web Applications. Advances in Intelligent Systems and Computing*, vol. 721 (Springer, 2018), pp. 95–104
6. CLUSIF, Web application security: Managing web application security risks. (March). https://crypto.stanford.edu/cs155old/cs155-spring11/lectures/10-web-site-sec.pdf(2010)
7. Web Application Attacks Statistics: 2017 Outcomes and Expectations for 2018 (2017). https://www.ptsecurity.com/ww-en/analytics/web-application-attack-statistics-2017/. Accessed 24 Oct 2019
8. Web Application Vulnerabilities: Statistics for 2018 (2018). https://www.ptsecurity.com/ww-en/analytics/web-application-vulnerabilities-statistics-2019/. Accessed 24 Oct 2019
9. Acunetix. Web Application Vulnerability Report 2019. 27. https://www.acunetix.com/blog/articles/acunetix-web-application-vulnerability-report-2019/. Last accessed 05.01.2022

10. ĐURIĆ, Z., WAPTT-Web application penetration testing tool. Adv. Electr. Comput. Eng. **14**(1), 93–102 (2014)
11. L. Kah Seng, N. Ithnin, S.Z. Mohd Shaid, Automating penetration testing within ambiguous testing environment. Int. J. Innovat. Comput. **8**(3), pp. 1–8 (2018). https://doi.org/10.11113/ijic.v8n3.180
12. X. Lei, J. Qu, G. Yao, J. Chen, X. Shen, Design and implementation of an automatic scanning tool of SQL injection vulnerability based on web crawler. In *Advances in Intelligent Systems and Computing* (vol. 895, pp. 481–488). https://doi.org/10.1007/978-3-030-16946-6_38
13. K. Manasa, L.V. Reddy, Open access designing a web application and detecting vulnerabilities using vega vulnerabilities scanner American journal of engineering research (AJER). Am. J. Eng. Res. (AJER) **5**(8), 227–232 (2016)
14. M. Mohammadi, B. Chu, H.R. Lipford, *Automated Detecting and Repair of Cross-Site Scripting Vulnerabilities through Unit Testing*. (April) (2018)
15. S. Patil, N. Marathe, P. Padiya, Design of efficient web vulnerability scanner. In 2016 In *International Conference on Inventive Computation Technologies (ICICT)* (vol. 2, IEEE). pp. 1–6
16. M. Parvez, P. Zavarsky, N. Khoury, Analysis of effectiveness of black-box web application scanners in detection of stored SQL injection and stored XSS vulnerabilities. In *2015 10th International Conference for Internet Technology and Secured Transactions (ICITST)* (IEEE, 2015), pp. 186–191

iReportNow: A Mobile-Based Lost and Stolen Reporting System

Bilkisu Larai Muhammad-Bello, Olatunde Petwilson Lewu, Sanjay Misra, Ajay Kumar Garg, Jonathan Oluranti, and Rytis Maskeliunas

Abstract Modern society is faced with a growing degree of security threats both internally and externally, but access to the right and timely information can be a defining factor. Technological advancements have been advantageous to the police in terms of reporting, monitoring and responding to crime and criminality. At present, crime reportage is being done manually in most police stations via the use of pen and paper to document statements. This method is not only slow and ineffective but could lead to loss of valuable time and productivity. This work presents the design and implementation of "iReportNow", a digital mobile application for reporting loss or theft. It leverages the current proliferation of mobile devices. The application makes information about reported cases of loss or theft readily available and accessible to the populace. The agile model of the system development life cycle (SDLC) was adopted throughout the stages of requirement gathering, to the design, analysis, implementation and testing. The application was implemented using Extensible Mark-up Language (XML) for the user interface, Java programming language for code binding, MySQL and PHP for a robust database. The application was deployed

B. L. Muhammad-Bello · O. P. Lewu
Federal University of Technology Minna, Minna, Niger, Nigeria
e-mail: bilkisu_bello@futminna.edu.ng

O. P. Lewu
e-mail: olatunde.lewu@st.futminna.edu.ng

S. Misra (✉)
Department of Computer Science and Communication, Ostfold University College, Halden, Norway
e-mail: sanjay.misra@hiof.no

A. K. Garg
Wipro Ltd, Gurgaon, India

J. Oluranti
Center of ICT/ICE, Covenant University, Ota, Nigeria
e-mail: Jonathan.oluranti@covenantuniversity.edu.ng

R. Maskeliunas
Kaunas University of Technology, Kaunas, Lithuania
e-mail: Jonathan.oluranti@covenantuniversity.edu.ng

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_57

on 20 Android mobile devices for usability testing and it achieved a mean score of 83.5 based on the system usability scale (SUS). iReportNow will enhance efficiency in reporting and accessing the information on reported cases of loss and theft. Furthermore, it would help to allay uncertainties of buying stolen properties as users can easily check up on an item put up for sale.

Keywords Information management system · Android mobile application · Crime reporting system

1 Introduction

Technological advancements in the telecommunication sector have engendered the proliferation of mobile devices. Of all the technological gadgets available to users, a digital mobile device is an important gadget that plays a prominent role as its functionalities have advanced beyond basic phone calls and short message services. Many now have the ability to run third-party applications and have functionalities to support electronic mail, multimedia messaging and surfing the web. These advancements in mobiles devices coupled with the increase in its application and service possibilities, which are oftentimes not fully exploited, have made it comparable to personal computers (PCs) in recent times [1, 2].

Crime is not an unfamiliar phenomenon to any person in the world irrespective of origin, race or cultural background. While individuals are responsible for reporting loss, theft and providing information about lost or stolen properties in a timely manner, the loss or theft of a property could be considered an emergency or non-emergency. In both cases, however, loss or theft of a property should be reported to the police. Thus, the innate challenge this research aims to address is efficient crime reporting to security operatives in real time to aid quicker response. Moreover, in modern societies, technology-driven security solutions are progressively being adopted by private corporate bodies and governments to track and minimize risk [3, 4]. The advancement in computing capabilities of mobile handheld devices has made them a worthwhile platform for accomplishing a wide variety of operations, including knowledge sharing. The identification, monitoring and response to crime is a new field in which mobile technology is becoming increasingly useful in the society. Reporting systems are characterized as processes, procedures and facilities used to produce and gather reports using one or more information management systems [5, 6].

A lost and stolen reporting system is an information system that offers a forum for the citizenry to quickly and easily report loss or theft of valuables to the appropriate authority. A significant aspect of this framework is that it provides an opportunity for fast and easy reporting while promoting rapid response on the part of law enforcement agencies. Sometimes, the missing and stolen items are sold as used items (or "second-hand" items as they a popularly called in Nigeria) to naïve consumers who are not fully aware due to lack of knowledge. Consequently, such unsuspecting innocent

buyers end up becoming criminal accomplices. Furthermore, because some vendors of used goods portray themselves as the legitimate business owners dealing in the sale of second-hand goods, if customers are not well informed, they could end up on the wrong side of the law if such second-hand goods turn out to be stolen. This fact is corroborated by [7] which reported that there are people who purchase and sell stolen goods, but they just place their buyers at the mercy of the law enforcement agencies.

On one hand, the loss or theft of a valuable property can have a terrible effect on the owner, on the other hand, such victims might have to travel long distances in search of the appropriate authority to report their ordeal. Furthermore, the victims now go through the painstaking process of manual statement documentation and may be required to drive back long distances to their homes to get copies of all necessary documentation because oftentimes, they have little to no knowledge of the required documentation. As a result, victims incur unnecessary transportation costs to and from their respective locations, resulting in a loss of valuable time and efficiency due to this manual reporting system. Thus, there is a need for an effective and efficient lost and stolen reporting system to aid victims in reporting their ordeals on a timely basis for quicker response and an increased chance of recovery. iReportNow is efficient in terms of aiding users to report loss or theft from any location. In addition, it gives access to available information and acts as a warning system, preventing unsuspecting innocent buyers from buying stolen items and becoming accomplices to criminality. The Android platform was chosen as the platform for our implementation because the open-source Android OS offers software developers the ability to access and use all of the source code of the device. There are several hundreds of Android products available in the market worldwide from various vendors, and the Android operating system (OS) accounts for 75.27% of the Mobile OS Market Share in the world, making it the most commonly deployed OS [8, 9].

The rest of the paper is structured as follows: related work in the field of crime reporting systems and mobile-based crime reporting systems is discussed in Sect. 2; Sect. 3 presents the research methodology for this work; we describe iReportNow and its system architecture in Sect. 4; Sect. 5 presents the user study and usability testing conducted with 20 participants; and Sect. 6 concludes the paper.

2 Related Work

In recent times, many works have implemented mobile-based applications for addressing present-day issues [10–12]. Specifically, the stolen vehicle reporting system has been widely addressed in the literature because it provides a quicker way of making a report. Using geo-crowding services, a vehicle can be probed at the point of registration for reported cases of theft. This way, officials can monitor and track vehicles reported as stolen at the point of reregistration [13]. In [14], for instance, crime can be reported in real time by using geolocation technology to

map the incident locale. Images, audio and video of crime scenes can be sent to the responding authorities in real time.

Many works have implemented mobile-based crime reporting system. Reference [5] uses a client–server architecture web-based mobile structure for identifying, reporting and monitoring criminals in real time, relying on geolocation technology. When reporting crimes and other illegal activities, anonymity is a concern so as to protect the identity of the concerned individual. The system proposed in [15] combined digital certificates, symmetric and asymmetric keys, and digital signature in protecting the identity of users. The cryptographic mechanism used for improved security and identity confidentiality also allows for data integrity and availability.

In [16], for instance, the proposed system is a web-based database that creates a dynamic list of lost and found items, while also automating the process of reporting incidences. The Locate It! Project [17] improved on this by implementing an online social community that connects people with lost and found items. The system leveraged the social interactions among registered users allowing them to report missing as well as found items. The system also leveraged the power of a matching algorithm to pair users of lost and found items with the watchful eyes of the system administrator. Reportage of crime via voice message, images, video and location capturing through the global positioning system (GPS) was implemented in [18]. The system thrives on communication using a network transporter or relay (communication between activated Wi-Fi-enabled devices in the absence of a network). It is broken into four modules: the user module which resides on the users' smartphone, a centralized database, a police station module and patrol module for first responders.

In [19], a mobile platform called Transafe was proposed to collect and analyze the public perceptual experience of safety via a cell phone program in the City of Melbourne, Australia. The program encourages users to record crimes and misdemeanors on the basis of their thoughts and perceptions at a given time and location. The crowdsourced data is then analyzed to establish the collective intelligence of different places in the city as well as their effective states at different times of the day. Furthermore, the framework merged the smartphone application with the social media platform to monitor and evaluate public users' views about a location in the city using "crowdsourcing" intelligence-enabled via the spatial movement of the large group of people that is voluntarily recorded on their smartphone app which can detect their positions using GPS and geo-tags. In [20], an iPhone application called "iPol" for crime-fighting in Geneva, Switzerland, was proposed. The program provides police officers with statistics and details about the most recent crime incidents, such as the most recent robbery or violent assault and the description of the suspects of the incidents. Police personnel is also able to trace and verify the ownership of a car by inserting its number plate in the iPol application. In addition, real-time information on patrol hours, location, police directories, attorneys, translators and on-the-job availability are also made available on-the-go to police officers via the mobile application.

Continued advances in technology have provided a vast incentive for engineers and technologists to delve further into discovering efficient and effective alternatives for a variety of inefficient manual procedures. In [21] a real-time vehicle inspection

and safety management system is introduced that would allow stolen vehicles to be identified during routine vehicle checks on the road. The real-time system facilitated quicker access to critical information on the authenticity and legitimacy of the vehicle documentation presented during the regular inspection. It leverages the instant messaging functionality found in any Global System for Mobile Communications (GSM) phone to access all vehicular information available on the system by law enforcement officials. It requires submitting the vehicle registration details to a special code that automatically responds by returning all the vital information on a car via a text message.

Palagiri [22] proposed an anti-theft application for Android smartphone. The system allows a user at the point of registration to designate a close acquaintance to act as an Administrator to monitor the activities of the user's phone. The Administrator's smartphone receives a periodic update from the activities of the user's phone, such as location and Subscriber Identification Module (SIM) International Mobile Equipment Identity (IMEI). In the case of theft and subsequent change of the SIM card, the user phone transmits a signal to the admin notifying the location and the IMEI of the new SIM inserted. The admin can notify the user of the new location of the phone via the location services and the sim card manager in the Android framework.

In [23], a real-time tracking system that leveraged a compact Internet of Things-based (IoT) GPS module was proposed. With the device inserted in valuable items such as a bag or car, the location of such valuables is tracked via GPS and GSM module that uploads data to an open-source ThingSpeak cloud API via General Packet Radio Service (GPRS). The exact location of the valuable item is shown on Google map either using a mobile application or through a web portal. The system also includes a direction feature that provides the shortest possible route to the location of a lost or stolen item.

3 Research Methodology

In order to improve the reporting and monitoring of crime in our cities, especially in stolen and loss of valuable items, this study researched the various approaches to reporting crimes in real time to the appropriate authorities with the aim of eliminating the bottleneck of time-wasting and paper-based crime reporting. Thus, a mobile platform for reporting and a web platform for monitoring are proposed. The system proposed in this paper has been developed through the agile methodology of the system development life cycle since it relies on an original, simpler implementation, which can eventually become more complicated and more detailed before the final framework is finished.

The development process started with an initial information gathering from the police authority on the ways and procedures of reporting lost and stolen items. Following this information gathering from the police authority, it became imperative to hear from the victims their experience in a situation like this so as to have

a clearer direction, though one of the researchers was once a victim of car theft himself. After the initial information gathering process, the initial interactive prototype was developed taking into consideration the two categories of user-targeted. Some scenarios were created to foster brainstorming and discussions on the app design considering the two paramount stakeholder requirements in this process: the victims of auto and electronic theft; and the unsuspecting victims who buy stolen gadgets as fairly used items unknowingly. Subsequently, we improved on the initial prototype and evaluated the different stages of the project.

A remarkable development we observed, which is distinctive of the evolution of systems and their users, is that we ended up developing a generalized app. At inception, the idea was to develop an application that would facilitate reporting an incident of theft or loss of valuables easily and quickly without the hassles involved currently in Nigeria. For this reason, it was imperative to make the app easy to use and learn since the app is not solely for the report of theft and loss, but also for allowing ordinary citizens access to information at their fingertips so as not to fall victims by becoming criminal accomplices as a result of buying stolen valuables. The app we ended up creating can work in any other country of the world with a properly populated underlying database.

4 iReportNow

The app iReportNow relies on the geolocation feature available on Android phones to allow mapping the user's current location with the nearest police station for faster response. The overall system architecture as depicted in Fig. 1 includes a MySQL database for storing all data related to the user, police station and valuable items. The PHP engine is used for code binding and robustness, the apache web server for the web interface administration located in every police station, as well as a native mobile application running on the Android operating system.

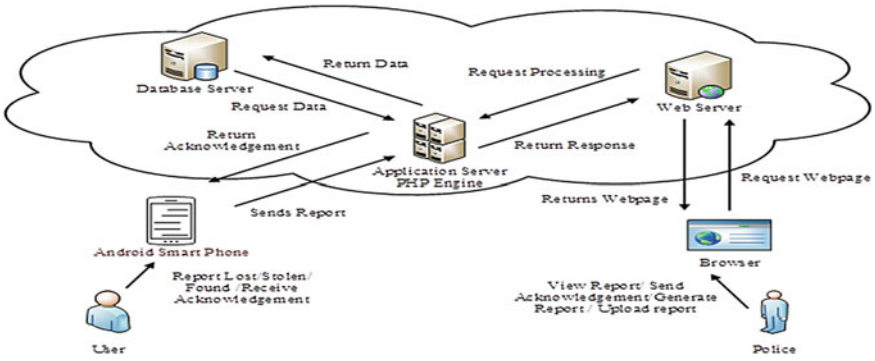


Fig. 1. iReportNow system architecture

The main features of the mobile application are described in the following subsections. A client (user and police) connects to the system with an Android smartphone and a web browser, respectively. User information is tested in the PHP engine and then submitted to the database server for persistence. The police request details from the web server with an HTTP link. There is a two-way link between the clients and the server that enables timely response to client requests. The visual interface of the iReportNow mobile app is described in the next section.

The activity diagram as depicted in Fig. 2 describes the dynamic aspects of the system in terms of the abstracted underlying important operations that have to be carried out to achieve all the system functionality of raising an alert. It indicates hypothetical yet essential actions that need to be taken in order to accomplish the goal of reporting or alerting and it physically represents the design of the device. After installing the app, a user can verify the status (missing or stolen) of an item via a simple database query prior to registering or signing up for an account/profile. This search functionality offers valuable information to customers of fairly used goods in an efficient manner. If there is a search hit (that is item found), the user can further send an alert to the police station with the nearest proximity to the location of the user's phone as captured on the GPS.

Registration of an account on the app by the user is mandatory for the user to access all other system functionalities with the exception of the functionality for "*Raising an Alert*" to the appropriate authority as shown in the system flowchart. Therefore, new users are prompted to create an account and then login to their profile to access other system functionalities. The platform also authenticates user credentials at the login stage before granting access to the system. The data model of iReportNow is depicted in Fig. 3 via an entity relationship diagram. The entity "***Items***" is a generalized superclass of the subclasses "***Electronics***" and "***Vehicles***"

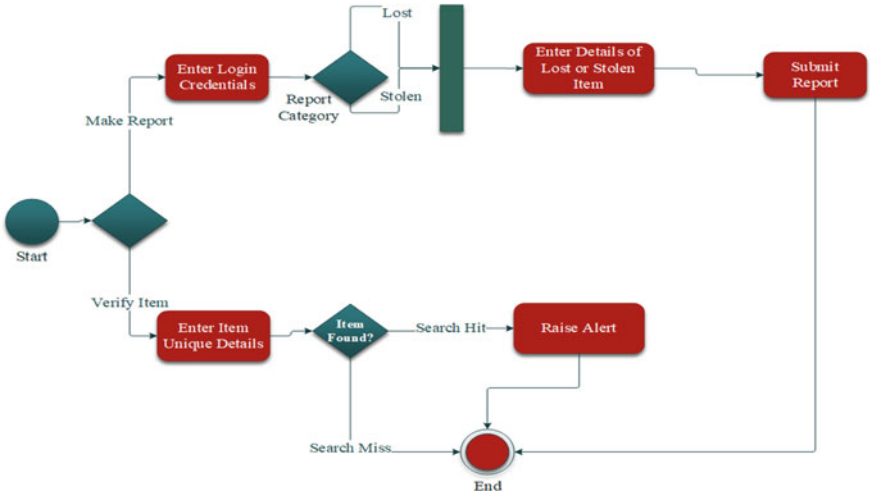


Fig. 2 Activity diagram

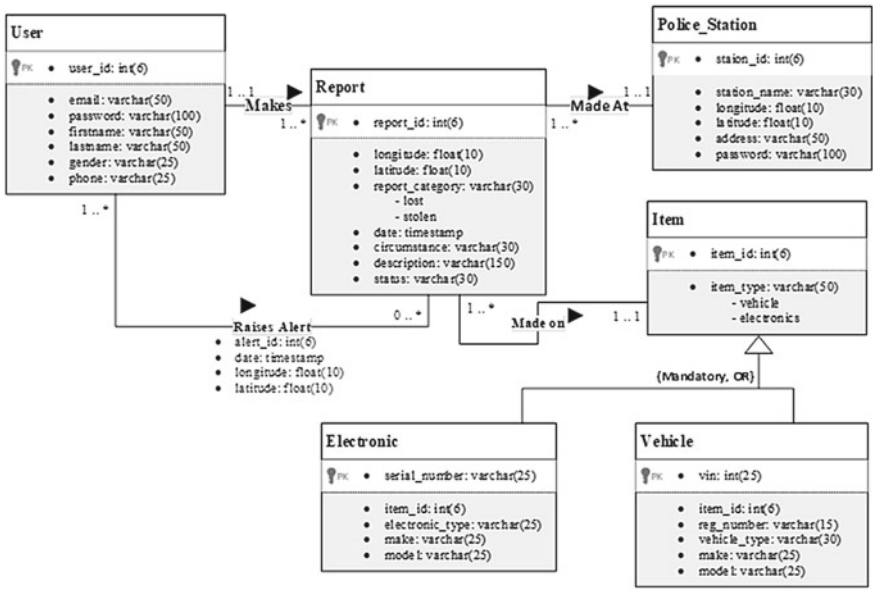


Fig. 3 Entity relationship diagram

both inheriting the attributes of the superclass "***Items***". The database is populated at the point of registration and reporting. To prevent multiple report of the same case, reports are channeled to the police station closest to the users' location.

a. **Mobile Application User Interface**

The application iReportNow is structured into three parts as follows: an onboarding screen (see Fig. 4a) shown during application startup, the core/essential part of the app and the navigation. The essential part of the application is further divided into three sections. First is the Dashboard (see Fig. 4a) area that gives users access to make a new report, view entries of the report made and an alert raised, as well as being able to access the police website.

In the report screen, users could select the category and type of item to report which presents to the user the opportunity to fill in unique attributes of the item being reported.

In the case of a vehicle, users have the option of entering the vehicle number plate and chassis or vehicle identification number (VIN) and other details of the vehicle for proper referencing. While in the case of an electronic, the user could enter the unique identifier for such electronic such as the IMEI of a mobile phone and product/serial number for other kinds of electronic gadgets, as seen in Fig. 5a, b.

The second section of the essential part of the application is item verification. This allows potential buyers to verify the authenticity of fair use of electronic or vehicle for any reported cases of theft or loss. A buyer can discretely input the VIN or chassis number of a vehicle to verify, or the IMEI number of a mobile phone which

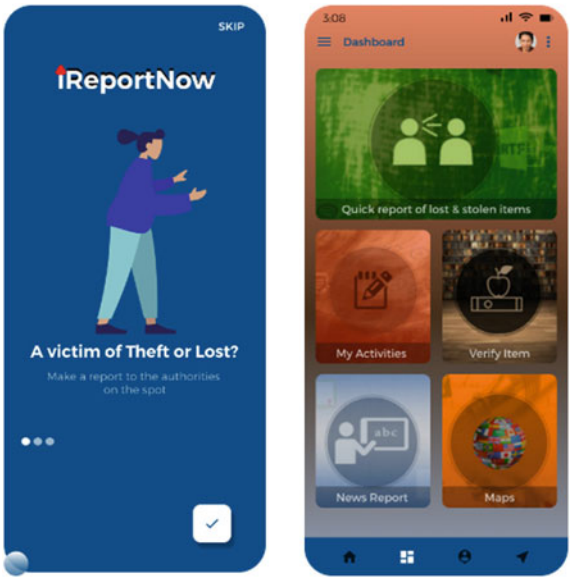


Fig. 4 **a** App onboarding **b** Application dashboard

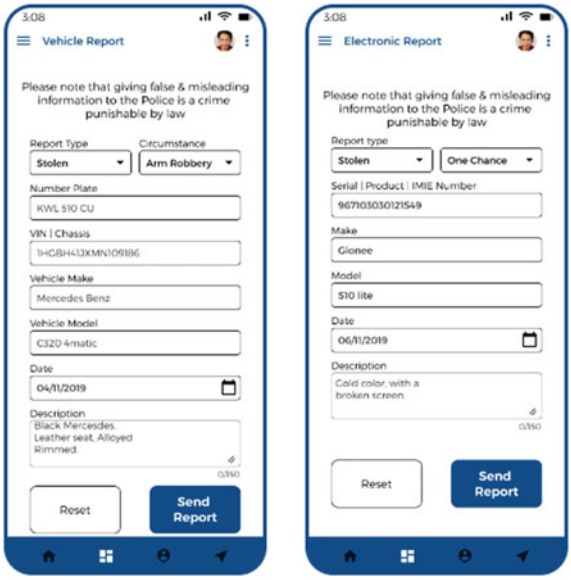


Fig. 5 **a** Vehicle report **b** Electronic report

can be queried by dialing the *#06# sequence, or the product/serial number of other electronic gadgets. If there is a hit, the information will be displayed, showing the various details of the item and the circumstances involved in the theft or loss. The police authority can be alerted immediately with the touch of the alert button and the location capture via GPS (Fig. 6a, b).

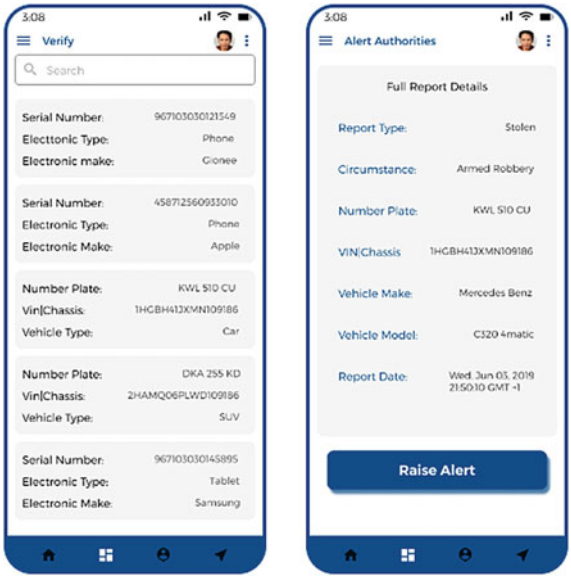


Fig. 6 **a** Item verification **b** Alert authorities

The last section is the menu icon at the top left of the Toolbar which allows users to access the Navigation Drawer for the basic application functionalities. Users could logout, carry out profile settings, as well as access the report and verification screens (Fig. 7).

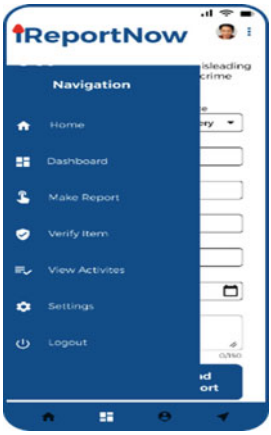


Fig. 7 Navigation drawer

5 User Study

a. **Methodology**

The usability testing experience was carried out with a total of 20 participants. The participants performed tasks of reporting, verifying, and navigating the application for user experience and testing other features and functionalities of the application. The application was installed on each participant's Android phone with a brief explanation of the goals and of the study in context.

We administered the system usability scale (SUS) questionnaire to assess the usability of the iReportNow. We were able to ascertain the apparent difficulty and completion rate of using the application through the collection of qualitative data from direct observation and participants' responses. We also got feedback from participants by administering questionnaire through well-structured questions as shown below:

Q1: Have you used this type of application before?

Q2: If yes, would say this is better?

Q3: How easy is it to use and figure out the app?

Q4: Is the app fast for query report?

Q5: What is the accuracy of the query result?

Q6: How likely are you to recommend this app to a friend or colleague?

Q7: Do you consider the application useful?

Q8: Do you have an overall comment/suggestion about the application?

b. Results

Using the SUS, we assessed the perceived difficulty participants experienced carrying out the various tasks based on three scenarios: (1) the ability of participants to complete tasks using the system and the effectiveness, (2) the level of resources consumed in performing the task (efficiency) and (3) the participants' subjective reactions to using the system (satisfaction) [24].

This was based on the qualitative Likert scale of {"very easy", "easy", "medium", "difficult", "very difficult"}. The majority of the participants found the task of making a report to be "very easy"; Task 2, the searching and verification of items for a possible report of theft or loss was also assessed as "very easy" or "medium" by most users. The navigation section was deemed easy by majority of the respondents. Table 1 summarizes the results of the SUS. The mean score was 83.5 with a standard deviation of 8.5.

The least score was 70, and a score of 70 is commonly considered as the acceptable threshold for positively adjudging that a system is easy to use [24].

After the participants' interactions with the application, we received feedback from the participants. The majority of users were excited for the ease of use and the

Table 1 Result of the system usability scale (SUS)

Participants	SUS score
P1	82.5
P2	90
P3	95
P4	85
P5	82.5
P6	87.5
P7	75
P8	72.5
P9	80
P10	85
P11	92.5
P12	90
P13	70
P14	95
P15	72.5
P16	75
P17	72.5
P18	87.5
P19	82.5
P20	97.5
Mean	**83.5**
SD	**8.5**

ability to make on-the-spot report and also verify fairly used items before buying. Respondents suggested improving the application by allowing the application to capture any type of report other than theft and loss of vehicle or electronic gadgets, while others proposed the addition of voice messages and text-to-speech mechanism for the visually impaired.

The graphical representation of all the responses we got from the questionnaire administered to the 20 users who interacted with iReportNow during the testing phase of the application development is shown in the clustered chart depicted in Fig. 8. The chart shows the total number of users on the vertical axis plotted against their response to each question on the horizontal axis. The green, orange and red columns indicate positive, neutral and negative responses, respectively. In general, the result shows that most of the participants assessed the app positively.

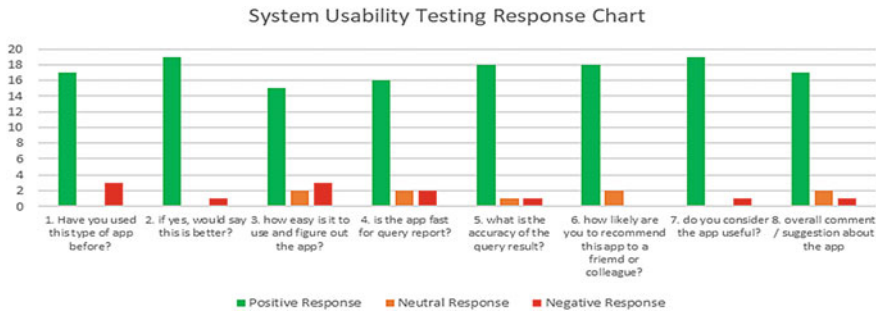


Fig. 8 System usability testing clustered chart

6 Conclusion

In this paper, we have presented iReportNow, an Android-based mobile application, for reporting and verifying stolen or lost valuables. The application aims to ameliorate the inefficiency inherent in the current system of reporting stolen or missing valuables. Therefore, aiding users to report loss or theft of valuables from any location on a timely basis for rapid response and an increased chance of recovery. In addition, the app grants access to available information on lost and stolen items to unsuspecting potential buyers. Consequently, acting as a warning system to prevent unsuspecting innocent buyers from buying stolen or lost items and becoming accomplices to the crime.

We have also presented the system usability test carried out with 20 participants who adjudged iReportNow to be useable. However, there are some limitations that should be looked into in the future. The system can be enhanced to allow it to capture more than just vehicle and electronic, making it an all-round reporting system. Furthermore, iReportNow could be integrated with functionalities such as text-to-speech and speech-to-text to cater to a more diverse group of users.

References

1. O. Jonathan, S. Misra, E. Ibanga, R. Maskeliūnas, R. Damaševičius, R. Ahuja, Design and implementation of a mobile webcast application with google analytics and cloud messaging functionality. In: Journal of Physics: Conference Series: the 3rd International Conference on Computing and Applied Informatics, vol. 1235, 1st edn. (IOP Publishing, Bristol), pp. 1–6 (2019)
2. F. Hacklin, J. Björkdahl, M.W. Wallin, Strategies for business model innovation: how firms reel in migrating value. Long Range Plan. **51**(1), 82–110 (2018). https://doi.org/10.1016/j.lrp.2017.06.009
3. C.M. Mubaraka, I.M. Jirgi, P.L. Nanyanzi, Integrating ICT in traffic police department in Uganda: design and development of traffic case management system (TCRIS). Innov. Syst. Design Eng. **4**(5) (2013)

4. A. Oludele, E.E. Onuiri, O.A. Olaore, O.O. Sowunmi, A real-time crime records management system for national security agencies. Eur. J. Comput. Sci. Inf. Technol. **3**(2), 1–12 (2015)
5. A.A. William, A.A. Millicent, Mobile solution for metropolitan crime detection and reporting. J. Emerg. Trends Comput. Inf. Sci. (2013)
6. A.J. Karim, The significance of management information systems for enhancing strategic and tactical planning. JISTEM-J. Inf. Syst. Technol. Manag. **8**(2), 459–470 (2011). https://doi.org/10.1590/S1807-17752011000200011
7. The Guardian Business Services, https://m.guardian.ng/business-services/fairly-used-items-two-sides-of-a-coin. Accessed 27 Auguest 2020
8. S. Roman, D. Phil, S. James, T. Nelson, *The Android™ Developers's Cookbook: Building Applications with the Android SDK*. 2nd edn. (Addison-Wesley, Indiana 2013)
9. Statcounter GlobalStats, https://gs.statcounter.com/os-market-share/mobile/worldwide. Accessed 27 Auguest 2020
10. O. Jonathan, C. Ogbunude, S. Misra, R. Damaševičius, R. Maskeliunas, R. Ahuja, Design and Implementation of a Mobile-Based Personal Digital Assistant (MPDA), In *Computational Intelligence, Communications, and Business Analytics. CICBA 2018. Communications in Computer and Information Science*, vol. 1031, ed. by J. Mandal, S. Mukhopadhyay, P. Dutta, K. Dasgupta (Springer, Singapore, 2019) https://doi.org/10.1007/978-981-13-8581-0_2
11. N. Vatanasuk, A. Chomputawat, S. Chomputawat, W. Chatwiriya, Mobile crime incident reporting system using UX dimensions guideline. In *2015 Asian Conference on Defence Technology (ACDT)* (2015), pp. 187–192. https://doi.org/10.1109/ACDT.2015.7111609
12. G. Jesil, S.P. Basant, R. Pratishvir, Crime reporting system using android application. Int. J. Pure Appl. Math. (IJPAM) **119**(7), 533–538 (2018)
13. A. Victor, Design and Implementation of Vehicle Security Alert System (V-SAS). [Unpublished], Department of Information and Media Technology, Federal University of Technology Minna (2018)
14. D. Adetunji, Realtime Crime Reporting System. [Unpublished], Department of Information and Media Technology, Federal University of Technology Minna (2017)
15. T.-F. Shih, C.-L. Chen, B.-Y. Syu, Y.-Y. Deng, A cloud-based crime reporting system with identity protection. Symmetry **11**, 255 (2019)
16. E. Al-Bataineh, B. Bataineh, S. Kindi, Design, development and usability evaluation of an online web-based lost and found system. Int. J. Digital Inf. Wirel. Commun. (IJDIWC) **5**, 75–81 (2015)
17. N. Alomar, W. AL-Rashed, The analysis and design of a web-based social network: locate it! project. Int. J. Comput. Sci. Issues (IJCSI) **10** (2013)
18. D. Lal, A. Abidin, N. Garg, V. Deep, Advanced immediate crime reporting to police in India. Proced. Comput. Sci. **85**, 543–549 (2016). https://doi.org/10.1016/j.procs.2016.05.216
19. M. Hamilton, S. Flora, C. Eva, S. Choy, Transafe: a crowdsourced mobile platform for crime and safety perception management, in *IEEE International Symposium on Technology and Society* (2011), pp. 32–37
20. Technocrazed iPol Mobile Smartphone application for Genevese Police to Fight Crime, http://www.technocrazed.com/ipol-mobile-smartphone-application-for-genevese-police-to-fight-crime,lA- 2020/08/27
21. B.N. Iduh, O.E. Ositanwosu, C.O. Ugwunna, Real-time vehicle inspection and security management system. Int. J. Comput. Trends Technol. (IJCTT) **63**(1), 11–17 (2018)
22. M.R. Palagiri, H. Revanna, V.C.P. Venkata, K. Nikhilesh, A. Ajil, Anti-theft application for android smartphone. Int. J. Adv. Res. Comput. Sci. **11**(1) (2020)
23. D.B. Dheez, S. Khushwant, A. Prashant, K.S. Hitesh, Real time tracking system: an iot based application. in *5th International Conference on Next Generation Computing Technologies (NGCT-2019)*
24. B. Simone, F. Stefano, L. Marco, On the dimensionality of the system usability scale: a test of alternative measurement models. Cogn. Process. **10**(3), 193–197 (2009)

Improving Security and Privacy in Attribute-Based Encryption with Anonymous Credential

Yuping Yan and Péter Ligeti

Abstract Attribute-based encryption [19] is developed based on fuzzy identity encryption. In the identity-giving cryptosystem, the user's public key can be any information that identifies the users. The original ABE schemes define a Central Authority (CA) for key management, allowing users to verify the ownership of attributes and thereby get access to sensitive resources. However, this CA construction is contradictory to the distributed computing requirement and compromises the users' privacy. Meanwhile, it does not fit the large-scaled computing environment. Thus, in our paper, we propose a new anonymous, decentralized, unlinkable access and collusion attack that prevented CP-ABE scheme with a pseudonym system by Chaum to improve user privacy and data security.

Keywords CP-ABE · Pseudonym system · Decentralization

1 Introduction

The concept of attribute-based encryption (ABE) is from public key encryption. It achieves fine-grain access control by its attributes verifying, which is more suitable in the cloud environment. In ABE schemes, ciphertext and secret key are in pairs, which are both associated with attributes. Data owners can encrypt the message by an access policy based on data users' attributes. Only when there is a match of access policy, the ciphertext can be decrypted [21].

This research has been partially supported by Application Domain Specific Highly Reliable IT Solutions project which has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the Thematic Excellence Programme TKP2020-NKA-06 (National Challenges Subprogramme) funding scheme).

Y. Yan (✉) · P. Ligeti
Department of Computeralgebra, Eötvös Loránd University, Budapest, Hungary
e-mail: yupingyan@inf.elte.hu

P. Ligeti
e-mail: turul@cs.elte.hu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_58

ABE was first proposed by Sahai et al. [19] as a one-to-many cryptosystem. ABE can be further classified as key-policy attribute-based encryption (KP-ABE) [12] and ciphertext-policy attribute-based encryption (CP-ABE) [1] by the difference of where the access policies are settled. In KP-ABE, the key policy represents as an access tree, and it is embedded into the key. Conversely, in the CP-ABE scheme, cipher policy is defined as an access tree, and it is deployed in the ciphertext. Both the schemes support attributes verification and complex tree policies. However, KP-ABE fits more that accessors determine access strategies, such as pay-TV systems and Publish and Subscribe systems (Pub/Sub), while CP-ABE works better when the data owner decides the access requirements, such as electronic medical systems and IoT applications. ABE can effectively fight data transparent and point-to-point key encryption mode problems compared with traditional public key infrastructure encryption schemes. Firstly, messages are decrypted by verifying visitors' attributes with access structure without paying attention to members' exact numbers and identity. It changes decryption mode from one-to-one to one-to-many, while the decryption cost is reduced, and visitor's privacy is preserved. Meanwhile, ABE mechanisms support flexible access control strategies by the implementation of various thresholds and logical operations.

However, these traditional schemes are not entirely realistic. The main technical reasons are as follows: (1) Centralized key management. There is always a central authority (CA) in transitional ABE schemes to generate and manage the public keys for all members. In this case, it is too heavy to distribute all key generation workload to a single institution. (2) Identity information leakage. Even though some papers proposed multi-authority ABE instead of a single third party, it is still possible for a collusion attack. (3) Keys and attributes revocation problem. In classic ABE mechanisms, keys and attributes updating are not straightforward, and it asks for enormous computational costs.

From the technical problems mentioned above, a cipher-policy attribute-based encryption scheme with a pseudonym system is proposed in this paper, which starts with the fundamental works of Chaum [11], and recent researches by Camenish and Lysyanskaya [7]. Pseudonym systems, also known as anonymous credential schemes, allow users to prove the procession of some attributes anonymously and unlinkably. Attributes are properties of individuals. For example, if we want to describe a refrigerator in IoT devices, the attributes can be location, device ID, temperature, etc. The attributes are grouped in a *credential*. In fact, one credential attribute is not enough, so a user can possess multiple distinct credentials [8]. We call this as attribute-based anonymous credential scheme (ABC in short).

We present a cipher-policy attribute-based encryption (CP-ABE) scheme with attribute-based credentials (ABC). In fact, it converts to a user-based attribute-based encryption. The contributions of this paper are in the following:

– **Removing the central authority:** In both original ABE and decentralized ABE, the central authority (CA) or multiple authorities are assumed to work honestly [10]. However, in our scheme, data users take over authorities' responsibilities, which turns into a decentralized encryption mechanism. Attribute authorities can

be semi-honest. In our case, AA can try to get some sensitive information of users or even revocation referee, but it will not influence the encryption scheme and data security of the data owner.
- **Anonymous user authentication:** There has been a little relationship between anonymous credentials and ABE [9]. In our construction, we take advantage of the anonymous credentials scheme for ABE. Thus, users can anonymously prove the procession of some attributes to access sensitive resources and decrypt data.
- **Collusion resistance:** Our scheme prevents both accessor collusion attacks and authority collusion attacks. The authorities do not communicate with each other, so authority collusion is impossible. Accessor collusion resistance can be reached by the original ABE scheme.
- **Revocations:** The proposed scheme is coupled to revoke credentials directly by a dedicated revocation referee. This process conducts in the ABC protocol, not in the ABE protocol anymore, which reduced the workload and algorithm complexity of authorities.

After precising some needed background of related works in Chap. 2, we informally describe, in Part 3, the notations and review classical definitions with some necessary primitives. In Sect. 4, we describe our protocol from involving entities, model security features, the algorithm, and protocol. Section 5 gives a model security analysis. Chapter 6 is our conclusion and future work part.

2 Related Works

In this part, the related works of decentralized attribute-based encryption and anonymous credential schemes are introduced.

2.1 Decentralized Attribute-Based Encryption

Decentralized attribute-based encryption has three entities: n attribute authority (AA), data owner (DO), and data user (DU). It can be categorized into authority-level decentralization and DO-level decentralization.

Authority-level decentralization: The typical feature of the authority-level decentralization approach is to distribute the workload of the CA among authorities. It is also known as multi-authority attribute-based encryption.

In Hur et al. [14] scheme, the central key authority institution is spitted into key generation authority and attribute center to solve the key escrow problem. In this two-party communication protocol, these two authorities issue private keys to data users (DU) without revealing their master keys to each other.

Yang et al. [22] suggested a universal attribute multi-authority ABE mode. However, in this scheme, a root authority is still requested for private key generation for each attribute authority.

To solve the security vulnerability bought by CA, Bozovic et al. [4] gave a multi-authority ABE scheme on the basis of the DBDH assumption. However, the server is still the honest-but-curious central authority.

DO-level decentralization: In the same paper, Bozovic et al. [4] listed several recipients who can decrypt the ciphertext. In this case, DOs cooperate with multiple authorities and take partial responsibility for decryption. Similarly, with Bozovic's model, Sandor et al. [20] developed an ABE model in which DOs collaborate with multiple authorities in the key generation phase.

2.2 Anonymous Credentials Schemes

Anonymous credentials were proposed by David Chaum in 1985 [11]. This construction enables to authenticate individuals without identification. It assumes that each user with a private key and interacts with each authority under a different pseudonym. These multiple pseudonyms cannot be linked together to the same user.

Recently, this notion was renamed attribute-based credentials (ABC in short). There are four main parties in the privacy-preserving ABC schemes, a *User*, an *Issuer*, a *Revocation Referee*, and a *Verifier*. *User* as the central entity receives credentials from *Issuer* and gets attributes updates from *Revocation Referee*. These updates are also sent to the *Issuer* and the *Verifier*. The *User* generates from his credentials a presentation token that contains the required information and the supporting cryptographic proof [17]. Thus, the *Verifier* can check *User's* identity by the *User* own credentials. After receiving the information, the *Users* will convey the credential to the Verifier, which we sometimes call service providers. The service provider will decide which information can be revealed and which should be kept private.

Lysyanskaya et al. [7] suggested a zero-knowledge proof with ABE and apply it in IBM's Idemix [5] project. Brands et al. [2] proposed a protocol for implementing digital credentials that include multiple attributes, and this protocol belongs now to Microsoft's U-Prove technology [3].

Figure 1 is the traditional ABC protocol infrastructure.

3 Preliminaries

Access structure. Let $A = \{A_1, A_2, ...A_n\}$ be a set of attributes. In set $\Gamma = 2^{\{A_1, A_2, ...A_n\}}$, for $\forall M, N$: if $M \in A$ and $M \subseteq N$, then $N \in \Gamma$, we say Γ is an monotonic access tree.

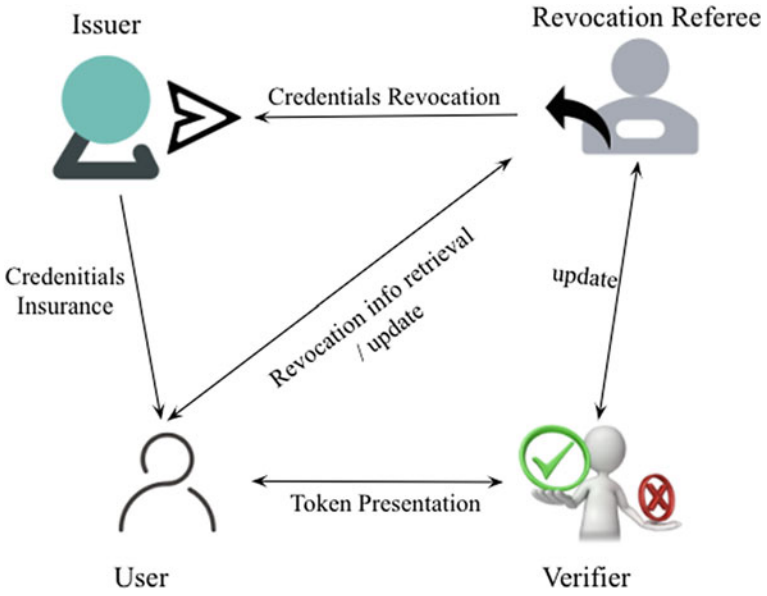


Fig. 1 Entities and interactions diagram of ABC protocol [6]

Review of CP-ABE. The framework of a CP-ABE proposed by Waters [1]. There are four algorithms: $Setup()$ to initialize all parameters, $KeyGen()$ to generate the keys, $Encrypt()$ to encrypt, and $Decrypt()$ to decrypt. More precisely,

Definition 1 A tuple of algorithms $\Pi = (Setup, Encrypt, KeyGen, Decrypt)$ is called a CP-ABE scheme if

a. $Setup(\lambda, U) \rightarrow (PK, MSK)$: In this initialization phase, the set of security parameters and attributes are taken as input to get master secret key MSK and the public initialization parameters PK.
b. $Encrypt(m, \mathcal{A}, PK) \rightarrow CT$: The task here is to generate the ciphertext CT from the inputs of the original message m, the access structure $\mathcal{A}$ over U, and the public parameters PK. We assume that CT contains the description of $\mathcal{A}$ itself and only users with attributes satisfying $\mathcal{A}$ are able to decrypt that.
c. $KeyGen(S, MSK) \rightarrow SK$: On the basis of the master key MSK and the attribute S, this algorithm outputs the secret key SK for S.
d. $Decrypt(PK, CT, SK) \rightarrow m$: In the final decryption phase, according to the public parameters PK, a ciphertext CT (containing $\mathcal{A}$), and the secret key SK for a set S of attributes, it produces an element of the message space. Here, we assume that the output is the original m if S satisfies $\mathcal{A}$.

4 Model Description

This section presents the model architecture in the aspects of enrolled entities, model security features, and the protocol scheme. The main achievements are the following:

- Decentralized authorities: It divides central authority (CA) responsibility from the traditional scheme into multi-attribute authorities, attribute issuer, verifier, and revocation referee.
- Flexible revocation: It distributes authorities' revocation stress into revocation referee, which enables users' attributes to update accordingly and directly.
- Fine-grained access control: Since there is a match of attributes between the ciphertext and the private key, a user can successfully decrypt the message.

4.1 Entities

In the proposed system, there are five entities: *Users*, *Issuer*, *Verifier*, *Revocation Referee*, and *Attribute Authorities* (Fig. 2). The roles of these entities are the following:

– **Users (U)**: Users or individuals who are eager to get access-sensitive data by providing their personal information are an essential part of the whole architecture. The ultimate goal is to make the attributes (credentials) confidential. Here, "Users" represents a group of individuals. It can be more Users, like User1, User2, etc., in the system. There are three actions of the Users in ABC protocol and one action in ABE protocol. In ABC protocol, Users get credentials and private keys from Issuer. If there is any change of the Users' attributes (credentials), Revocation

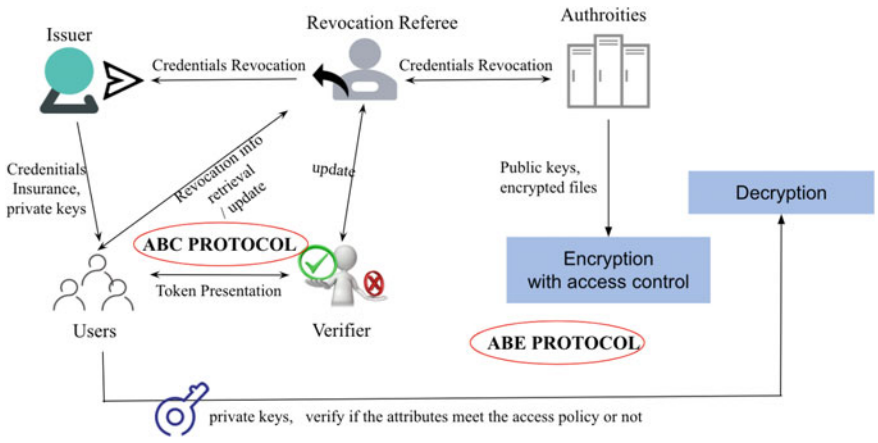


Fig. 2 Entities and interactions diagram

Referee will update users. Meanwhile, Verifier validates the User's attributes again by cooperating with the Revocation Referee. Users' updating attributes are gotten ensured in this round way. On the other hand, based on attributes and private keys, Users will be verified if it is eligible or not to decrypt the encrypted file by following ABE decryption protocol policies.

- **Issuer (I)**: For one thing, Issuer issues attributes (credentials) and distributes private keys to Users. And for another, it also cooperates with the Revocation Referee for revocation and dynamic update.
- **Revocation Referee (RR)**: This entity responses to revoke issued attributes. It is enviable to update users' attributes in a big-data-driven society and cloud environment. It concerns privacy protection, and attributes after revocation can not be used. Issuer and Verifier will work together with Revocation Referee to update Users' credentials to protect Users' legitimate rights.
- **Verifier (V)**: Verifier provides a service to the Users. As a relying party or service provider, it checks that the credentials or attributes are valid. If the revocation is supported, Verifiers update the Users' credentials accordingly with the Revocation Referee.
- **Authorities (A)**: It belongs to a portion of ABE protocol which we have already mentioned in the previous chapter. They have also been named Attribute Authorities (AA). We have multiple attribute authorities to manage different attributes. For example, Attribute Authorities $a.$ represents "name" while Attribute Authorities $b.$ represents "school" if we describe students by different attributes. Different AAs do not communicate with each other. There are no overlapping responsibilities among them and different authorities' responses to various attributes. Authorities also respond to generate system public keys, system master keys to encrypt sensitive data.

4.2 Model Security Features

Here we clarify the security properties of our model. Later we will prove them in the last section.

Users can get access to the registered resources data only when its attributes meet the requirements of access policy A_{c-cp}. This MA-CP-ABE algorithm is built on the top of Water's algorithm [23]. This security proof is simulated by a challenger Ch and an adversary Ad. This simulation game is considered similar to [1]:

Setup phase: The challenger runs the *System Setup* algorithm and sends the generated public parameters, public key from the authorities to the adversary.

Secret key queries phase: The adversary repeat the secret key queries as many as possible corresponding to the sets of attributes A_{c-cp_1},...A_{c-cp_n}. Then the challenger responses the corresponding secret keys to the adversary.

Challenge phase: The adversary gives two equal length messages M_0 and M_1. Meanwhile, it gives a challenge to access structure A^* such that none of the sets of attributes from *Secrete key queries phase* satisfy the access structure. Then, the

challenger flips a random coin $b \in 0, 1$, encrypts the respective message M_b under the chosen access structure A^*, and sends the ciphertext to the adversary.

More secret key queries phase: *Secret key queries phase* is repeated with the constriction that the attribute sets cannot satisfy the access structure A^*.

Guess: The adversary outputs a guess b' of b.

We claim that the adversary Ad is successful if the above guess is correct, i.e., $b = b'$. We define the advantage of an adversary in this game as $Pr[b' = b] - \frac{1}{2}$.

Definition 2 A multi-authority ciphertext-policy attribute-based encryption scheme is secure if all PPT adversaries have at most negligible advantage in the above simulation game.

4.3 Protocol

Here we describe the main framework of the protocol. Assume that some attributes are updated. Hence some secret keys should be updated. After the key update, the messages will be encrypted. Additionally, we assume that there exists an ABC protocol generating ABC token t_S for Users based on their attributes such that the Verifier can verify the tokens. Additionally, we assume that there exists a CP-ABE scheme $\Pi = (Setup, Encrypt, KeyGen, Decrypt)$ and the goal is to modify that scheme to achieve security after key revocation in the multi-authority setting. The more precisely proposed protocol Π^* is the following:

a. $Setup^*(\lambda, U) = Setup(\lambda, U) \rightarrow (PK, MSK)$: This step runs in the multiple authorities sides. It takes the universe U of the attributes, and the security parameter λ as inputs and outputs the public initialization parameters PK and the main system private key (or master key) MSK.
b. $Encrypt^*(m, \mathcal{A}, PK) \rightarrow CT$: Based on the message m, it takes the updated access structure describing the revoked attributes $\mathcal{A}$ over U and the public parameters PK as inputs and outputs a ciphertext CT. In the CP-ABE mode, we claim that CT contains the description of $\mathcal{A}$ itself and only users with attributes satisfying $\mathcal{A}$ are able to decrypt that.
c. $KeyGen^*(S, t_S, MSK) \rightarrow SK$: This algorithm runs in the multiple authorities sides as well. First, the authority runs $KeyGen(S, MSK)$ resulting in updated private key SK. Next, the authority verifies the validity of t_S and outputs an updated private key SK only if it is valid. Note that this point is the main difference between the original and the proposed method.
d. $Decrypt^*(PK, CT, SK) = Decrypt(PK, CT, SK) \rightarrow m$: A ciphertext CT (containing $\mathcal{A}$), the public parameters PK, and a secret key SK for an attribute set S are taken as input, and outputs an element of the message space. We assume that the output is the original m if S satisfies $\mathcal{A}$.

As shown in Fig. 3, we can find the interactions flow in different entities. There is a call of ABCE (attribute-based credential engine) to obtain the mechanism-specific

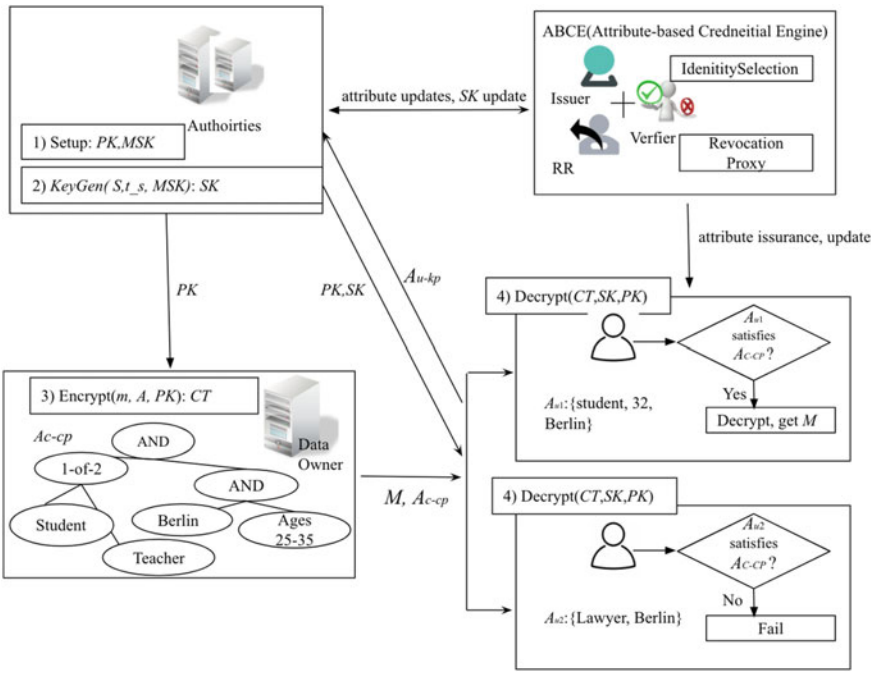


Fig. 3 Protocol

cryptographic data in the Users' sides. To be more specific, the $IdentitySelection$ from $Issuer$ and $Verier$ distributes attributes to $Users$ and provides different user interfaces to support a User in choosing a preferred credential A_{u-cp}, since there are different combinations of pseudonyms to satisfy a given access policy. In this process, Users generate the secret keys and credentials. In our case, there are two Users, A_{u1} represents User1 which has the credentials of $\{student, 32, Berlin\}$. A_{u2} represents User2 which has the credentials of $\{Lawyer, Berlin\}$. The $RevocationProxy$ together with $RevocationReferee$ deals with the generation or presentation of tokens/credentials (t_s) subjected to Revocation in the update of Users' attributes.

Moreover, the authorities generate the public keys. The data owner is the access policymaker. Here we have the access policy A_{c-cp}, and only if the user satisfies this access policy he can get access to sensitive resources. Here our access policy A_{c-cp} specifies that a User should be a $Student$ or $Teacher$, located in Berlin in the age of 25–35. By comparing A_{u-cp} and A_{c-cp}, it verifies whether the User can decrypt the data or not.

5 Security Analysis

In this part, we will prove the security of our protocol from the algorithm itself and prevention of collusion attacks.

5.1 Security

Theorem 1 *Let* Π *be a secure CP-ABE,* Π^* *be the construction above, and assume that every input token* t_S *in* $KeyGen^*$ *are generated by an ABC satisfying soundness. The protocol* Π^* *is secure after key revocation.*

Proof Let $CT = Encrypt(M_b, \mathcal{A}, PK)$ be a ciphertext and let Ad be a PPT adversary with $A_{Ad} \notin \mathcal{A}$. Then Ad can get the respective secret key SK with negligible probability only since, as a consequence of soundness of the ABC, she can present a valid token t_{Ad} with negligible probability only. Then Ad can guess b from the inputs CT, PK, A_M with negligible probability only from the security of ABE. This completes the proof.

5.2 On Collusion Attacks

The traditional single-server attribute encryption schemes are vulnerable to collusion attacks through key randomization. However, our scheme can efficiently prevent both accessor collusion and authority collusion in cloud environment and IoT devices.

Authority collusion denotes that any number of authorities may try to access the confidential data even if they are not qualified. However, in our scenario, different authorities do not communicate with each other; thus, it is impossible to launch authority attack.

Accessor collusion means different clients might combine their attributes together in order to get access to some inaccessible resources. Our applied scheme can efficiently prevent accessor collusion attack. Firstly, for each authority, there is a random nonce during the system setup and key generation. In order to decrypt a ciphertext, it requires the same random value in client's private keys. Thus, different accessors cannot integrate their private keys to strengthen their decryption power apart from some negligible probability events. Secondly, in ABC protocol, users can prove their identity (credentials) in an unlinkable and anonymous way, which makes covert decryption impossible for collusion attacks.

6 Conclusion

In conclusion, we propose a CP-ABE scheme without the trusted central institution, and an attribute-based credential engine shares the stress of attribute issuance and revocation, which maintains the data privacy and even improves users' identity privacy. As a result, it provides a privacy-friendly authentication with efficient attributes and user revocation.

Future work: It contains a proof-of-concept protocol only. In the future, it is necessary to formally validate in the practical aspect, such as in the IoT devices, electronic health record systems.

References

1. J. Bethencourt, A. Sahai, B. Waters, Ciphertext-policy attribute-based encryption, in *2007 IEEE Symposium on Security and Privacy (SP'07)* (IEEE, 2007), pp. 321–334
2. S. Brands, *Rethinking Public Key Infrastructures and Digital Certificates: Building in Privacy* (MIT Press, 2000)
3. J. Brown, P. Stradling, C.H. Wittenberg, U-prove ctp r2 whitepaper. Technical report (Microsoft Corporation, 2011)
4. V. Božović, D. Socek, R. Steinwandt, V.I. Villányi, Multi-authority attribute-based encryption with honest-but-curious central authority. Int. J. Comput. Math. **89**(3), 268–283 (2009)
5. J. Camenisch, *Specification of the Identity Mixer Cryptographic Library* (IBM Research-Zuric, 2010), pp. 1-48
6. J. Camenisch, Concepts around privacy-preserving attribute-based credentials, in *IFIP PrimeLife International Summer School on Privacy and Identity Management for Life* (Springer, Berlin, 2013), pp. 53–63
7. J. Camenisch, A. Lysyanskaya, An efficient system for non-transferable anonymous credentials with optional anonymity revocation, in *International Conference on the Theory and Applications of Cryptographic Techniques* (Springer, Berlin, 2001), pp. 93–118
8. J. Camenisch, S. Fischer-Hübner, K. Rannenberg, (eds.), *Privacy and Identity Management for Life* (Springer Science & Business Media, 2011)
9. M. Chase, S.S. Chow, Improving privacy and security in multi-authority attribute-based encryption, in *Proceedings of the 16th ACM Conference on Computer and Communications Security* (2009), pp. 121–130
10. M. Chase, Multi-authority attribute based encryption, in *Theory of Cryptography Conference* (Springer, Berlin, 2007), pp. 515–534
11. D. Chaum, Security without identification: transaction systems to make big brother obsolete. Commun. ACM **28**(10), 1030–1044 (1985)
12. V. Goyal, O. Pandey, A. Sahai, B. Waters, Attribute-based encryption for fine-grained access control of encrypted data, in *Proceedings of the 13th ACM Conference on Computer and Communications Security* (2006), pp. 89–98
13. H. Lin, Z. Cao, X. Liang, J. Shao, Secure threshold multi authority attribute based encryption without a central authority, in *International Conference on Cryptology in India* (Springer, Berlin, 2008), pp. 426–436
14. J. Hur, D. Koo, S.O. Hwang, K. Kang, Removing escrow from ciphertext policy attribute-based encryption. Comput. Math. Appl. **65**(9), 1310–1317 (2013). ISSN 0898-1221, https://doi.org/10.1016/j.camwa.2012.02.005

15. D.S. Kasunde, A.A. Manjrekar, Verification of multi-owner shared data with collusion resistant user revocation in cloud, in *2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT)* (IEEE, 2016), pp. 182–185
16. M. Koning, P. Korenhof, G. Alpár, J.H. Hoepman, The ABC of ABC: an analysis of attribute-based credentials in the light of data protection, privacy and identity (2014)
17. K. Rannenberg, J. Camenisch, A. Sabouri, *Attribute-Based Credentials for Trust* (Identity in the Information Society, Springer, 2015)
18. R. Gennaro, S. Jarecki, H. Krawczyk, T. Rabin, Secure distributed key generation for discrete-log based cryptosystems. J. Cryptol. **20**(1), 51–83 (2007)
19. A. Sahai, B. Waters, Fuzzy identity-based encryption, in *Annual International Conference on the Theory and Applications of Cryptographic Techniques* (Springer, Berlin, 2005), pp. 457–473
20. V.K.A. Sandor, Y. Lin, X. Li, F. Lin, S. Zhang, Efficient decentralized multi-authority attribute based encryption for mobile cloud data storage. J. Netw. Comput. Appl. **129**, 25–36 (2019)
21. Y. Yan, M.B. Kamel, P. Ligeti, Attribute-based encryption in cloud computing environment, in *2020 International Conference on Computing, Electronics and Communications Engineering (iCCECE)* (IEEE. 2020), pp. 63–68
22. K. Yang, X. Jia, Expressive, efficient, and revocable data access control for multi-authority cloud storage. IEEE Trans. Parallel Distrib. Syst. **25**(7), 1735–1744 (2013)
23. B. Waters, Ciphertext-policy attribute-based encryption: an expressive, efficient, and provably secure realization, in *International Workshop on Public Key Cryptography* (Springer, Berlin, 2011), pp. 53–70

Next Generation Wireless Communication: Facilitated by Machine Learning

Praveen Kumar Singh

Abstract Currently, the fifth generation (5G) Wireless Communication Networks (WCN) with enhanced capabilities are in the process of deployment all over the world. With ever increasing mobile traffic, they are likely to offer better coverage, security and spectrum efficacy in addition to more cost-effective infrastructural network. Emergence of Artificial Intelligence (AI), e-health, cognitive techniques, automation of industries and many other Machine Learning (ML) applications are likely to flourish with next generation WCN. This paper deals with a significant aspect of next generation WCN and how it will be facilitated by the ML [2–4]. Initially, the overview of both the WCN and ML has been given by bringing out the fundamentals of both the concepts in this paper which is followed by the brief outline about the noteworthy parameters facilitating the next generation mobile communication data networks. Further, it provides a brief account on the promising prospects in relation to future WCN applications. An added technological paradigm associated with these prospects even beyond 5G networks shall enable such applications that the limitations between the cyber and the physical domains are likely to be wiped out. A recommendation on deployment parameters of future generation WCN applications supported by ML has also been provided to evaluate the impact of this promising technology [21]. Lastly, this paper explores the future research possibilities wherein the ML can contribute in surmounting the limitations of such widespread mobile network deployments.

Keywords Communication · WCN · Cyber security · ML · Cellular · IoT · Spectrum · Data rate · Latency · MIMO

1 Introduction

The frequency bands in wireless communication have quite a large spectrum. While it commences from very low frequency (VLF) spectrum which covers 3–20 kHz and extends up to 5 GHz SHF scientific and industrial bands for varied commercial

P. K. Singh (✉)
Department of Computer Application, Shri Ramswaroop Memorial University, Lucknow-Deva Road, Barabanki, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_59

applications [1]. Presently, mobile communication of fifth generation (5G) wireless networks is taking place which is likely to offer better dividends in terms of enhanced coverage, data rates, quantum of communication systems to get connected, sizeable reduction of power consumption, etc. In addition to substantial traffic capacity of 10 Mbps/m^2, power consumption efficacy of around 100 times, data speed up to 20 Gbps, the end user latency of just 1 ms and likewise, there are numerous advantages which will be facilitated by the next generation wireless communication technology. The next generation wireless communication networks will facilitate promising Machine Learning (ML) and Artificial Intelligence (AI) technologies while confronting with a wide array of spectrum, large quantum of antennas, prevailing service needs and extensive heterogeneous networks [3]. Moreover, AI employs the distributed control mechanism demonstrated through a promising distribution of co-relation algorithms of datasets and eventually aggregated by cloud computing.

2 Overview of Machine Learning

Machine Learning (ML) allows computers to learn as well as progress on or after occurrences constantly. While we provide the machine additional data, it enables the associated algorithms to "learn and further develop against the outcomes. These algorithms learn through an iterative course. Data extraction through Machine Learning too has undergone for big transformation and also in the manner the data is interpreted by linked algorithms which change the conventional statistical methods [5]. The processing of ML has been illustrated in Fig. 1 below which is done in two different stages, In the first stage, raw training data is taken and prior to its processing, it is normalized by dimension reductions or image processing. Further, it undergoes different types of learning process and the error analysis is done. In the second stage

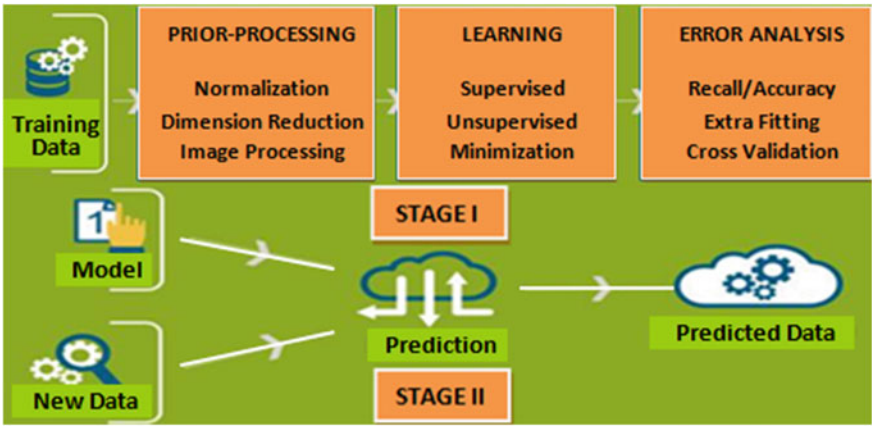


Fig. 1 Stages of data processing of machine learning (ML)

the new data under a particular model is undertaken and the prediction is as per the associated learning mechanism. There are various ML techniques with specific requirements to offer a definite outcome by employing data in varying forms. Some of the prominent ML techniques are discussed in succeeding paragraphs.

2.1 Supervised Learning

In supervised learning, for a given data set, there is an aim to learn a function for mapping. It can be categorized into support vector and the K Nearest Neighbors (KNN). Such classification depends upon the stability of data output. In support vector machine, a linear classifier form attempts to segregate an x-dimensional vector's data points by employing a $\times$ 1 hyper plane. Likewise, in KNN model refers to a receding pattern algorithm to classify it as regression in which no hypothesis in respect of distribution of data is warranted [11]. The fundamental principle of KNN pertains to select a test point akin to a feature vector depending upon the choices of its K neighbors in the vicinity as illustrated below in Fig. 2. Assuming that the data chosen for the training purpose indicates a repetitive class that can influence the test data prediction, a weigh methodology may be incorporated for each vector in vicinity that are in inverse proportion from the test point.

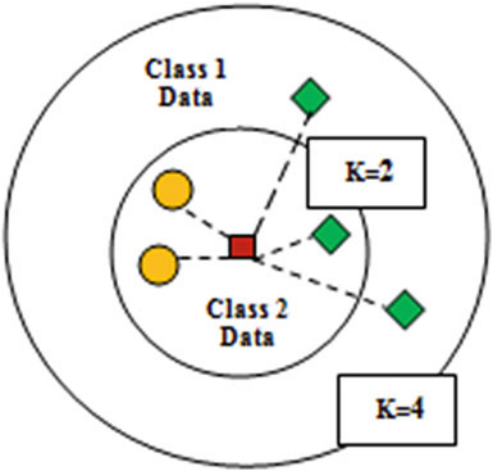


Fig. 2 A fundamental KNN model

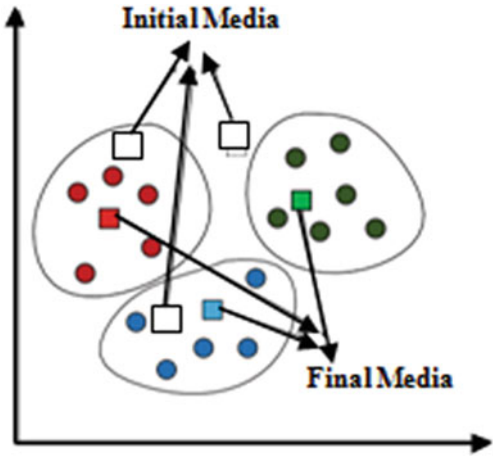


Fig. 3 A basic K-media model

2.2 Unsupervised Learning

In this ML mechanism, there is an attempt to learn a form to elaborate a veiled configuration from an unmarked data. Different forms of unsupervised leanings include an algorithm of K-media clustering and the examination of the principle vector component. A basic K-media model is illustrated above in Fig. 3. In K-media clustering, there is an attempt to divide K clusters out of n number of data points wherein nearest mean cluster represents against every data point. An iterative sophistication indicates the K-media algorithm that may be quite familiar. Initially, it begins with random trigger, thereafter one cluster is allocated by every data point and the mean of the cluster value is recorded [14]. In principal component examination, there is a reduction of variables which are quite large in size. It is represented mathematically and commonly known as dimensional reduction apparatus in which a separate coordinated structure is transformed through the associated data.

2.3 Reinforcement Learning

In this ML method, there is an attempt to define a long-term aim through trials as well as errors by associating self to the surroundings. The most common method relates to Q-learning in which the Q values are learnt and referred to as a discounted amassed incentive by adhering to a distinct policy. There is another form of reinforce learning known as Multiple Armed Bandit (MAB) learning in which there is a sequential action through proportionate distribution in an attempt to gain largest incentive [6]. In this method, there is a tradeoff between the maximum possible future incentive (exploration) and the best feasible action undertaken concurrently as illustrated in

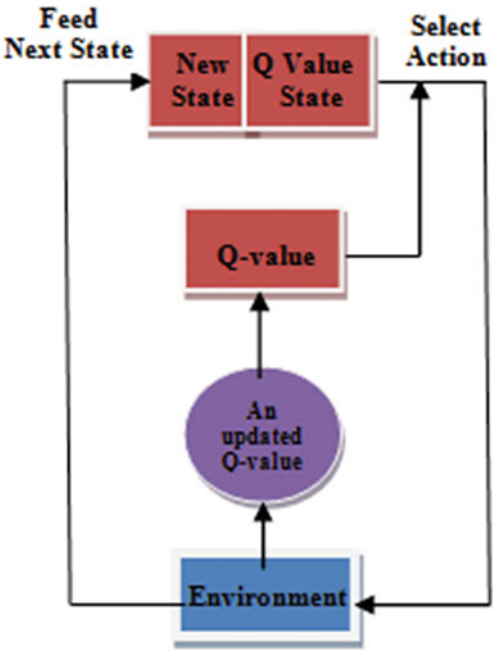


Fig. 4 The Q-learning methodology

Figs. 4 and 5 below through Q-learning and MAB learning methodologies. In this form, a state of equilibrium or a steady state is likely to be accomplished. In addition, the other prominent reinforce learning methods include actor critic, strategy evaluation and joint assessment method and the deep reinforcement method.

2.4 Dense Neural Network

The fundamental constituent in this ML method is a neuron in which the weight corresponds to the input value whereas an activation occupation refers to the output. The conversion of input is through different network layers wherein two consecutive levels have no direct co-relation. There are two major neural networks known as recurrent and convolution. In the recurrent neural network, a concealed layer is inherent in the network structure which is visible afterwards in the output and it remembers the echo state networks. In the convolution neural network, two key levels refer to convolution as well as pooling activities [9]. Additionally, there are auto encoders and intense learning machine methodologies too which represent the dense neural networks.

Fig. 5 The MAB learning methodology

3 Machine Learning Facilitating Future Generation Wireless Communication Networks

With evolving ML, there is a realization that an abundance of opportunities lies for advancement of future generation communication networks. The huge amount of information concealed within the data can be extracted by resolving complex models/algorithms through ML. It is widely considered that ML or AI-driven data can facilitate enormously to control, automation and data analysis for future networks. However, the strength of prevalent ML needs to be duly analyzed to make it applicable for a particular application [10]. There are primarily two major types of services in this context which are supposed to be examined. The first is augmented mobile broadband and the second can be referred to as low latency ultra reliable communication. These services are briefly outlined in succeeding paragraph.

3.1 Augmented Mobile Broadband (AMBB)

Augmentation of the mobile broadband services can meet the larger data speeds requirements for upcoming applications over a standardized areas covered by such applications. Virtual reality, HD video streaming, complex data computations, etc., are the services which can be duly supported by AMBB. There will be certain parameters requirements which need to be met for AMBB services and are enumerated as follows:

3.1.1 Maximum Data Rate

In order to facilitate higher data speed, max data rate for each associated applications are planned out. With growing technological advancements, existing technological up-gradations too demand higher data speeds in future. For example, IMT services running for fourth generation services at 1 Gb/s need to be upgraded at 20 Gb/s for the fifth generation services [7]. The applications like Multiple Input Multiple Output (MIMO), centimeter or milimeter waves will need much higher data speeds with better spectrum efficiency to support their services in future [8].

3.1.2 Network Energy Efficacy

It is significant for an augmented mobile broadband and is likely to increase almost 100 times for future generation wireless networks in comparison to existing 4G applications in India. There will be number of new services generated by future applications which will warrant for huge power consumption. It necessitates to accomplish the desired level of energy efficacy with future services and likely to be a very challenging affair. If the required amount of energy efficacy is not realized then the available network may not be exploited for the existing traffic capacity. A solution lies in putting the sensor nodes of MTC applications in sleep mode when not being utilized. As there will only be a definite level of energy available at sensor nodes and may not be rechargeable, therefore provisioning of carefully scheduling of sleeps and awareness modes is extremely important [12–15].

3.1.3 Spectrum Efficacy

For an ideal future generation wireless systems with IMT applications, the maximum uplink spectral efficacy of 15 bit/s/Hz and the downlink of 30 bit/s/Hz spectral efficacy is desired. It indicates that under ideal circumstances, the max data speed to be channelized through the individual bandwidth of channel in bit/s/Hz. In 5G itself, availability of spectrum is likely to increase up to 30 G Hz from existing 3 G Hz. Therefore, in order to optimally utilize the available spectrum for enhanced spectral efficacy, instead dedicating it for an individual access, multifold access needs to be permitted in the network [13]. In the new 3.5 G Hz citizen bands for broadband radio services, the apprehensions of licensing issues have been addressed by incorporating Q-learning ML for dynamic spectrum allocation. The shared bandwidth arrangement facilitates better spectrum efficacy and the negative impacts can be mitigated.

3.2 *Low Latency Ultra Reliable Communication (LLURC)*

There are a number of services which will be offered through new generation wireless communication networks. These services will range from healthcare, public transportation, academics, business transactions, telecommunications, power generation, environment control and many other such sectors which have direct influence over the public lives. Such services will require quite enlarged bandwidth to accommodate the technical advancements of future generation communication networks. To name a few areas bullet train connectivity's with rapid speeds, autonomous driving, end-to-end mobile vehicle communication, mission-oriented applications, smart applications and remote medical surgeries [14]. Figure 6 above illustrates some of the services. These kinds of applications will need better mobility, high reliability, enhanced data speeds and the low latency. Two critical needs of low latency and relatively better mobility are being discussed as follows:

3.2.1 Low Latency

The performance efficiency of any data network can be measured through the level of latency. Lower the latency level, higher is the consideration of efficiency of that

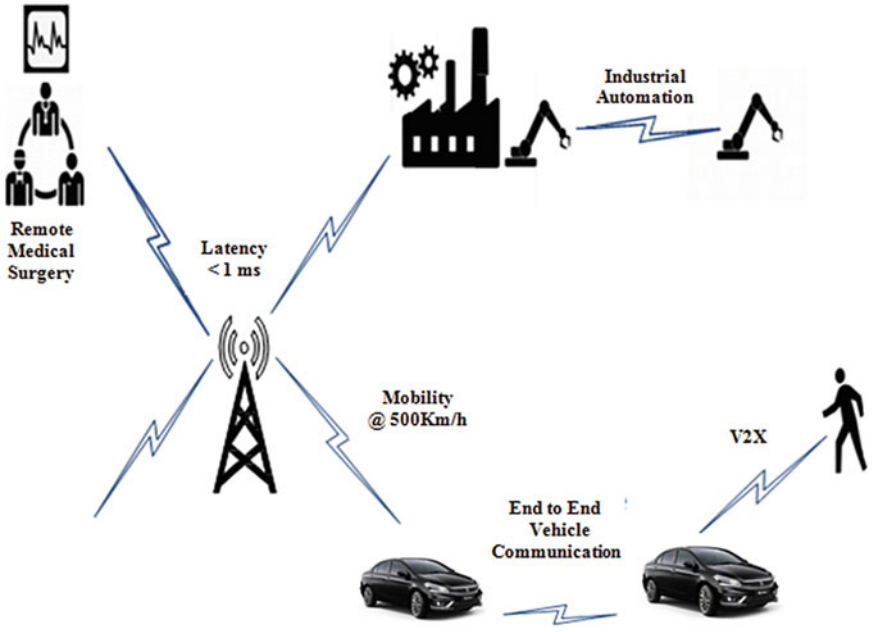


Fig. 6 Low latency ultra reliable communication (LLURC) services for future generation wireless communication networks

network. Therefore, all future generation wireless communication networks will demand extremely low latency for their efficient employments. Existing 4G services possess 10 ms latency level and even with rolling out of 5G services in India are likely to witness its reduction at just 1 ms level. The LLURC will largely be dependent upon the better anticipation of the network requirements and its more efficient management [13]. The cloud servers may not need to keep all the data which may in turn increase the system latency of the associated applications. Therefore, it is a mandatory requirement of the nearest base station (BS) to keep the desired level of information to ensure low system latency. Since BSs may not have sufficient storage capacity, so in order to predict the user needs, ML-based algorithms may be employed.

3.2.2 Enhanced Mobility

Mobility is referred to the maximum feasible speed of a mobile station along with desired level of quality of service (QoS) which can be accomplished by the particular application. For the future generation mobile vehicle, it is considered that there will be a requirement to achieve vehicular mobility up to 500 km/h instead of 350 km/h as envisaged in 4G services. Since such high mobile applications of future generation networks are likely to operate in a very dense mm wave environment in which user will need switching over among BSs frequently, therefore an appropriate beam forming vector has to be identified for that application [17]. Such applications will warrant a suitable mechanism to employ ML to hand over control to the clusters of centralized controllers.

4 Promising Prospects for Future Generation Wireless Communication Networks

Currently, though 5G has already seen some of its active service applications in USA, China, South Korea and Europe, India and some other nations all over the globe are in the process of rolling out. Meanwhile, keeping in view of the enormous prospects in future, a good research work has already commenced to bring out the nuisances of sixth generation (6G) applications and its future requirements. In a likelihood of extreme complex wireless communication environment, the latency level for 6G services are expected to further reduce on the tune of 0.1 ms. In order to ensure appropriate level of equipments density, incorporating 3D features in the associated applications will become quite significant [18]. The 6G is expected of having a device density of 100 machines per m^3 as compared to just 10 machines per m^3 in 5G services.

In order to achieve such capabilities, it will be crucial that future networks can examine the network usage and accurately learn environment uncertainties and

should also be able to hold capacity of self-maintenance of the associated data network. ML will be the key in supporting future generation services by facilitating the data networks to be dynamic, self-managing and to become relatively more autonomous. There is also a need to keep a track of all emerging technologies of big data, machine learning (ML), artificial intelligence (AI), cloud computing, Internet of Things (IoT), block chain, etc., which can contribute for maximum exploitation of the potential services of next generation wireless communication networks. In a dynamic environment, the spectrum will be the key in future and are likely to be used on shared basis by employing its cognitive properties. In order to predict its precise availability, ML will play a vital role in successful implementation of the associated services.

5 Next Generation Wireless Applications Deployment Challenges by Inclusion of Machine Learning

Inclusion of ML to 5G and subsequent next generation wireless communication technologies though quite handy, however the associated challenges need thorough examination. The issues may require some serious deliberation before we incorporate the ML algorithms in our applications. Before we proceed further, let us have a look on relative comparison of ongoing 4G (over most part of the globe), 5G (running only in USA, China, South Korea, etc.) and other next generation (6G and 7G) wireless mobile communication technologies as illustrated in Table 1 below. The

Table 1 Relative comparison of different Cellular Communication Technologies

Parameters	Cellular technologies			
	4G	5G	6G	7G
Service	VoIP, LTE, Wi-Fi, WiMAX	WWWW	Global and secured cellular services	Global and secured cellular services
Frequency bands	2–8 GHz	4G Frequency	95 GHz-3THz	95 GHz-3THz
Core network	Internet	Internet	Internet	Internet
Multiplexing	OFDMA	MIMO, AI capabilities, CDMA	CDMA	CDMA
Data rate	100–300 Mbps	Approx 100 + Mbps	Approx 11 Gbps	Approx 11 + Gbps
Switching type	Packet switching	IPv6 but advancements under progress	All packet	All packet
Time period for Implementation	Current	Under progress	Probably by 2030	Probably by 2030

comparison indicates about wide reach of future 6G and 7G future cellular services over a large spectrum from 95 GHz-3THz. The major likely dividend from those future generation wireless communication technologies will relate to tremendously enhanced bandwidth upto 11 Gbps and even more. That is the most noteworthy gain we are going to accrue out of these future technologies. Though the comprehensive details may not be feasible in this paper, however a brief outline of some of the prominent challenges in this respect is enumerated in succeeding paragraphs.

5.1 The Data Criticality

For all Machine Learning applications, data has major role to play in information sharing over data communication network. The type of different data chosen has implications on suitability of leanings required by the particular application in 5G and other future generation data networks. Availability of real data sets is extremely important for the ML applications in wireless communication [21]. Generating the data sets from the simulators may not be helpful in this case as ML applications pick up real-time data to perform its functions whereas simulators release a pre-fed programmed data. That is where a domain expertise with featured engineering will become a mandatory requirement to correctly represent the data sets with future generation wireless application which will be supported by ML applications.

5.2 No Idealistic Theorem

When all feasible data sets generated in each ML algorithm are collectively examined, we find that after averaging the entire datasets, there is hardly any variation in the overall relative performance in all the available datasets. It signifies that all the ML algorithms need to be treated with equal importance and no single algorithm can be considered better than compared to other ML algorithms. It also indicates that we should not look for any ideal ML algorithm with expectation to have an absolute solution for the given application. Rather, we have to find out that as per the specific associated application, the relevant distribution of data has been taken into consideration while choosing the appropriate ML algorithm because it'll have serious implications over the performance of the services offered by the application [23].

5.3 Precision and Interpretability Tradeoff

In the ongoing fifth generation mobile communication networks deployment, it has been observed that a particular Base Station (BS) may allocate additional resources to a subscriber compared to others during the implementation of an ML algorithm.

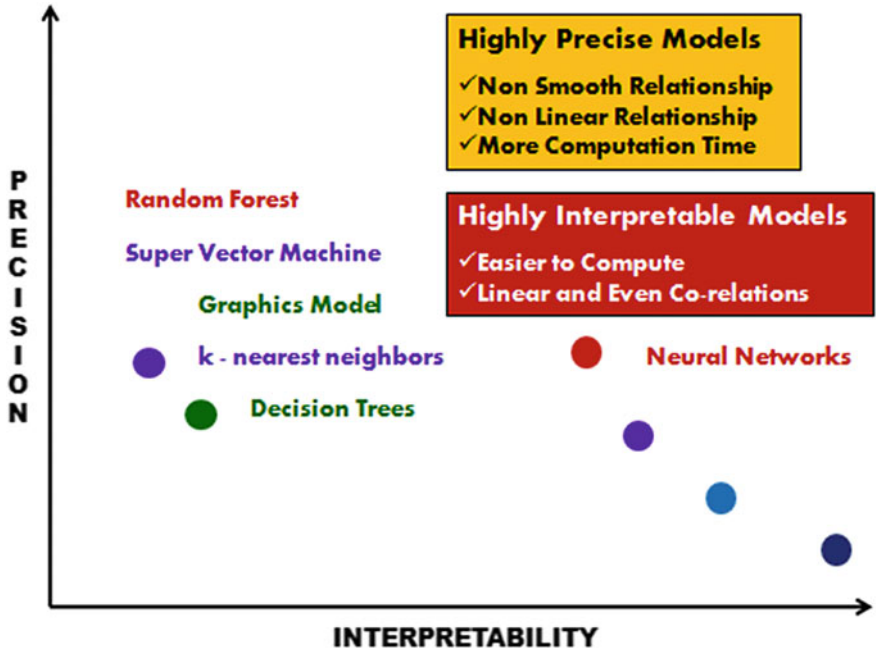


Fig. 7 The Precision and Interpretability tradeoff in an ML algorithm

Such complexities are still under examination and may not stand to any logic. In order to comprehend why a particular model is more appropriate to a peculiar state of affairs and how a certain ML algorithm is chosen, a deep understanding in respect of interpretability and precision tradeoff is required. It has also been illustrated in Fig. 7. Aim should be to strike a balance chord between precision and the interpretability in the subject model [11]. The functionality of Deep Neural Network (DNN) ML model is a little peculiar wherein it is must that the various concealed layers structure and the node activation methodology to be understood.

5.4 Privacy and Security

The ML capability to quickly negotiate against rapid changing environmental data has facilitated it to contribute as a stringent tool against intrusion detection and cyber security violations in wireless communication networks. However, such adaptability may be counterproductive at times with some unexpected outcomes. For example, a deliberately initiated malicious noise signal may mislead to the associated CNN. Another important issue with ML algorithms may relate with the adversarial cyber attack against it. The solution lies in provisioning of a well tested and trained mechanism to ensure adequate safeguard for privacy and security of the system [12].

Certain other measures in this respect which can be effective are establishing the rotation of algorithm key, regulation and the proper authentication. In a wireless communication network where even a slight variation or error in signal may lead to catastrophic results, formidable efforts will be required by all stakeholders to ensure its due privacy and security.

6 Future Research Scope

While analyzing the prevalent and future generation wireless mobile communication technologies, in order to explore the future research possibilities, let us examine the key issues where there is a scope for improvement in these technologies. There are number of shortcomings in widespread fourth generation (4G), however a definite need to identify feasibility of progress in 5G and subsequent future generation mobile wireless communication technology exist. In 5G services, the endeavor has been for lower power consumption and reduced suspension. In addition, IP addresses are allocated by IPv6 to mobile nodes randomly based on the location management. It leads to wastage of some precious resources that may be looked into for future research. In sixth generation (6G), roaming and the hands off would be the major issue primarily because of varying network standards likely to be adopted by different satellite systems.

The major hurdle which may haunt the implementation of all mobile future generation communication technologies is subjected to non-compatibility of prevalent devices or handsets. These existing devices will be in all likelihood being obsolete and will be incompetent to support the desired next generation services. It will warrant replacing these existing handsets at mass level and need to manufacture the new handsets amass to support the future generation services. It implies in creation of big hardware infrastructure for mass level production at global level. The key barrier is the expenditure incurred in creating such infrastructure. In general, cost of infrastructure creation or expansion of 5G and subsequent mobile future generation communication technologies is very high and the sour issue is that their privacy and security issues still need to be confronted to resolve.

7 Conclusion

The roll out of fifth generation wireless mobile communication services is currently happening in most parts of the world. The nations like USA, China and South Korea have already started to offer these services to their citizens. In order to accomplish the uniformity in all subsequent generation (6G, 7G, 7.5G and 8G) mobile communication technologies, a unison global technical standard will be required. 5G is being considered as a genuine wireless paradigm wherein 6G is expected to integrate satellite networks with 5G. 7G wireless networks are likely to target for space

roaming. Different trials and research are already under progress to accomplish the desired set of goals worldwide [7]. In addition, there is enough assurance which can be extended by the Machine Learning (ML) to support the future generation mobile wireless communication services. There are large numbers of dividends facilitated by ML which include mitigation interference over mobile relay or nomadic nodes, precise predictions, pattern recognition or feature extraction, etc. In addition, mobile wireless communication management by ML is likely to open new horizons in URLL networking [15].

References

1. M. Chen, U. Challita, W. Saad, C. Yin, M. Debbah, Machine learning for wireless networks with artificial intelligence: a tutorial on neural networks, arXiv:1710.02913v1, Accessed 21 April 2021
2. Z.H. Huang, X. Xu, H.H. Zhu, M.C. Zhou. An efficient group recommendation model with multi-attention based neural networks. IEEE Trans, Neural Netw. Learn. Syst. **31**(16), 4461–4474 (2020)
3. G. Cao, Z. Lu, X. Wen, T. Lei, Z. Hu, AIF: an artificial intelligence framework for smart wireless network management. IEEE Commun. Lett **22**(2), 400–403 (2018)
4. W. Lee, M. Kim, D. Cho, Deep power control: transmit power control scheme based on convolution neural network. IEEE Commun. Lett. **22**(6), 1276–1279 (2018)
5. Y. Sun, W. Jiang, G. Feng, P.V. Klaine, L. Zhang, M.A. Imran, Y.C. Liang, Efficient handover mechanism for radio access network slicing by exploiting distributed learning. IEEE Trans. Netw. Serv. Manag. **17**, 2620–2633 (2020)
6. P.V. Klaine, M.A. Imran, O. Onireti, R.D. Souza, A survey of machine learning techniques applied to self-organizing cellular networks. IEEE Commun. Surv. Tuts. **19**(4), 2392–2431 (2017)
7. Y. He, N. Zhao, H. Yin, Integrated networking, caching and computing for connected vehicles: a deep reinforcement learning approach. IEEE Trans. Veh. Technol. **67**(1), 44–55 (2018)
8. C. Zhong, M. Gursoy, S. Velipasalar, A deep reinforcement learning-based framework for content caching, in *Proceedings of CISS* (Princeton, USA, 2018), pp. 1–6
9. Z.H. Huang, X. Xu, H.H. Zhu, M.C. Zhou, An efficient group recommendation model with multi-attention-based neural networks. IEEE Trans. Neural Netw. Learn. Syst. **31**, 4461–4474 (2020)
10. M. Chen, W. Saad, C. Yin, M. Debbah, Echo state networks for proactive caching in cloud-based radio access networks with mobile users. IEEE Trans. Wireless Commun. **16**(6), 3520–3535 (2017)
11. Y. Sun, M. Peng, S. Mao, Deep reinforcement learning based mode selection and resource management for green fog radio access networks. IEEE Internet Things J. **6**(2), 1960–1971 (2019)
12. G. Zhao, S. Qin, G. Feng, Y. Sun, Network slice selection in softwarization-based mobile networks. Trans. Emerg. Telecommun. Technol. **31**(1), e3617 (2020)
13. M. Feng, S. Mao, Dealing with limited backhaul capacity in millimeter-wave systems: a deep reinforcement learning approach. IEEE Commun. Mag. **57**(3), 50–55 (2019)
14. S. Han, C.I, G. Li, S. Wang, Q. Sun, Big data enabled mobile network design for 5G and beyond. IEEE Commun. Mag. **55**(9), 150–157 (2017)
15. Y. Yu, T. Wang, S.C. Liew, Deep-reinforcement learning multiple access for heterogeneous wireless networks. IEEE J. Sel. Areas Commun. **37**(6), 1277–1290 (2019)
16. K. David, H. Berndt, 6G vision and requirements. IEEE Veh. Technol. Mag. **13**(3), 72–80 (2018). https://doi.org/10.1109/MVT.2018.2848498

17. P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, D. Meger, Deep reinforcement learning that matters, in *Proceedings of the 32nd AAAI Conference on Artificial, Intelligence* (2018), pp. 1–26
18. W. Saad, M. Bennis, M. Chen, A vision of 6G wireless systems: applications, trends, technologies, and open research problems. IEEE Netw. **34**, 134–142 (2020)
19. S.Z. Chen, Y.C. Liang, S.H. Sun et al., Vision, requirements, and technology trend of 6G: how to tackle the challenges of system coverage, capacity, user data-rate and movement speed. IEEE Wireless Commun. **27**, 218–228 (2020)
20. Z.Q. Zhang, Y. Xiao, Z. Ma et al., 6G wireless networks: vision, requirements, architecture, and key technologies. IEEE Veh. Technol. Mag. **14**, 28–41 (2019)
21. L. Zhang, Y.C. Liang, D. Niyato, 6G visions: mobile ultra-broadband, super internet-of-things, and artificial intelligence. China Commun. **16**, 1–14 (2019)
22. S. Sharma, S.J. Darak, A. Srivastava, Energy saving in heterogeneous cellular network via transfer reinforcement learning based policy, in *Proceedings of COMSNETS* (Bangalore, India, 2017), pp. 397–398
23. Y. Feng, M.H. Wang, D.M. Wang et al., Low complexity iterative detection for a large-scale distributed MIMO prototyping system, in *Proceedings of IEEE International Conference on Communications, Shanghai*, vol. 64(6) (2019), pp. 1–6
24. Z. Chen, D. Chen, Y. Zhang, X. Cheng, M. Zhang, C. Wu, Deep learning for autonomous ship-oriented small ship detection. Saf. Sci. **130**(12), 797–812 (2020)

Simulation-Based Method for Analyzing Timing Attack Against Pass-Code Breaking System

Shaminder Kaur, Balwinder Singh, and Lipika Gupta

Abstract Rate of diversification and evolution among various embedded and IOT devices have brought new concerns and challenges in the field of security of embedded devices. Not only do these devices need protection against software attacks or network attacks but equal attention needs to be paid against emerging type of attacks known as hardware attacks. Hardware attacks also known as side channel attacks extract the secret information through side channels which act as leakage parameters and convey information such as power emanation of devices, EM emanation, etc. Least studied timing attack needs attention due to its trait of less expertise and easy conduction. In this work methodology is proposed for Simple (SPA) and Differential Power Analysis (DPA) against evaluation and testing of timing attack for pass-code key breaker. A model depicting timing attack as real-time scenario which captures timing information as a side channel leakage parameter against microcontroller is presented. We also presented assembly code (*.asm*) for generation and evaluation of timing attack. We conclude our paper by presenting a model of timing attack that can be modeled practically for deeply investigating and analyzing experimental results against timing attack. After performing over several timing-based iterations we were able to recover secret key with 100% match.

Keywords Side channel attacks (SCA) · Simple power analysis (SPA) · Differential power analysis (DPA) · Timing attack · Hardware Security

S. Kaur (✉) · L. Gupta
School of Engineering and Technology, Chitkara University, Himachal Pradesh, India

L. Gupta
e-mail: lipika.gupta@chitkarauniversity.edu.in

B. Singh
Centre for Development of Advanced Computing, Pune, India
e-mail: balwinder@cdac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_60

1 Introduction

Side channel attacks are extremely dangerous in breaking security of various cryptosystems. It is a known fact that every embedded device is vulnerable to side channel attacks. Adversary exploits side channel leakage information like power, temperature, EM emanation of embedded devices and successfully breaks the security of IOT and cryptosystems. Among these attacks, timing attacks have become a realistic cyber threat because they are non-invasive and can be launched remotely. Timings attacks are gaining popularity these day since such attacks generally target weak devices such as smart cards. Timings attacks are becoming more damaging attacks these days which exploits timing as leakage parameter to break the key. They enable the attacker to gather secret information based on the fact that there is a strong relationship between secret key and execution time of various cryptosystems. Cryptosystems often take slightly different amounts of time to process different inputs. In cryptography, a timing attack is a side-channel attack in which the attacker attempts to compromise a cryptosystem by analyzing the time taken to execute cryptographic algorithms. By carefully measuring the amount of time required to perform different operation attacker can easily steal secret information such as pass key of embedded devices. Existing study reveals that unintentional timing characteristics would only reveal a small amount of information from a cryptosystem (such as the Hamming weight of the key).

Our research unfolds a fact that timing attacks can be even more damaging than existing. Timing attacks if combined with an advanced technique known as cryptanalysis can increase the rate of timing information leakage which further can be more damaging to embedded devices. Timings attacks are mostly neglected during design phase which needs a special attention. There are various approaches for performing timing attacks: hardware, software and simulation based. Hardware- and software-based techniques are generally used upon availability of prototype. Also when system to be used is too large to be modeled and simulated at an acceptable cost. On the contrary, other technique called as "simulation based fault injection" proves to be potential and practical in doing exhaustive analysis of embedded products. Trait of this technique, i.e., non-availability of prototype can be exploited, thus allowing the analysis of practically any possible fault. One of the worries about recreation-based strategy is that it utilizes high CPU time to reenact the model of the framework given that it is accessible. Approach used by "*Hardware-controlled techniques*" is it uses external equipment to inject fault deep inside the hardware so that fault may propagates to the victim software. Related mechanism is observed by other platform (software platform) which is running on the hardware device. Contrary "*software controlled*" strategy utilizes a hazardous/destructive programming (i.e., shortcoming infusion and control programming), which runs on a similar equipment stage as the objective programming does. It furthers hampers the external conditions of the embedded hardware to produce faulty output. Another category is "simulation based approach." This paper uses simulation-based approach due to its several advantages: no need of prototype, less complex, less costly. Lastly it allows early

analysis of detection of side channel attacks. This trait of this technique makes it one of the best option for doing early analysis of attacks in order to build suitable countermeasures in a timely manner.

In this research work we have presented simulation-based real-time demonstration of timing attack which depicts how attacker can recover pass-code by exploiting timing information channels. We have demonstrated with an example of 4 bit pass-code key lock system where users need to enter correct key to access the hidden data. This experimental study will give upcoming engineers a platform to evaluate and analyze timing attacks deeply so that suitable countermeasures can be built against them in an efficient and timely manner.

This paper is organized as follows: Sect. 2 describes the use of simple power analysis as statistical method, Sect. 3 demonstrates timing attack model, Sect. 4 represents algorithm used for generating timing attack, Sects. 5 and 6 present results and conclusion.

2 Related Work

Various authors have either worked on software-based techniques or hardware-based techniques as discussed below. Limited authors worked on simulation-based approach which proves to be one of the best approaches in doing early analysis. Through this study we are the first one to investigate practical simulation-based approach on timing attacks against embedded devices. Table 2 presents advantages/disadvantages among various approaches used so far for generating side channel attacks. An attempt has been made to differentiate between various techniques based on factors such as: cost, performance, complexity, risk of damage, controllability, observe-ability, portability. Based on this information hardware engineers can decide which technique to experiment with in order to get suitable countermeasures against side channel attacks.

In [12] authors worked on software-based timing attacks which is cache-based timing attack that gathers information about timings or interrupt information. Remote attackers use the cache-based timing information and infer secret data such as typed words or sentences which can be dangerous. Authors have presented interrupt-based attack which is again software-based technique used for gathering timing information with an expertise required in handling software–hardware code interfacing.

In [13] authors studied and analyzed how RSA algorithm can be exploited remotely through cache-based timing attack. RSA algorithm is generally used online when there is exchange of digital signatures, etc. Security of cryptosystems can be broken remotely by attackers through side channels attacks. RSA algorithm uses two basic optimizations: Chinese Remainder Theorem (CRT) and a sliding window exponentiation. To perform such cache-based timing attacks related to GPU a theoretical foundation skill set is required for performing timing attacks in several cryptosystems which is again quite challenging and cumbersome.

In [14] authors proposed another timing attack based on assumption that numbers of extra reduction are known to the attacker. Timing attack proposed is not a practical approach. In [15, 16, 17, 18, 19] authors proposed interrupt-based timing attacks which are applicable only on software part. For such attacks to be successful adversary needs to have access to cycle accurate timing source. Also one needs to have thorough knowledge of JavaScript. In [20] timing attack attempt on cache-based memory has been demonstrated. Cache memory is used to increase the performance of processor since buffering data before it gets used and saves a lot of time.

Until now timing attacks have been applied in pretext to software codes. Also it has been observed that very few authors have worked upon an approach of extracting secret key through analyzing execution time. Experimental analysis performed indicates that there is strong relation between execution time and the inferring of secret key. Also most of the authors have worked on software techniques. Drawbacks of software techniques are that an expertise is required in software platforms such as javascript, etc., which becomes tedious for a beginner to perform experiments related to timing attacks. Only expert hand or software engineer can analyze such experiments whereas hardware-based techniques require a prototype before performing such experiments which is not a good option if timely countermeasures is the requirement. Such techniques are complex and time consuming and even they require an expert hand in performing such experiments. To perform software attacks successfully adversary needs to have knowledge about software running on cryptosystem. Also sometimes it becomes tedious to interact between software–hardware target machines. An engineer new to this field may find difficulty in understanding such complex techniques. So there is requirement of simulation-based techniques which are one of the best choices for doing early analysis. Detection before production level can be really a cost saving activity for today's semiconductor industry. It is a known fact that every embedded device is vulnerable to hardware attacks known as side channel attacks. It is to be emphasized that protection and detection against hardware attacks plays an important role in securing a diverse set of IOT devices and embedded devices. It is only possible if hardware engineers are successfully able to critically analyze and experiment with such attacks. Simulation-based platform presented in this paper can be helpful for upcoming engineers to perform experiments with hardware attacks which will further help them to build timely countermeasures.

3 Contributions

1. Assessment of the threats related to timing attacks as side channel attack mechanism which affects the security of integrated circuits or embedded devices.
2. We demonstrated Timing leakage as variation in computation time as compared to prior attacks that exploits timing difference related to memory access which is quite cumbersome [21].

3. Our research unfolds a fact that timing attacks can be even more damaging than existing attacks. Timing attacks if combined with an advanced technique known as cryptanalysis can increase the rate of timing information leakage which further can be more damaging to embedded devices. Simulated approach used in this paper does not need any prototype or physical circuit for analysis purposes as existing. Thus our approach proves to be best for doing early analysis and building timely countermeasures against side channel attacks which is demand of today's semiconductor industry.

4 Timing Attack Using SPA as Statistical Method

To conduct the attack, the adversary needs to collect a set (sample) of messages together with their processing time by the cryptographic device. Statistical method used for analyzing timing attack is SPA tool (simple power analysis). Each and every instruction takes a particular time to execute. Based on that timing this information is further exploited as side channel leakage parameter to extract the secret data. Output waveforms clearly indicate exploitation of timing information as side channel leakage parameter. An attacker captures this timing information signal and exploits it for extracting unintended data.

5 Timing Attack Set Up Model

The timing attack model known as pass-code key model has been set up. An algorithm1 is presented which allows user to enter correct key. This assembly code shown as algorithm 1 is interfaced with 8051 microcontroller. Timing attack model is depicted in Fig. 1. Four switches are connected with microcontroller. The correct key can be entered through four switches. SW1, SW2, SW3 and SW4 are four input switches used to enter pass-code key. As soon as the correct key is entered an access to secret information is granted. Flow diagram shown in Fig. 2 explains the timing attack clearly with help of flow diagram. User enters the key and it is compared byte by byte. In our code we have used 4 byte long key. When user enters first byte it is compared against the correct 4 byte long key. Guessed key is compared byte by byte with correct key. As long as entered guessed key is correct, microcontroller will take long to do the comparison. Key will be compared till wrong match is found and timing information representing such execution phenomena are captured by attacker (Fig. 3).

Longer is the comparison if guessed key matches with correct key and shorter will be the comparison if there is mismatch between two keys. Execution time of embedded device is based on this comparison. Execution time information is depicted in Table 1. Attacker maintains such data base and analyzes such timing information later. The longer the comparison, longerthe execution time. To conduct timing attack

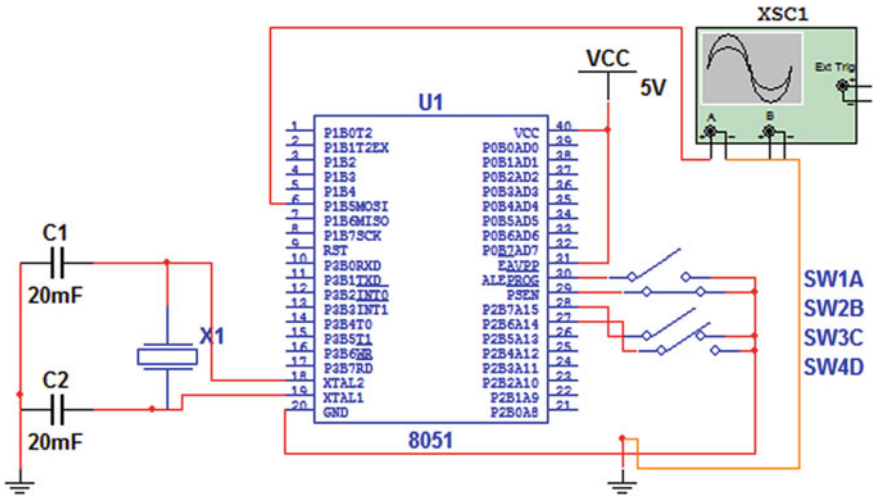


Fig. 1 Timing attack pass-code key model

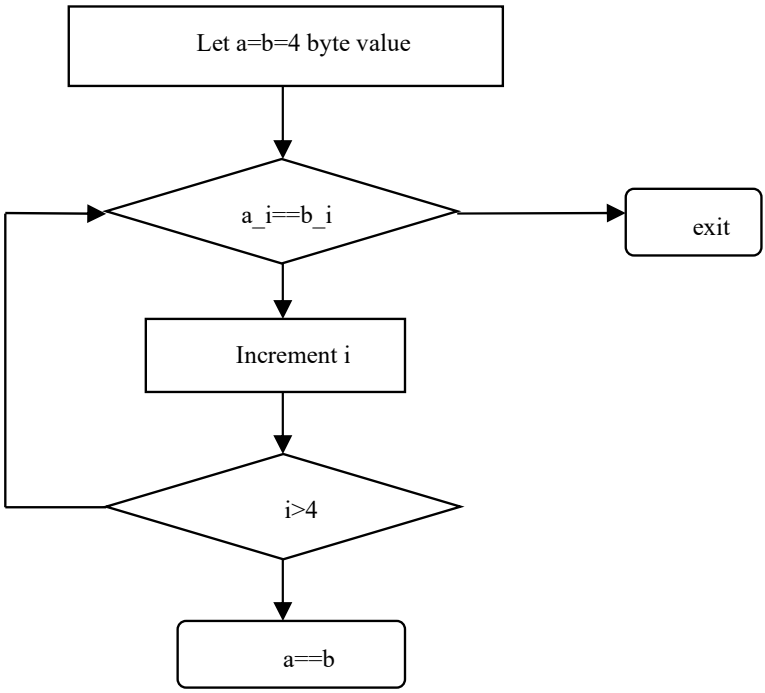


Fig. 2 Flow diagram for timing attack

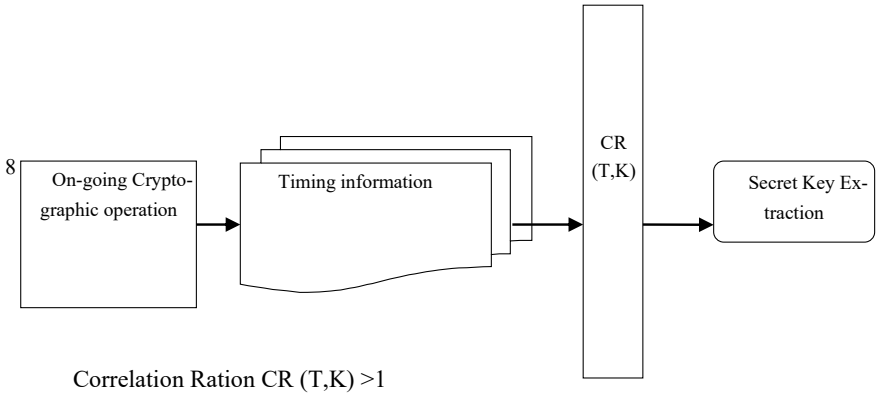


Fig. 3 Model used for performing Timing attack

Table 1 Timing measurement

Time (T)	Attempt to guess key	Successful
T1 = 2.170 μs	1 byte guessed correctly	✗ (Unsuccessful)
T2 = 4.340 μs	2 byte guessed correctly	✗ (Unsuccessful)
T3 = 5.085 μs	1 and 2 byte guessed correctly	✗ (Unsuccessful)
T4 = 10.08 μs	3 byte guessed correctly	✗ (Unsuccessful)
T5 = 7.059 μs	1 and 3 byte guessed correctly	✗ (Unsuccessful)
T6 = 13.021 μs	4 byte guessed correctly	✓ (Successful)

successfully adversary keeps on collecting timing information based on different execution time. As shown in Fig. 4 timing information received was $\boldsymbol{T = 13.02}$ μs when attacker guessed all four keys correctly. Before that attacker keeps on gathering timing information and keeps a data set to analyze it later. Attacker then performs brute force analysis and can keep on trying different combination of keys.

Different keys combinations guessed by attacker will take different execution time. Attacker can easily maintain such data in order to reach correct key. Table 1 indicates parameters used by attacker while using timing information as side channel leakage information. Attacker maintains a database as depicted in Table 1. Before guessing the correct key different key combinations were tried which results in different timing information. $\boldsymbol{T1 = 2.17}$ **μs** for 1 byte guessed correctly. $\boldsymbol{T2 = 4.34}$ **μs** for 2 bytes guessed correctly. $\boldsymbol{T3 = 5.085}$ **μs** for 1 and 2 bytes guessed correctly.

Likewise attackers keep on collecting timing information and at last an attempt to timing attack is successful with $\boldsymbol{T6 = 13.02}$ **μs** for correct guessed key. Based on

Table 2 Advantages/disadvantages of four techniques

Technique	Cost	Performance	Specific hardware required	Complexity	Risk of damage	Controllability	Observability	Portability
Hardware based	High	High	Yes	High	High	High	High	Low
Software based	Low-medium	Medium	No	Medium	Low	High-medium	Medium	Low-medium
Simulation based	Low	Medium	No	Low	Low	High	High	High
Emulation based	Medium	High	Yes	Medium	Low	Medium	High	Medium

timing information acquired attacker can guess the secret key. Longer the comparison, i.e., longer the execution time, more possibility of attacker to guess the correct key. By doing simple power analysis on captured timing waveforms attacker can easily extract the correct key.

Statistical analysis method: *Simple Power Analysis (SPA) for capturing timing information:*

Statistical method used is *SPA* tool (simple power analysis) and *timing information* captured during ongoing cryptographic operation. Let us understand this model with help of an example where user needs to enter the correct key which is 4 byte long. For the sake of simplicity we have chosen correct pass key as "4 4 4 4". For analyzing timing attack model deeply we have discussed cases below. Case 1 indicates attacker attempt for entering secret pass-code. For each and every attempt attacker captures timing waveforms. Based on that captured timing waveforms and by doing simple power analysis of observed waveforms attacker is successful enough to crack the secret key and can get access to secret data.

Where Y is the logic state of the circuit after execution, i.e., it is faulty output. F is the logic state of the circuit, i.e., correct output. T is timing information acquired during ongoing cryptographic operation and K is secret key to be extracted based on doing cryptanalysis of timing information as side channel leakage parameter. The target of SPA attack can be expressed as:

$$Y = F(T, K).$$

It is the attackers attempt to find correlation between correct output F and faulty output Y based on parameters T and K.

6 Correlation Ratio (CR)

It is defined as correlation between correct output and faulty output. If CR > 1 then attack is successful. Faulty trace is having highest correlation ratio with original output means attacker is close enough to guess the secret key. CR ratio is performed in the time domain while performing timing attack. Timing attack is based on exploiting timing information of instruction being executed while a cryptographic operation is going on. It is known that every instruction takes a particular amount of time to execute. Attacker tries to capture this timing information and exploits it in order to reveal secret information. Timing attack is successfully performed having CR > 1. Attacker captures faulty waveforms and time against them. After performing several iterations on timing information adversary can successfully get the information not intended for him. This research work.

Algorithm 1: Algorithm for generating Timing Attack

```
ORG 0000 H
CLR A
SETB P1.5
MOV A, # 04h
CJNE A, # "enter first byte", Loop
CJNE A, # "enter second byte", Loop
CJNE A, # "enter third byte", Loop
CJNE A, # "enter fourth byte", Loop
Loop: CLR P1.5
END
```

Let us discuss CASE 1 from attacker's viewpoint: Various timing-based iterations were performed and some of the cases were captured and discussed here. For sake of simplicity few cases are discussed. Case 1 is an attempt by adversary where attacker is successful enough to crack the secret key. Waveform presented in Fig. 4 is captured from logic analyzer tool. It shows that when attacker enters all the guessed key correctly timing information was equal to $T = 13.02$ μs. As long as the comparison between guessed and correct key is longer more timing information will be captured by attacker as depicted in Fig. 4 . Case 1 shows a successful attack attempt with maximum value of timing information. Shorter the comparison less will be execution time. Table 1 shows the values based on timing information gathered by attacker. Attacker keeps on gathering such huge amount of data until the correct match is found.

Case 1 depicts the scenario where adversary enters all the values of correct pin. Other cases like case 2 and 3 depict wrong attack attempt by adversary. Attack attempt along with timing information is depicted. Many such waveforms were captured with depicts different timing information against execution time of different key combinations. Different cases were captured and analyzed. For the sake of simplicity we have just represented few waveforms with their corresponding execution time. Attacker can keep on doing different trials and can gather huge amount of information based on these timing attack. Adversary uses the timing information as side channel leakage information and exploits the embedded devices in order to extract the secret key not intended for him. Table 1 shows timing information to be exploited by adversary as a side channel leakage parameter.

CASE 1: When All 4 Key Inputs Are Correct

See Fig. 4

CASE 2: When 3 Key Inputs Are Correct

See Fig. 5

CASE 3: When 2 Input Key Is Correct

See Fig. 6.

Graph in Fig. 7 shows the gap between guessed key and correct key. It is the attempt of attacker to reduce this gap by getting close to secret information and making a timing attack attempt a successful one. Adversary may perform brute force analysis and keep on gathering timing information against the target system in order

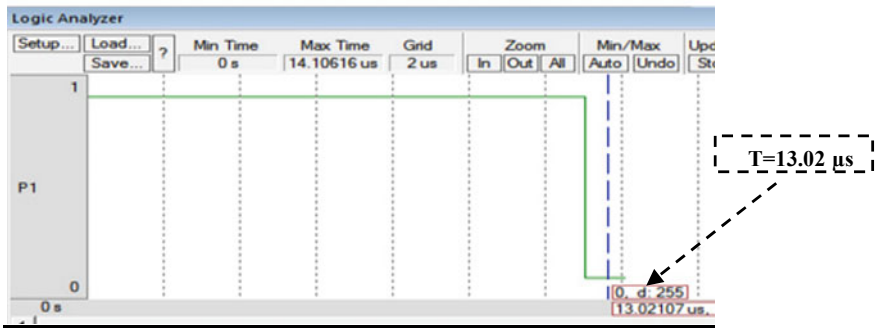


Fig. 4 Timing information as side channel leakage with T1 = 13.02 µs

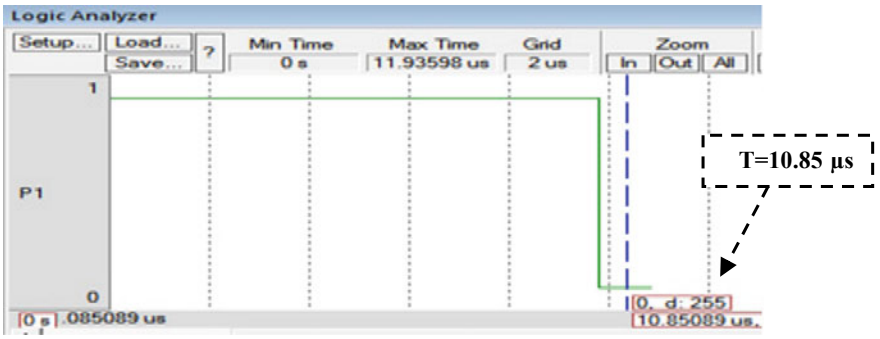


Fig. 5 Timing information as side channel leakage with T2 = 10.85 µs

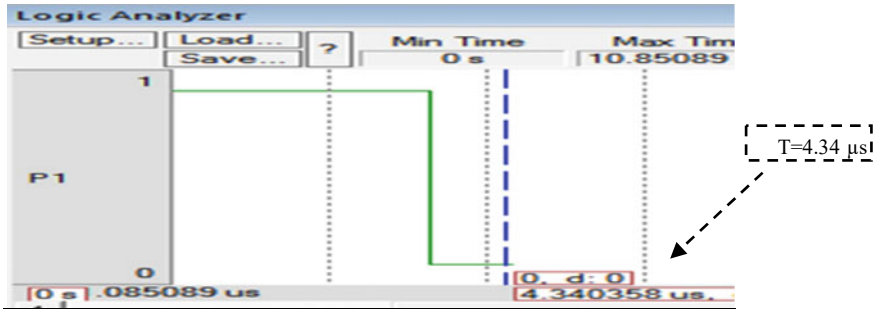


Fig. 6 Timing information as side channel leakage with T3 = 4.34 µs

to minimize this gap. Large amount of database is captured by attacker in order to minimize the gap. Correct guess key will make the gap equal to zero.

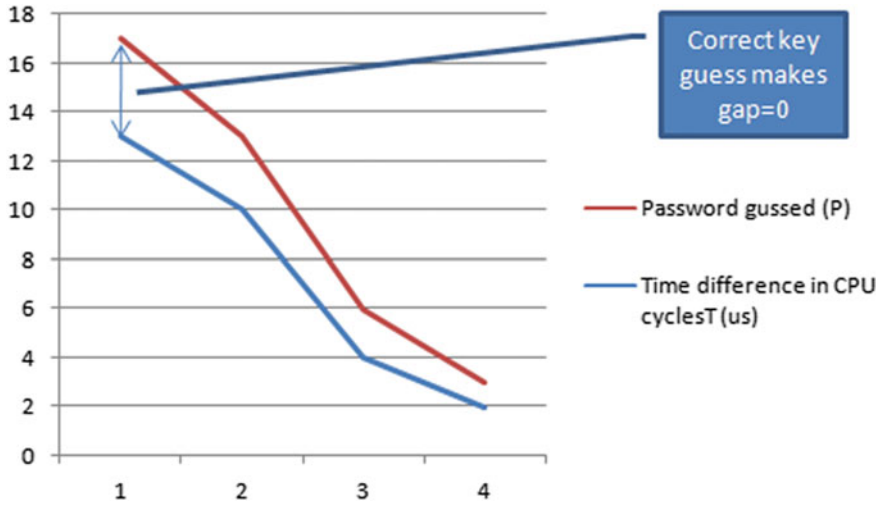
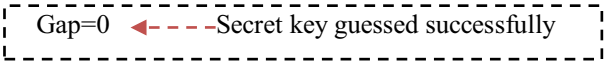


Fig. 7 Time difference between correct and guessed key



7 Conclusion

In this paper an attempt to emulate a real timing-based attack has been made which will provide an efficient evaluation platform for upcoming hardware security specialists to study and analyze effect of timing attack. We tried to show how timing leakage as variation in computation time is exploited. Proposed simulation-based method is successful in doing early analysis of attacks so that in future timely countermeasures can be built as compared to existing hardware and software techniques. Existing research states that unintentional timing characteristics would only reveal a small amount of information from a cryptosystem (such as the Hamming weight of the key). Our experimental work breaks this stereotype and reveals a fact that timing attacks can be even more dangerous. Timing attacks if combined with an advanced technique known as cryptanalysis can increase the rate of timing information leakage which further can be more damaging to embedded devices. Our solution has an edge over existing in a way that it allows to emulate a real timing-based attack framework which will act as evaluation platform in critically analyzing timing attacks and providing suitable countermeasures it. Simulated approach used in this paper does not need any prototype or physical circuit for analysis purposes as existing. Thus our approach proves to be best for doing early analysis and building timely countermeasures against side channel attacks which is demand of today's semiconductor

industry. It can be used as testing experiment during early step of design, thereby shortening and helping the designers to find the best solution during a preliminary phase and potentially without additional costs**.**

References

1. P.C. Kocher, Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems, in *Annual International Cryptology Conference* (1996) , pp. 104–113
2. O. Acıiçmez, W. Schindler, C.K. Koç, Cache based remote timing attack on the AES, in *Cryptographers Track at the RSA Conference* (2007) , pp. 271–286
3. W. Cilio, M. Linder, C. Porter, J. Di, D.R. Thompson, S.C. Smith, Mitigating power-and timing-based side-channel attacks using dual-spacer dual-rail delay-insensitive asynchronous logic. Microelectron. J. **44**(3), 258–269 (2013)
4. D. Brumley, D. Boneh, Remote timing attacks are practical. Comput. Netw. **48**(5) 701–716 (2005)
5. U.M. Sadique, D. James, A novel approach to prevent cache-based side-channel attack in the cloud. Proc. Technol. **25**(2), 32–39 (2016)
6. R.S. Nair, S.C. Smith, J. Di, Delay insensitive ternary CMOS logic for secure hardware. J. Low Power Electron. Appl. **5**(3), 183–215 (2015)
7. J.D. Mayer, J. Sandin, Time trial: "Racing towards practical remote timing attacks" Black Hat US Briefings (2014)
8. W. Schindler, A timing attack against RSA with the Chinese remainder theorem, in *International Workshop on Cryptographic Hardware and Embedded Systems* (2000), pp. 109–124
9. W. Schindler, A combined timing and power attack, in *International Workshop on Public Key Cryptography* (2002), pp. 263–279
10. B.B. Brumley, N. Tuveri, Remote timing attacks are still practical, in *Proceedings of the European Symposium on Research in Computer Security* (2011)
11. C. Arnaud, P.-A. Fouque, Timing attack against protected RSA-CRT implementation used in PolarSSL, in *Proceedings of the Cryptographers' Track at the RSA Conference* (2013), pp.18–33
12. M. Schwarz, M. Lipp, D. Gruss, S. Weiser, S. Maurice, R. Spreitzer, S. Mangard, Keydrown: eliminating software-based keystroke timing side-channel attacks (2018)
13. C. Luo, Y. Fei, D. Kaeli, GPU acceleration of RSA is vulnerable to side-channel timing attacks, in *Proceedings of the International Conference on Computer-Aided Design* (2018), pp. 1–8
14. R. Tóth, Z. Faigl, M. Szalay, S. Imre, An advanced timing attack scheme on RSA, in *International Telecommunications Network Strategy and Planning Symposium*, vol. Supplement (2008), pp. 1–9
15. M. Lipp, D. Gruss, M. Schwarz, D. Bidner, C. Maurice, S. Mangard, Practical keystroke timing attacks in sandboxed javascript, in *European Symposium on Research in Computer Security* (2017), pp. 191–209
16. M. Schwarz, M. Lipp, G. Gruss, S. Weiser, C. Maurice, R. Spreitzer, S. Mangard, Keydrown: eliminating keystroke timing side-channel attacks (2017)
17. D. Gruss, D. Bidner, S. Mangard, Practical memory deduplication attacks in sandboxed javascript, in *ESORICS'15* (2015)
18. B. Gras, K. Razavi, E. Bosman, H. Bos, C. Giuffrida, ASLR on the line: practical cache attacks on the MMU, in *NDSS'17* (2017)
19. M.Schwarz, C. Maurice, D. Gruss, S. Mangard, Fantastic timers and where to find them: high-resolution microarchitectural attacks in javascript. in *FC'17* (2017)
20. Y. Lyu, P. Mishra, A survey of side-channel attacks on caches and countermeasures. J. Hardware Syst. Sec. **2**(1), 33–50 (2018)

21. Z.H. Jiang, Y. Fei, D. Kaeli, A complete key recovery timing attack on a GPU, in *HPCA* (2016), pp. 394–405
22. A.C. Aldaya, C.P. García, L.M.A. Tapia, B.B. Brumley, Cache-timing attacks on RSA key generation, in *IACR Transactions on Cryptographic Hardware and Embedded Systems* (2019), pp. 213–242
23. C. Luo, Y. Fei, D. Kaeli, Side-channel timing attack of RSA on a GPU. ACM Trans. Architect. Code Optimizat. (TACO) **16**(3), 1–18 (2019)
24. F. Tramèr, D. Boneh, K. Paterson, Remote side-channel attacks on anonymous transactions, in *29th {USENIX} Security Symposium* (2020), pp. 2739–2756
25. M. Lipp, A. Kogler, D. Oswald, M. Schwarz, C. Easdon, C. Canella, D. Gruss. PLATYPUS: software-based power side-channel attacks on x86, in *IEEE Symposium on Security and Privacy (SP)* (2021)

Quantum Dot Cellular Automata-Based Design of 4 × 4 TKG Gate and Multiplier with Energy Dissipation Analysis

Soha Maqbool Bhat, Suhaib Ahmed, and Vipan Kakkar

Abstract With the advancements in electronics industry and continuous scaling of VLSI devices, the current CMOS technology is starting to reach its limits in the nano regime and hence various other technologies are being explored for future devices fabrication and production. Quantum dot Cellular Automata (QCA) is one such technology which has a good potential to operate under low power, high speed and small area requirements. In this paper, QCA technology has been used to design a 4 × 4 reversible TKG gate. This gate is designed using 34 cells with a latency of clock cycles. In addition to this, a logic level design has also been proposed for a 4 × 4 Multiplier using this TKG gate along with another 3 × 3 reversible SSG-1 gate. The performance evaluation of the proposed designs affirms the efficiency of these designs and validates the low power consumption by QCA technology-based designs in nano regime.

Keywords Quantum computing · Quantum dot cellular automata · Multiplier · Reversible computing · Nanoelectronics

1 Introduction

The scaling down of traditional CMOS semiconductors (silicon based) has been the main thrust behind the quick development of the electronic industry since it assisted with expanding the chip thickness, lessen power dispersal and furthermore to speed up coordinated circuits. Be that as it may, in the profound nanometer system, a quantum impact comes into the image and assumes a significant part in the semiconductor activity. Different issues in planning circuits with profound nanometer semiconductor are high power utilization and electron relocation/migration alongside leakage current [1–4]. Different choices to address the issues being looked at by CMOS innovation are being explored by ITRS [2, 5, 6]. Given the future interest of

S. M. Bhat · V. Kakkar
School of ECE, SMVD University, Katra, India

S. Ahmed (✉)
Department of ECE, BGSB University, Rajouri, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_61

advanced frameworks, arising nanotechnologies seem to be a decent market to put resources into [1, 4, 7]. In this specific situation, Quantum dot Cellular Automata (QCA) is an innovation which is being explored for digital designing due to its ability to provide solutions for different issues looked at by CMOS innovation in the nano system [8–10]. QCA circuits can be created by utilizing molecules and thus the versatility of QCA is vastly improved contrasted with customary silicon CMOS semiconductor scaling.

1.1 QCA Technology

One of the most encouraging electronic innovations being studied a lot by the analysts is nanotechnology dependent on QCA which was proposed by C.S. Lent in 1993 [3]. QCA is created on the idea of utilizing cells and the connection between charges in these cells is adequate to perform calculation and data change [1, 2]. Another benefit of electronic circuit configuration based on QCA technology is that for the cell's interconnection, QCA does not require traditional wires. QCA cell is the essential segment in a QCA circuit and is depicted in Fig. 1a. It comprises four quantum wells in which just two electrons are confined which are isolated by quantum tunnel junctions. The coulombic connections between these QCA cells are liable for the electron flow in QCA circuits. The electrons dwell just at diagonal destinations, i.e., corner to corner, in this manner accomplishing least repugnance [11–13]. These electron living positions bring about two states of polarization in QCA which are binary '1' and '0' as depicted in Fig. 1a. Contrasted with CMOS innovation, there is generally low energy dissemination as there is negligible and irrelevant energy scattering during the state change and propagation in QCA [3, 8, 9]. The QCA cells

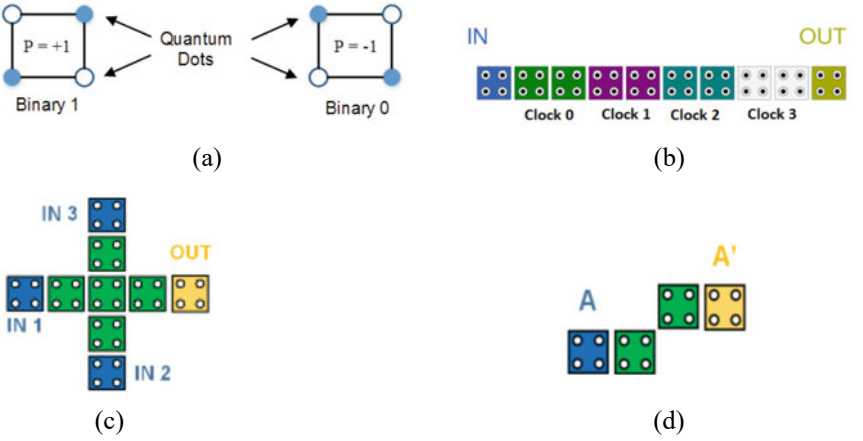


Fig. 1 **a** QCA cells as binary 1 and binary 0 [14], **b** binary wire with different clock zones [15] **c** majority voter gate with three inputs [15, 16] **d** inverter in QCA [15, 16]

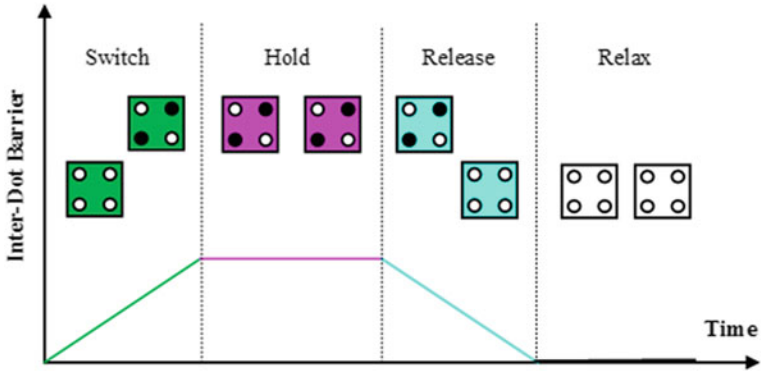


Fig. 2 QCA clocking illustration depicting its various clock zones and phases [14]

are often connected together to form a binary wire as depicted in Fig. 1b. QCA circuits also operate by utilizing four clock zones (clock 0 through 3) where each clock zone has phase difference of 90° with previous clock. Some of the basic circuits use in QCA are majority voter and inverter shown in Fig. 1c and d, respectively. The clock used also has four operating phases which are switch, hold, release and relax, shown in Fig. 2. The detailed description on these clock phases can be accessed from [1, 2].

1.2 Reversible Logic

One of the main components of Reversible Computing is a Reversible Logic which is achieved when one-to-one mapping is possible in the logic along with condition that number of inputs and outputs are same. By doing so, a logic gate can be called as a Reversible Gate. Different authors over the years have proposed different reversible Gates of different sizes such as 2 × 2, 3 × 3, 4 × 4, etc. Some of these gates have been given in Table 1.

Table 1 Few existing Reversible Gates in literature

Gate Size	Gate
2 × 2	Feynman [17]
3 × 3	RG-QCA [6], Fredkin [18], IMG [19], SSG-QCA [9], RUM [20] NNG [21], TR [22], Peres [23], QCA1 [24], SSG-1 [25]
4 × 4	MRLG [26], HNG [27], PFAG [28], RAM [29], TKG [30], HNFG [31]

1.3 Paper Contributions and Organization

The main contributions of this work are:

- First ever reported QCA design of 4 × 4 TKG gate
- Logic development of partial product generator for 4 × 4 reversible multiplier
- Logic development of addition sub-circuit 4 × 4 reversible multiplier
- Energy dissipation analysis of proposed QCA design

The remaining paper is organized as: description of TKG gate and its proposed QCA implementation along with design verification and performance evaluation is presented in Sect. 2. This section is followed by proposed logics of partial product generator and addition sub-circuit for 4 × 4 reversible multiplier in Sect. 3 The energy dissipation analysis for proposed QCA design of TKG gate is presented in Sect. 4 with concluding statements in Sect. 5.

2 QCA Implementation of TKG Gate

Rather et al. in [30] proposed a 4 × 4 TKG gate which is reversible in nature. The equations of TKG gate along with their quantum implementation are shown in Fig. 3 and the truth table for TKG gate is given in Table 2. The proposed QCA circuit of TKG gate is shown in Fig. 4. The TKG gate is designed using 34 cells occupying total area of 0.0356 μm^2 and having 0.5 clock cycle latency. The cell area occupied by the proposed design is 0.0356 μm^2. The functionality of the design is checked using QCADesigner tool's coherent vector simulation engine [32].

The waveform obtained from the simulating the proposed design TKG gate in QCADesigner tool is shown in Fig. 5. It is seen that when input A = 1, B = 0, C = 1 and D = 1, the outputs of the proposed TKG design in the waveform are P = 1, Q = 1, R = 1 and S = 0.

These values coincide with the values given in Table 2. Similarly, all other 15 input and corresponding output combinations for the proposed design can be verified from the simulation waveform and matched with the truth table.

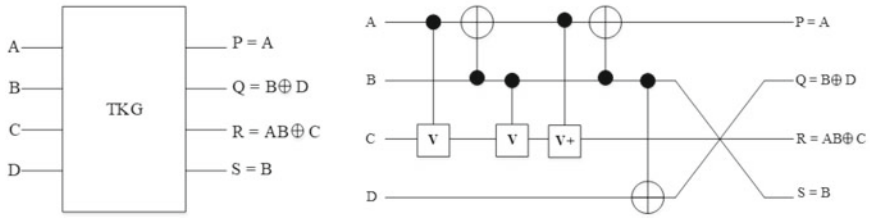


Fig. 3 Illustration of **a** TKG Gate and **b** its quantum implementation [30]

Table 2 Truth Table of TKG Gate [30]

Input				Output			
A	B	C	D	P	Q	R	S
0	0	0	0	0	0	0	0
0	0	0	1	0	1	0	0
0	0	1	0	0	0	1	0
0	0	1	1	0	1	1	0
0	1	0	0	1	1	0	1
0	1	0	1	0	0	0	1
0	1	1	0	0	1	1	1
0	1	1	1	0	0	1	1
1	0	0	0	1	0	0	0
1	0	0	1	1	1	0	0
1	0	1	0	1	0	1	0
1	0	1	1	1	1	1	0
1	1	0	0	1	1	1	1
1	1	0	1	1	0	1	1
1	1	1	0	1	1	0	1
1	1	1	1	1	0	0	1

The performance of QCA design proposed for TKG gate is estimated by comparing its performance parameters with existing designs. This performance comparison is presented in Table 3. Here QCA is product of total area and square of latency.

It is seen that the proposed TKG design has better performance in almost all parameters with low QCA cost, thereby making it appropriate for application in different nano circuits.

3 Multiplier Design

In this paper, logic of a reversible gate-based 4 × 4 multiplier has been proposed. A 4 × 4 multiplier consists of two 4-bit inputs (X, Y) which leads to an 8-bit output Z. Figure 6 illustrates the mechanism of multiplication of these 4-bit numbers. It is a two-step process viz., partial product generation and addition of these partial products. Here MSB bit Z_7 stores the carry generated from the partial product of MSB bits of both inputs, i.e., X_3 and Y_3.

Here, a logic level implementation of partial product generator is proposed using the TKG gate and another 3 × 3 reversible SSG-1 gate. A total of 16 reversible gates have been used to design the partial product generator. This logic is depicted in

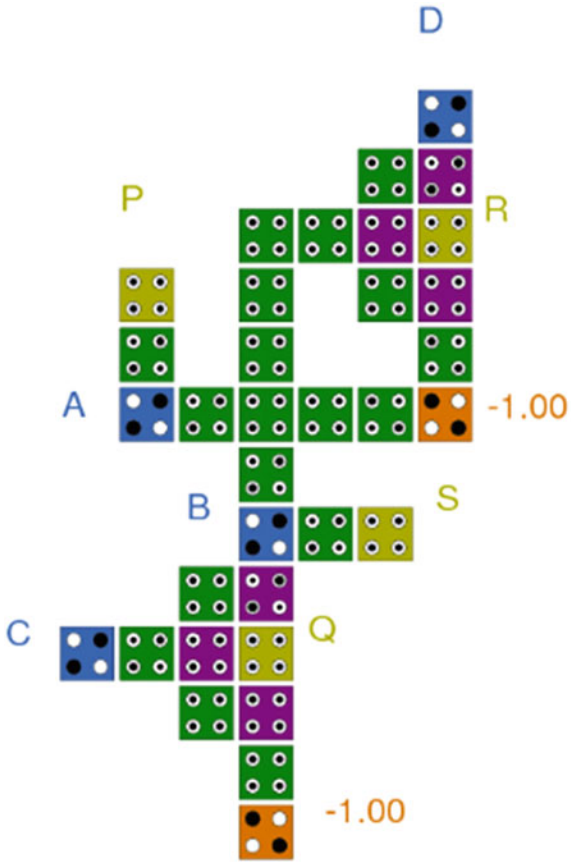


Fig. 4 Proposed QCA circuit design of TKG gate

Fig. 7. Out of these 16 gates, 4 are TKG gate and 12 are SSG-1 gate. The proposed structure is suitable to designing low delay multiplier circuits. Similarly, a logic for addition sub-circuit has also been proposed. This sub-circuit is used to add, as per the mechanism illustrated in Fig. 6, the individual products generated from the partial product generator. The proposed sub-circuit logic, shown in Fig. 8, consists of 12 gates of which 8 are TKG and 4 are SSG-1 gates. In future work, these proposed logics can be implemented using the proposed TKG designs in QCA in order to develop a QCA-based 4×4 reversible multiplier for application in various arithmetic circuits and processors.

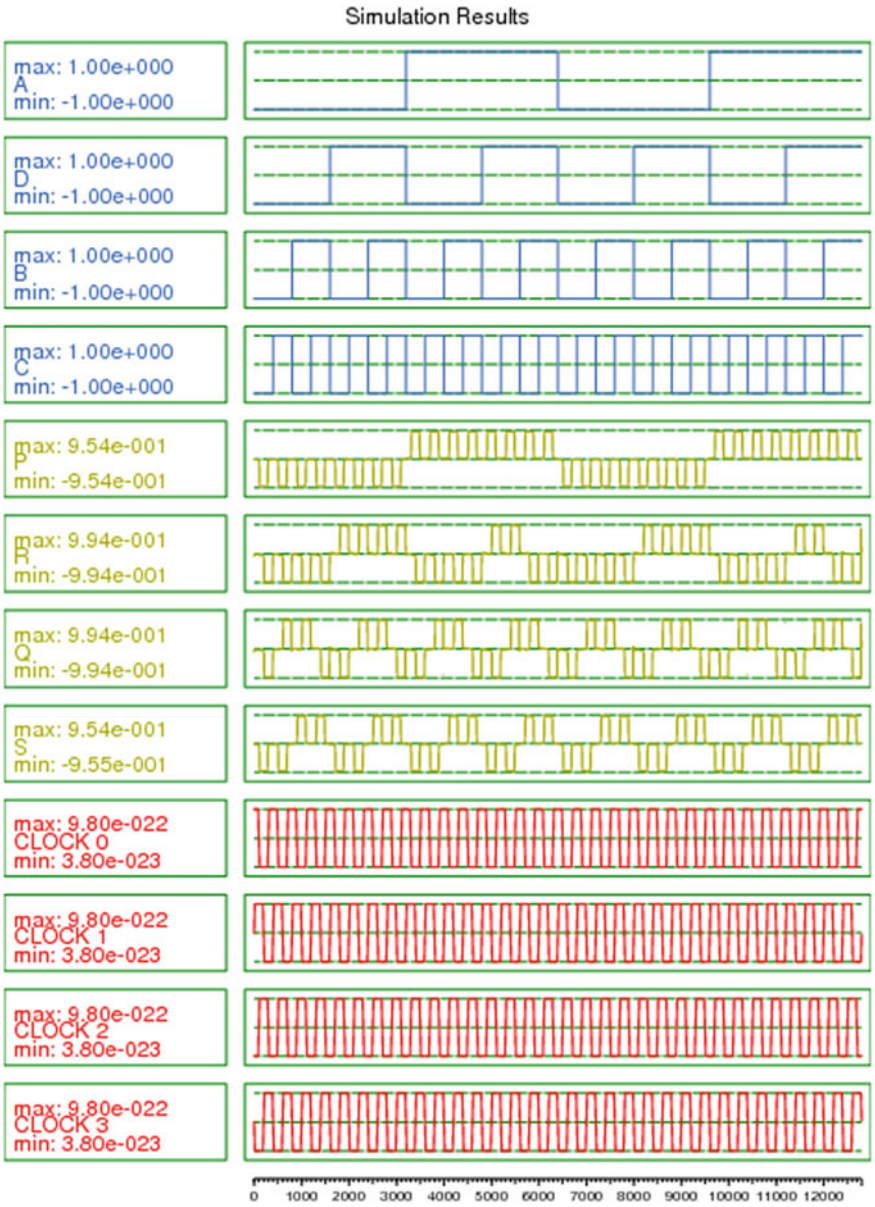


Fig. 5 Waveform of QCA design of TKG gate

Table 3 Comparison of Reversible Gates

Parameter	[33]	[34]	[6]	[20]	Proposed Design
Cell Count	178	171	157	142	34
Cell Area (μm^2)	0.0577	0.055	0.051	0.046	0.01102
Total Area (μm^2)	0.23	0.19	0.19	0.201	0.0356
Latency	3.25	3	1.75	2	0.5
QCA cost	2.43	1.71	0.582	0.804	0.0089

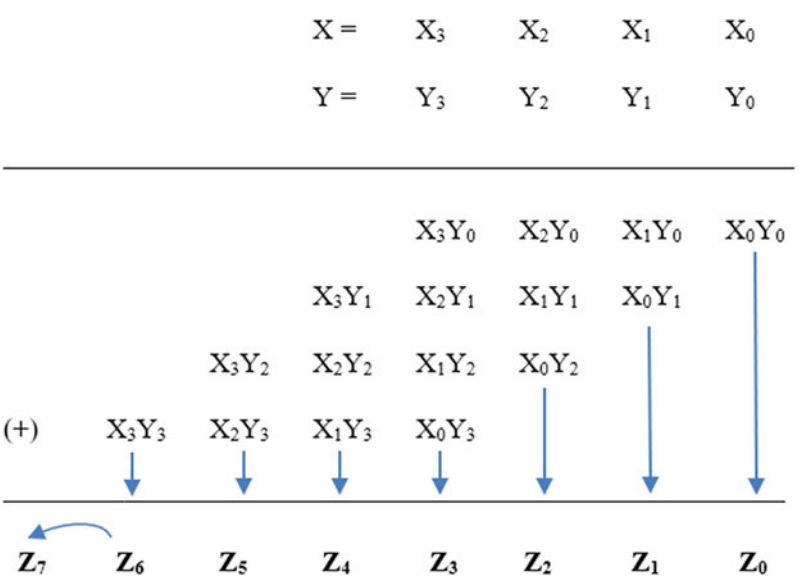


Fig. 6 Operation of 4 × 4 multiplier

4 Energy Estimation Analysis

The energy dissipation of the QCA design of proposed for TKG gate is analyzed using QCADesigner-E tool [35] which estimates energy dissipated by QCA circuits as per the mechanism presented in [36–38]. In this tool, total energy dissipation (E_{bath_total}) is approximated as sum of all 'bath' of energies (E_{bath}) for each clock cycle formed by each cell in the design. The total energy dissipation of the proposed design is equal to 3.44×10^{-2} eV with an error of $\pm 3.60 \times 10^{-3}$ eV and the per cycle average energy dissipation is 3.12×10^{-3} eV with an error of $\pm 3.27 \times 10^{-4}$ eV.

These values have been computed from the energy dissipation computed for complete structure and presented in Table 4. E_{clk} gives the sum of energy transfer within the cells in the circuit and the clock which is separated to each clock cycle. For understanding of these values, energy dissipated by output cells P, Q, R and S

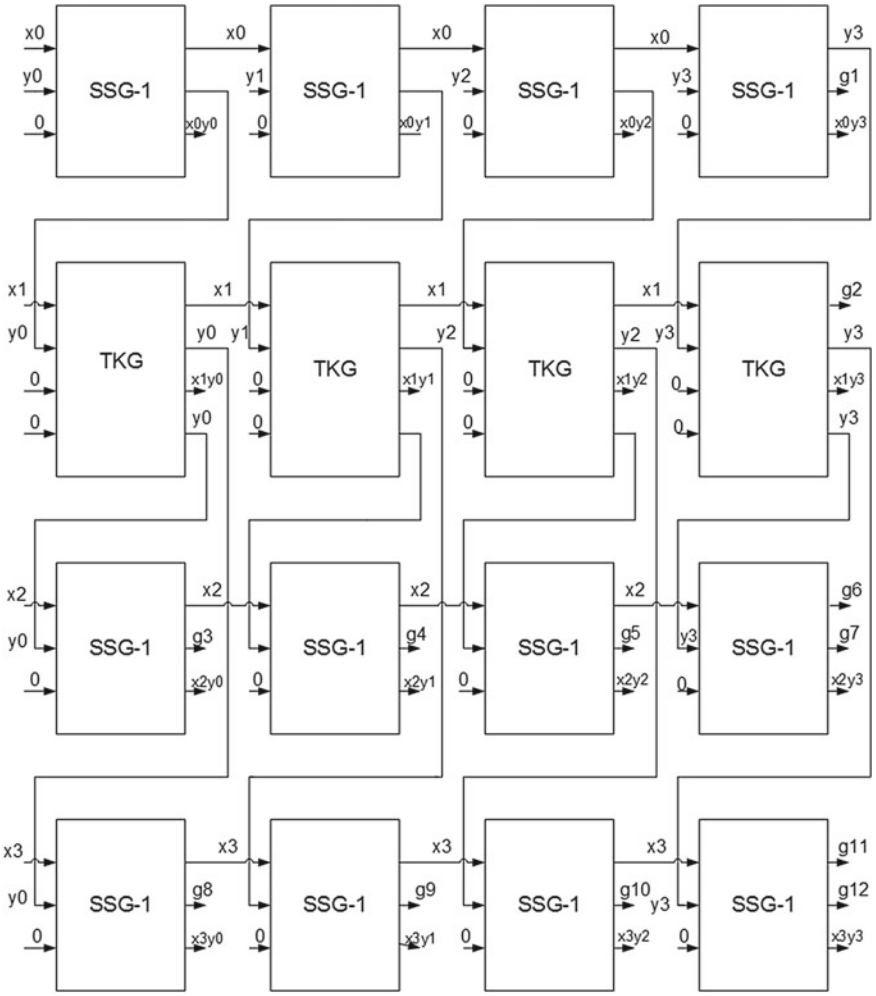


Fig. 7 Proposed partial product generator for 4 × 4 multiplier

has also been presented in Tables 5, 6, 7 and 8, respectively. Similarly, these values are obtained for all cells in the QCA design and are used to obtain the values given in Table 4.

5 Conclusion

In this paper, QCA technology has been explored to design reversible logics. A new design of 4 × 4 TKG reversible gate has been proposed in QCA. It is designed using 34 cells occupying total area of 0.0356 μm^2 and 0.5 clock cycle latency. The

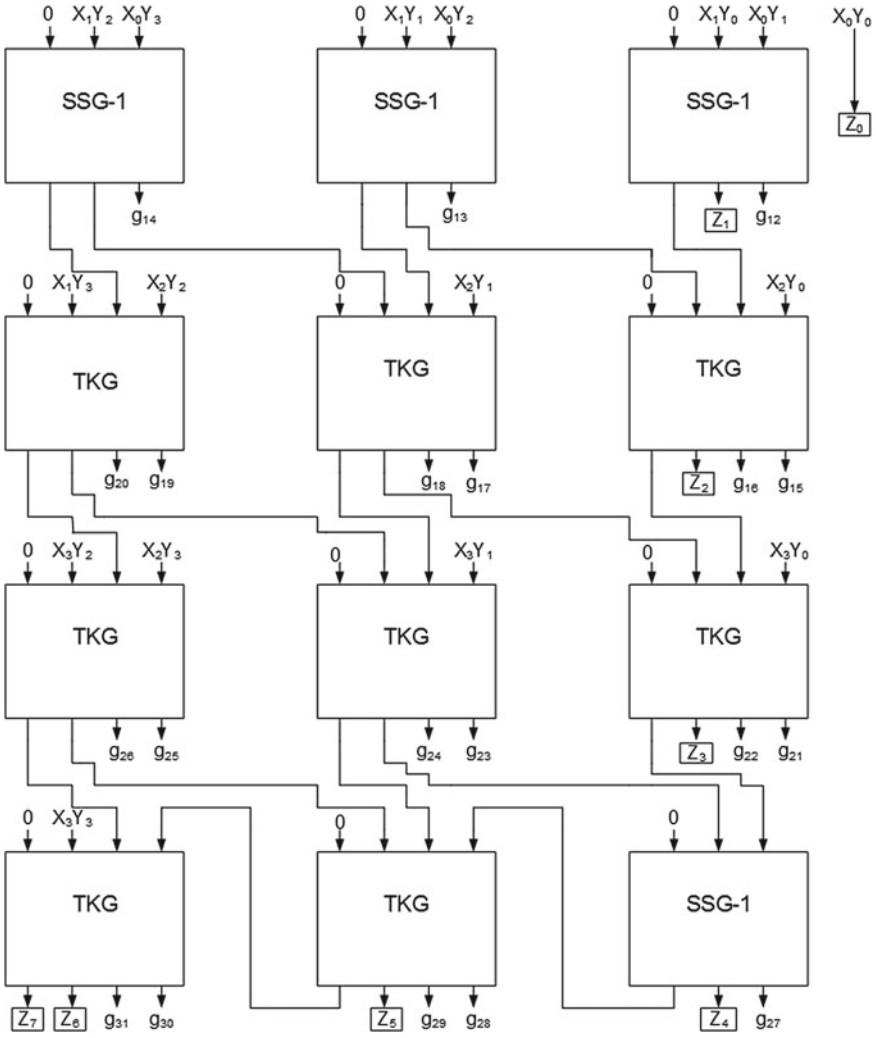


Fig. 8 Proposed addition sub-circuit for 4 × 4 multiplier

proposed QCA design dissipates energy of 3.44×10^{-2} eV with an error of $\pm 3.60 \times 10^{-3}$ eV and the per cycle average energy dissipation is 3.12×10^{-3} eV with an error of $\pm 3.27 \times 10^{-4}$ eV. Also, a new logic of 4 × 4 reversible multiplier has also been proposed wherein TKG gate along with another reversible SSG-1 gate have been used to proposed partial product generator and addition sub-circuits. In future work, the proposed QCA design of TKG gate can be used to design QCA circuit for this 4 × 4 reversible multiplier for application in various nano regime arithmetic circuits.

Table 4 Energy dissipation of TKG gate

Energy	Value (eV)										
E_bath_total	4.6876×10^{-3}	4.2530×10^{-3}	2.9093×10^{-3}	2.6703×10^{-3}	2.2771×10^{-3}	3.7690×10^{-3}	1.8604×10^{-3}	3.8317×10^{-3}	3.4376×10^{-3}	2.9975×10^{-3}	1.6608×10^{-3}
E_clk_total	-5.9844×10^{-4}	3.8458×10^{-4}	3.7028×10^{-4}	1.3519×10^{-3}	2.4615×10^{-3}	5.6349×10^{-3}	1.8076×10^{-3}	-5.3921×10^{-4}	8.6413×10^{-4}	1.8589×10^{-3}	1.8314×10^{-3}
E_Error_total	-4.9960×10^{-4}	-4.4790×10^{-4}	-3.0366×10^{-4}	-2.8125×10^{-4}	-2.3478×10^{-4}	-4.0092×10^{-4}	-1.8769×10^{-4}	-4.0520×10^{-4}	-3.6256×10^{-4}	-3.1026×10^{-4}	-1.6678×10^{-4}

Table 5 Energy dissipation by output P of TKG gate

Energy	Value (eV)										
E_bath	1.0830×10^{-5}	3.7964×10^{-5}	1.0846×10^{-5}	3.8806×10^{-5}	2.7316×10^{-5}	3.7935×10^{-5}	2.7841×10^{-5}	3.7970×10^{-5}	1.0848×10^{-5}	3.6048×10^{-5}	1.0862×10^{-5}
E_clk	1.1176×10^{-4}	1.4508×10^{-4}	1.1174×10^{-4}	1.4514×10^{-4}	1.4768×10^{-4}	1.4613×10^{-4}	1.4762×10^{-4}	1.4623×10^{-4}	1.1245×10^{-4}	1.4576×10^{-4}	1.1243×10^{-4}
E_io	-1.2326×10^{-4}	-1.8677×10^{-4}	-1.2327×10^{-4}	-1.8776×10^{-4}	-1.7754×10^{-4}	-1.8779×10^{-4}	-1.7806×10^{-4}	-1.8793×10^{-4}	-1.2397×10^{-4}	-1.8532×10^{-4}	-1.2397×10^{-4}
E_in	0	0	0	0	0	0	0	0	0	0	0
E_out	2.6542×10^{-6}	3.6261×10^{-6}	2.4988E × 10^{-6}	3.7601E × 10^{-6}	4.8930×10^{-6}	2.6865×10^{-6}	4.8811×10^{-6}	2.6435×10^{-6}	2.6213×10^{-6}	4.5580×10^{-6}	2.4913×10^{-6}
E_Error	-6.7354×10^{-7}	-3.7258×10^{-6}	-6.7532×10^{-7}	-3.8195×10^{-6}	-2.5405×10^{-6}	-3.7220×10^{-6}	-2.5988×10^{-6}	-3.7261×10^{-6}	-6.7597×10^{-6}	-3.5132×10^{-6}	-6.7748×10^{-7}

Table 6 Energy dissipation by output Q of TKG gate

Energy	Value (eV)										
E_bath	5.8293×10^{-5}	5.8385×10^{-5}	3.4796×10^{-5}	3.4795×10^{-5}	3.4700×10^{-5}	3.4987×10^{-5}	2.3055×10^{-5}	2.3062×10^{-5}	5.8293×10^{-5}	5.8385×10^{-5}	3.4796×10^{-5}
E_clk	5.9756×10^{-5}	6.0449×10^{-5}	-2.3232×10^{-5}	-2.2855×10^{-5}	6.4790×10^{-5}	6.3165×10^{-5}	5.8679×10^{-5}	6.2713×10^{-5}	5.9761×10^{-5}	6.0450×10^{-5}	-2.3233×10^{-5}
E_io	-1.2355×10^{-4}	-1.2580×10^{-4}	-1.4508×10^{-5}	-1.5702×10^{-5}	-1.0240×10^{-4}	-1.0030×10^{-4}	-8.3310×10^{-5}	-8.5866×10^{-5}	-1.2356×10^{-4}	-1.2580×10^{-4}	-1.4507×10^{-5}
E_in	1.4068×10^{-5}	1.2609×10^{-5}	-1.1652×10^{-4}	-1.1627×10^{-4}	1.7184×10^{-4}	1.7175×10^{-4}	1.5821×10^{-4}	1.5983×10^{-4}	1.4066×10^{-5}	1.2609×10^{-5}	-1.1652×10^{-4}
E_out	9.0248×10^{-6}	9.0290×10^{-6}	-9.4410×10^{-6}	-9.4286×10^{-6}	9.3669×10^{-6}	9.4085×10^{-6}	$-–9.4058 \times 10^{-6}$	-9.4359×10^{-6}	9.0248×10^{-6}	9.0290×10^{-6}	-9.4410×10^{-6}
E_Error	-5.5051×10^{-6}	-6.9614×10^{-6}	-2.9437×10^{-6}	-3.7621×10^{-6}	-2.9059×10^{-6}	-2.1526×10^{-6}	-1.5754×10^{-6}	-9.0756×10^{-8}	-5.5052×10^{-6}	-6.9613×10^{-6}	-2.9437×10^{-6}

Table 7 Energy dissipation by output R of TKG gate

Energy	Value (eV)										
E_bath	5.5200×10^{-5}	5.5134×10^{-5}	5.5200×10^{-5}	5.5134×10^{-5}	2.2767×10^{-5}	5.5210×10^{-5}	2.2767×10^{-5}	5.5210×10^{-5}	3.4743×10^{-5}	3.4744×10^{-5}	3.4743×10^{-5}
E_clk	1.4261×10^{-4}	1.4147×10^{-4}	1.4261×10^{-4}	1.4147×10^{-4}	2.4011×10^{-4}	1.4250×10^{-4}	2.4011×10^{-4}	1.4248×10^{-4}	1.0067×10^{-4}	1.0087×10^{-4}	1.0067×10^{-4}
E_io	-2.0296×10^{-4}	-2.0174×10^{-4}	-2.0296×10^{-4}	-2.0174×10^{-4}	-2.6441×10^{-4}	-2.0285×10^{-4}	-2.6442×10^{-4}	-2.0399×10^{-4}	-1.3831×10^{-4}	-1.3852×10^{-4}	-1.3831×10^{-4}
E_in	-6.0013×10^{-6}	-5.4443×10^{-6}	-6.0002×10^{-6}	-5.4425×10^{-6}	2.0426×10^{-4}	-5.8642×10^{-6}	2.0425×10^{-4}	-5.8639×10^{-6}	1.6722×10^{-4}	1.6695×10^{-4}	1.6722×10^{-4}
E_out	0	0	0	0	0	0	0	0	0	0	0
E_Error	-5.1487×10^{-6}	-5.1312×10^{-6}	-5.1487×10^{-6}	-5.1292×10^{-6}	-1.5380×10^{-6}	-5.1395×10^{-6}	-1.5380×10^{-6}	-6.2968×10^{-6}	-2.8990×10^{-6}	-2.9050×10^{-6}	-2.8990×10^{-6}

Table 8 Energy dissipation by output S of TKG gate

Energy	Value (eV)										
E_bath	1.9989×10^{-4}	2.0142×10^{-4}	9.3374×10^{-6}	9.5964×10^{-6}	2.5830×10^{-4}	3.9350×10^{-5}	9.1805×10^{-6}	9.1698×10^{-6}	1.9956×10^{-4}	2.0105×10^{-4}	9.3047×10^{-6}
E_clk	9.3723×10^{-5}	9.2040×10^{-5}	9.4339×10^{-5}	9.3115×10^{-5}	6.6882×10^{-5}	1.3641×10^{-4}	9.9653×10^{-5}	6.3756×10^{-5}	9.4133×10^{-5}	9.2455×10^{-5}	9.4030×10^{-5}
E_io	-3.1532×10^{-4}	-3.1534×10^{-4}	-1.0418×10^{-4}	-1.0324×10^{-4}	-3.5338×10^{-4}	-1.7963×10^{-4}	-1.0934×10^{-4}	-7.3384×10^{-5}	-3.1536×10^{-4}	-3.1534×10^{-4}	-1.0383×10^{-4}
E_in	-3.2483×10^{-4}	-3.2453×10^{-4}	-1.0697×10^{-4}	-1.0570×10^{-4}	-3.3968×10^{-4}	-1.7323×10^{-4}	-1.0664×10^{-4}	-7.1607×10^{-5}	-3.2420×10^{-4}	-3.2386×10^{-4}	-1.0660×10^{-4}
E_out	2.2882×10^{-7}	2.2923×10^{-7}	1.1631×10^{-7}	1.1638×10^{-7}	-2.3205×10^{-7}	-1.7671×10^{-7}	-1.1386×10^{-7}	-1.0718×10^{-7}	-2.6622×10^{-7}	-2.6607×10^{-7}	-2.5465×10^{-7}
E_Error	-2.1706×10^{-5}	-2.1876×10^{-5}	-5.0241×10^{-7}	−5.3197E-07	-2.8198×10^{-5}	-3.8717×10^{-6}	-5.0815×10^{-7}	-4.5883×10^{-7}	-2.1669×10^{-5}	-2.1835×10^{-5}	-4.9870×10^{-7}

References

1. B. Bilal, S. Ahmed, V. Kakkar, An insight into beyond CMOS next generation computing using quantum-dot cellular automata nanotechnology. International Journal of Engineering and Manufacturing **8**, 25 (2018)
2. B. Bilal, S. Ahmed, V. Kakkar, Quantum Dot Cellular Automata: A New Paradigm for Digital Design. International Journal of Nanoelectronics and Materials **11**, 87–98 (2018)
3. C.S. Lent, P.D. Tougaw, W. Porod, G.H. Bernstein, Quantum cellular automata. Nanotechnology **4**, 49 (1993)
4. C.G. Smith, Computation without current. Science **284**, 274–274 (1999)
5. F. Ahmad, S. Ahmed, V. Kakkar, G.M. Bhat, A.N. Bahar, S. Wani, Modular Design of Ultra-Efficient Reversible Full Adder-Subtractor in QCA with Power Dissipation Analysis. Int. J. Theor. Phys. **57**, 2863–2880 (2018)
6. B. Bilal, S. Ahmed, V. Kakkar, Modular adder designs using optimal reversible and fault tolerant gates in field-coupled QCA nanocomputing. Int. J. Theor. Phys. **57**, 1356–1375 (2018)
7. P.D. Tougaw, C.S. Lent, Logical devices implemented using quantum cellular automata. J. Appl. Phys. **75**, 1818–1825 (1994)
8. Ahmed, S., Bhat, S.M., Kakkar, V.: Design of Efficient 1-bit Comparator in Quantum dot Cellular Automata Nano-computing. In: 2020 IEEE-HYDCON, pp. 1–6. IEEE, (2020)
9. S.M. Bhat, S. Ahmed, Design of Ultra-Efficient Reversible Gate Based 1-bit Full Adder in QCA with Power Dissipation Analysis. Int. J. Theor. Phys. **58**, 4042–4063 (2019)
10. M. Raj, S. Ahmed, L. Gopalakrishnan, Subtractor circuits using different wire crossing techniques in quantum-dot cellular automata. J. Nanophoton. **14**, 026007 (2020)
11. I. Amlani, A.O. Orlov, G. Toth, G.H. Bernstein, C.S. Lent, G.L. Snider, Digital logic gate using quantum-dot cellular automata. Science **284**, 289–291 (1999)
12. S.E. Frost, A.F. Rodrigues, A.W. Janiszewski, R.T. Rausch, P.M. Kogge, Memory in motion: a study of storage structures in QCA, in *First Workshop on Non-Silicon Computing*
13. M.T. Niemier, P.M. Kogge, Logic in wire: using quantum dots to implement a microprocessor, in *ICECS'99. Proceedings of ICECS'99. 6th IEEE International Conference on Electronics, Circuits and Systems (Cat. No. 99EX357)* (IEEE, 1999), pp. 1211–1215
14. S. Ahmed, S.M. Bhat, S.-B. Ko, Design of reversible gate-based fingerprint authentication system in quantum-dot cellular automata for secure nanocomputing, in *The International Conference on Recent Innovations in Computing* (Springer, 2020), pp. 729–740
15. S. Ahmed, S.F. Naz, S.M. Bhat, Design of quantum-dot cellular automata technology based cost-efficient polar encoder for nanocommunication systems. Int. J. Commun. Syst. **33**, e4630 (2020)
16. S. Ahmed, S.F. Naz, S. Sharma, S.B. Ko, Design of quantum-dot cellular automata-based communication system using modular N-bit binary to gray and gray to binary converters. Int. J. Commun. Syst. **e4702** (2020)
17. R.P. Feynman, Quantum mechanical computers. Found. Phys. **16**, 507–532 (1986)
18. E. Fredkin, T. Toffoli, Conservative logic. Int. J. Theor. Phys. **21**, 219–253 (1982)
19. I. Manzoor, N. Nafees, M.I. Baba, S.M. Bhat, V. Puri, S. Ahmed, Logic design and modeling of an ultraefficient 3×3 reversible gate for nanoscale applications, in *International Conference on Intelligent Computing and Smart Communication 2019* (Springer, 2020), pp. 1433–1442
20. S. Ahmed, M.I. Baba, S.M. Bhat, I. Manzoor, N. Nafees, S.-B. Ko, Design of reversible universal and multifunctional gate-based 1-bit full adder and full subtractor in quantum-dot cellular automata nanocomputing. J. Nanophoton. **14**, 036002 (2020)
21. N. Nafees, I. Manzoor, M.I. Baba, S.M. Bhat, V. Puri, S. Ahmed, Modeling and logic synthesis of multifunctional and universal 3×3 reversible gate for nanoscale applications, in *International Conference on Intelligent Computing and Smart Communication 2019* (Springer, 2020), pp. 1423–1431
22. S. Saravanan, I. Vennila, S. Mohanram, Design and implementation of an efficient reversible comparator using TR gate. Circuits Syst. **7**, 2578 (2016)

23. A. Peres, Reversible logic and quantum computers. Phys. Rev. A **32**, 3266 (1985)
24. X. Ma, J. Huang, C. Metra, F. Lombardi, Reversible and testable circuits for molecular QCA design, in *Emerging Nanotechnologies* (Springer, 2008), pp. 157–202
25. S.M. Bhat, V. Kakkar, Design and modeling of an ultra-efficient 3x3 SSG-1 reversible gate for nanoscale applications, in *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)* (IEEE, 2021), pp. 720–723
26. V. Kumar, D. Dhawan, Design of reversible adder subtractor using multifunction reversible logic gate (MRLG). Int. J. Adv. Comput. Electron. Eng. **1**, 5–11 (2016)
27. A.K. Biswas, M.M. Hasan, A.R. Chowdhury, H.M.H. Babu, Efficient approaches for designing reversible binary coded decimal adders. Microelectron. J. **39**, 1693–1703 (2008)
28. M.S. Islam, M.M. Rahman, Z. Begum, M.Z. Hafiz, Low cost quantum realization of reversible multiplier circuit. Inf. Technol. J. **8**, 208–213 (2009)
29. H. Rangaraju, A.B. Suresh, K. Muralidhara, Design and optimization of reversible multiplier circuit. Int. J. Comput. Appl. **52** (2012)
30. T.A. Rather, S. Ahmed, V. Kakkar, Modelling and simulation of a reversible quantum logic based 4× 4 multiplier design for nanotechnology applications. Int. J. Theor. Phys. **59**, 57–67 (2020)
31. M. Haghparast, K. Navi, A novel reversible BCD adder for nanotechnology based systems. Am. J. Appl. Sci. **5**, 282–288 (2008)
32. K. Walus, T.J. Dysart, G.A. Jullien, R.A. Budiman, QCADesigner: a rapid design and simulation tool for quantum-dot cellular automata. IEEE Trans. Nanotechnol. **3**, 26–31 (2004)
33. P. Kumar, S. Singh, Optimization of the area efficiency and robustness of a QCA-based reversible full adder. J. Comput. Electron. **18**, 1478–1489 (2019)
34. A. Roohi, R. Zand, S. Angizi, R.F. DeMara, A parity-preserving reversible QCA gate with self-checking cascadable resiliency. IEEE Trans. Emerg. Top. Comput. **6**, 450–459 (2016)
35. F.S. Torres, R. Wille, P. Niemann, R. Drechsler, An energy-aware model for the logic synthesis of quantum-dot cellular automata. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **37**, 3031–3041 (2018)
36. C.S. Lent, M. Liu, Y. Lu, Bennett clocking of quantum-dot cellular automata and the limits to binary logic scaling. Nanotechnology **17**, 4240 (2006)
37. J. Timler, C.S. Lent, Power gain and dissipation in quantum-dot cellular automata. J. Appl. Phys. **91**, 823–831 (2002)
38. J. Timler, C.S. Lent, Maxwell's demon and quantum-dot cellular automata. J. Appl. Phys. **94**, 1050–1060 (2003)

Artificial Intelligence with Enhanced Prospects by Blockchain in the Cyber Domain

Praveen Kumar Singh

Abstract The various modern state-of-the-art applications with increasing utility of Artificial Intelligence (AI) have enabled human lives to thrive in a prevalent and convenient digital world. It has also led to the usage of huge amount of sensors data emanated out of different sources. A precise and scalable data analysis performed by AI to examine the colossal amount of data in a cyber space has potential to open yawing inroads for many future applications. At the same time, Blockchain being another promising technology can support a secure decentralized data communication network comprised of diverse independent nodes in a cyber domain. It has potential to overcome prevalent associated challenges of AI through resourceful data sharing. This paper brings out the specific facets of AI in a cyber domain facilitating IoT after giving brief overview on AI and blockchain. A marked progress in cryptographic techniques with strong support of AI in a cyber domain has opened the new avenues in digital applications. This paper aims in exploring viable data communication architecture with support of blockchain and AI by proposing it to support an extensive range of existing and upcoming smart applications. Both qualitative and quantitative examinations have been carried out to deal with computational complications and to evaluate the system performance to support the proposal. Lastly, the paper discusses about the future prospects and the possible challenges to be confronted in the cyber domain over an AI and blockchain integrated data communication network architecture.

Keywords Artificial intelligence (AI) · Blockchain · Cyber domain · Machine learning (ML) · IoT · Architecture · SDN · IPFS

P. K. Singh (✉)
Department of Computer Application, Shri Ramswaroop Memorial University, Lucknow-Deva Road, Barabanki, Tindola, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8_62

1 Introduction

Various prevalent smart applications have become a very convenient means in several routine human activities. Many applications warrant data processing in a cyber domain to enjoy the enormous dividends of faster speed with enhanced bandwidth. At the same time, they also become vulnerable to a few issues which include data connectivity, security, data centralization, GIS visualization, etc. In a cyber domain, AI holds the capability to thwart any potential data breaches [1]. Continued evolution in AI makes it an attractive tool for developers to employ it in a broad array of tasks like forecasting, training, decision making, predicament resolutions, etc. The major issues confronted by the common users relate to spams and distributed phishing over IoT applications. A colossal amount of high speed data being processed on internet demands a faster and precise analysis to support desired level of system security. Different architectures and Machine Learning (ML)-based data frameworks are being employed for this purpose.

AI-based solutions can offer an effective computation and system data analysis. It may be categorized in different groups of cloud, edging, fog and the apparatus wherein cloud bases system analysis employ a centralized cloud server which restricts the data speed with reduced latency and gets plagued with constraints of data storage space. The edging and fog-based system analysis use data load balancing with distributed data network architecture which facilitate in a scalable data analysis. The apparatus base AI system analysis offers more secure and optimized power consumption solution [2, 3]. The issues of data precision, centralized architecture, latency and other pertinent user concerns can be taken care of by integrating it with blockchain network where all the nodes are connected with hash function in a distributed network framework. In such systems, since all the blocks are interconnected, therefore, system hacking in blockchain is not possible for the inimical elements.

2 Overview of Artificial Intelligence and Blockchain

There is a huge potential of an integrated data network comprised of AI and blockchain in a cyber domain. These two are the core technologies in many widespread digital applications. Processing and the deliberate examination of the data transmission in the smart applications are being done through the AI which in turn provides an intelligent decision taking means to the human by the machines. In order to facilitate a better degree of security, blockchain over distributed and decentralized platform is employed [3]. A brief overview of these two technologies is given in the succeeding paragraph.

2.1 Artificial Intelligence

A set of process employed by AI consists of the aggregation of data, examine it to figure out the learning pattern and then to utilize such patterns to produce with some tangible predictions and solutions. ML, deep learning and robotics are considered a segment of AI. There are numerous examples in healthcare, communication, automated vehicle driving, television recording, mechanized farming, power conservation, weather forecasting, banking sector, etc. [4]. In all these solutions, an apparatus is used to record some events and then based on the thorough examination of the patterns occurred in the particular event, predictions are evolved to determine the outcome of the AI solution. It enables the system to detect unusual activities, if any to mitigate the security threats.

In other words, any machine which performs an automatic function to include comprehends, learn, examine and offer an automated solution can be referred to as AI. A basic AI model has been illustrated below in Fig. 1. A different set of AI applications like data processing, speech, vision, robotics, system expertise and its analysis and machine learning have been indicated to provide a general overview to conceptualize AI [4, 5]. Certain challenges confronted by AI in a cyber domain relate to privacy and security issues, efficient usage of energy, traffic blockage, etc., are addressed through deep learning and ML.

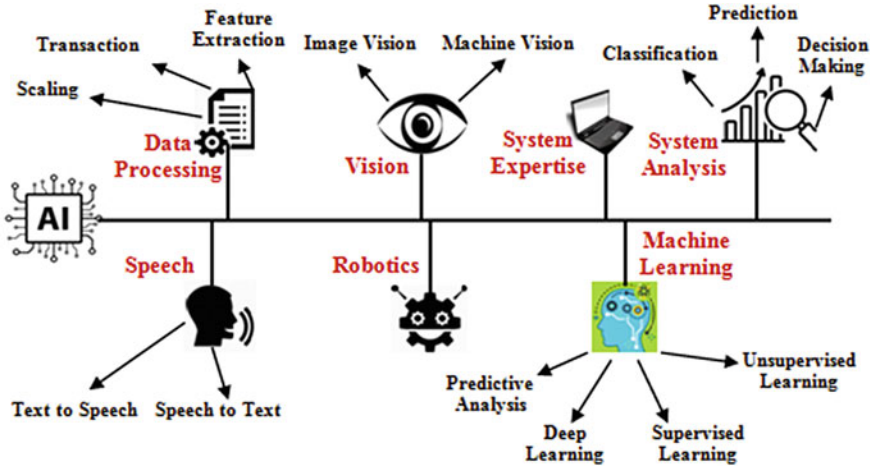


Fig. 1 Basic Model of Artificial Intelligence

2.2 *Blockchain*

Blockchain is comprised of a chain of four essential groups of blocks which are linked together namely transaction details, value of hash both in existing and preceding blocks and the timestamp. Instead of centralized data network architecture with its inherent drawbacks, blockchain supports a distributed and decentralized architecture which assists in conserving number of transactions taking place at different nodes within that data network [6]. A formidable security system is provisioned in these blocks as they possess the cryptographic codes consisting of hash function of own as well as of other blocks as well. It prevents any possibility of data alteration by an unknown evader in the system. The transaction linkages of the different blocks have been illustrated in Fig. 2. In blockchain, the provision of decentralized data storage by interplanetary File System (IPFS) has tremendous advantage with higher throughput to store a large amount of data compare to any centralized storage system. Various blocks are linked here in a peer-to-peer distributed network through a contract code.

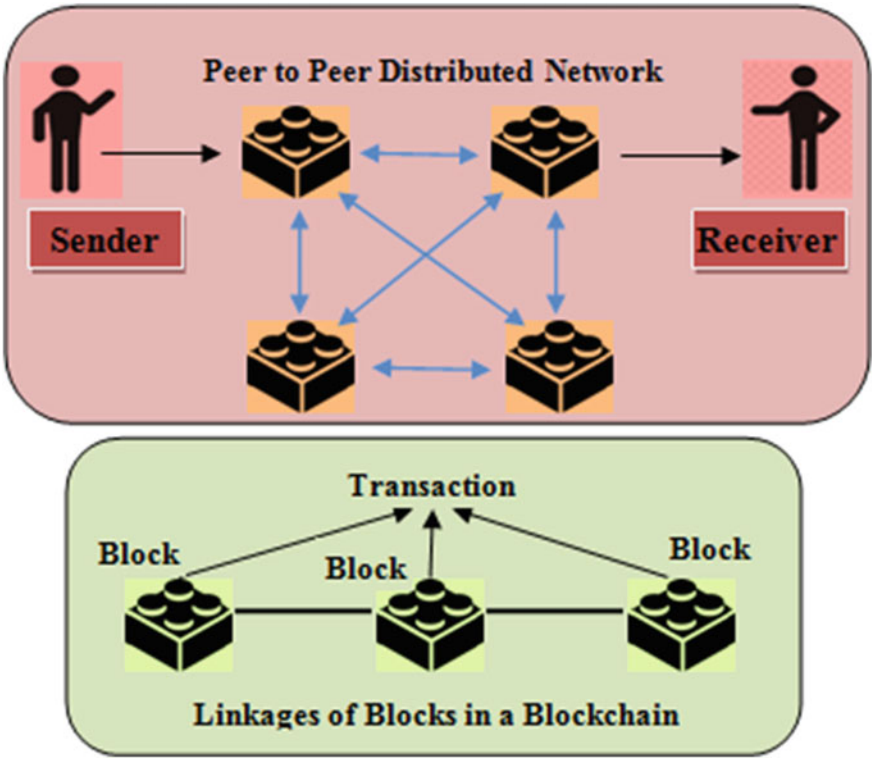


Fig. 2 Basic model of blockchain

A large quantum of data in cyber domain has an inherent drawback to have its own security challenges. Blockchain offers relatively better security architecture in the distributed data network with enhanced fault tolerance and computational speeds.

3 3. Artificial Intelligence (AI) Facilitated by IoT in a Cyber Domain

In the cyber domain with increased usage of internet, corresponding vulnerability against the cyber attacks also increases. In order to provide AI-based solutions adequate safeguards, cyber security is a significant aspect to protect the network peripherals from any impending cyber threats. In fact, AI itself has a vital role to play to ensure desired level of cyber security. The growth of AI applications is largely dependent upon the confidence gained by the users in its security system. With constant evolution of the internet, AI techniques in the cyber domain too are likely to be evolved to support all the critical applications in human well-being [6, 7]. The approach to address the issues in digital world has been that the machines should not only be able to offer tangible solutions, they should be able to examine them if required like humans. The major diverse AI application realms in cyber domain are being discussed in succeeding paragraphs.

3.1 Cyber Attacks and Human Linkages

The most common means adopted by the cyber attackers to forward fake websites and emails to common users which are generally deemed as the legitimate ones by them. In order to discern the legitimacy of the target, an expertise is needed and that is where AI can play a key role in offering an alternative to human intelligence [7]. Different features of the application like SSL certificate, web URL and its length, IP address, predefined blacklisted keywords are employed to detect linkages which could lead to cyber frauds. For instance, in a Java scripted AI solution, coding characters distribution, word length, sensitive calls, frequency of code occurrences are just few examples which can be employed to surmount human limitations in examining and detecting the cyber attacks.

3.2 Data Traffic in IoT

In order to succeed any AI solution, the accuracy with which the wicked patterns of the cyber attackers can be identified has to be of highest level. Human weaknesses in terms of lack of cyber security awareness are required to be addressed in AI

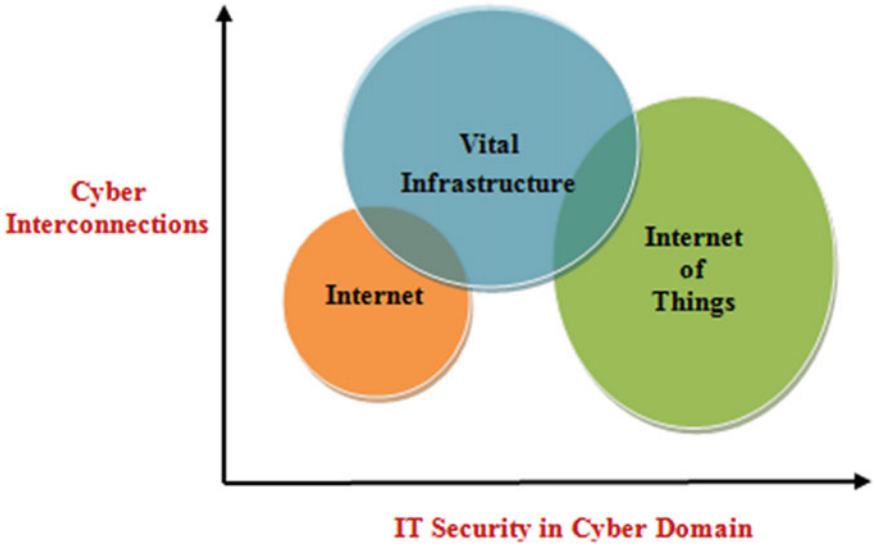


Fig. 3 AI applicability in different cyber applications domain

solutions where the precision of predictions in the associated applications have strong implications in its accomplishment. In IoT, the malicious patterns appear just like the legitimate ones [8]. The significant aspect of such AI-based solution relates to its capability to examine the massive quantum of data traffic over the internet. Figure 3 below illustrates applicability of AI over various cyber application domains. Performance of cyber interconnections in relation to their security in cyber domain is measured in overall paradigm of IoT and its associated infrastructure.

3.3 Vital Network Infrastructure

There are certain infrastructures which are considered vital from national security perspective. It might be power, telecommunications, nuclear, water resources, gas, oil, etc. Since almost all such vital assets are linked one way or other through the cyber connectivity, therefore provisioning of adequate safeguards are quite crucial for them. AI has potential to become a major facilitator for these assets. A robust intrusion recognition mechanism with good cyber security arrangement within the Software-Defined Network (SDN) becomes an indispensable requirement for all such prevalent vital applications [9]. The advantage of an SDN system is that it examines the network data against any cyber security violation before routing it further within the network. It employs AI to analyze the data traffic pattern for making further security predictions. AI facilitates the desired output through different real-time computational iterations to recognize any illegitimate data traffic.

4 The Proposed AI Architecture Supported by Blockchain

The primary function of an AI supported system is to learn from the existing pattern, make necessary amendments based on system's requirements and then carry out the associated function as human automatically. Various simulated intelligent functions in a cyber domain are undertaken by AI-based systems to perform different predictive, adaptive, prescriptive and continuous analysis. Still there are certain limitations in the output of such systems which may warrant its integration with blockchain for a scalable and secure solution for the post as well as the real-time system processing [6, 9]. It necessitates devising system architecture by integrating AI with blockchain which can incorporate various parameters to address the issues of technology, cryptographic algorithm, software, assessment parameters, etc.

A flow chart above in Fig. 4 below illustrates the contours of required system architecture. The system hierarchy in a cyber domain must conform to certain essential parameters like data effectiveness, security and privacy, data access, optimized energy consumption, etc. Blockchain facilitated AI architecture must be dynamic, productive and technically effective. It should have provision of adequate security and to protect the interests of users by addressing their privacy concerns [10]. It should also cater to a scalable data acquisition which can be tracked during its entire transmission process. In the whole data network architecture, efficiency of the system also relies on the minimum power consumption right from its inception to convergence and data aggregation process. In order to formulate the proposed AI data network architecture which can be facilitated by the blockchain in a cyber domain, we need to identify different sets of network applications. Accordingly, diverse intelligence

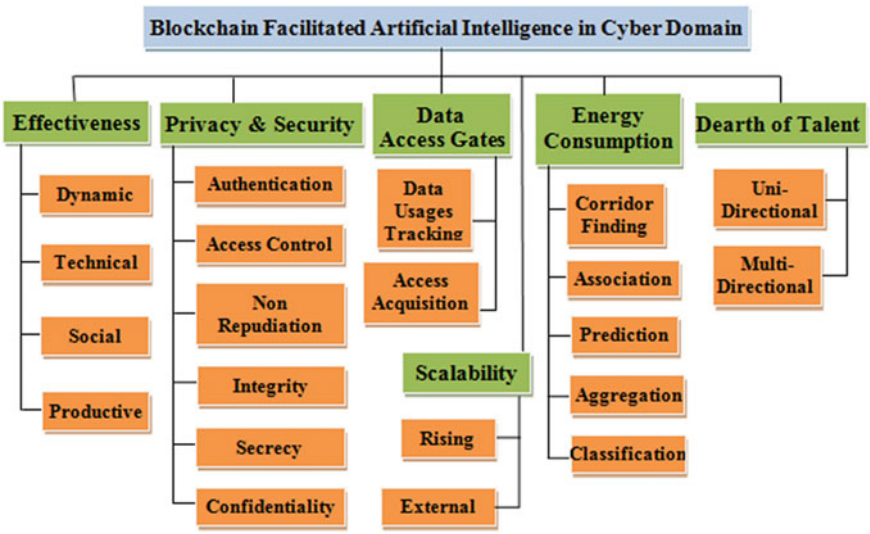


Fig. 4 Blockchain facilitated AI Hierarchy in a cyber domain

applications are proposed in four separate intelligence architecture in the proposal namely apparatus, edge, and fog and cloud intelligence [11]. These are the intelligence applications over cyber domains which assist in security, data examinations and other data centralization matters. In an intelligence data network hierarchy, the initial apparatus intelligence layer is comprised of different IT peripherals which generate enormous quantum of data that is subsequently passed on to edge intelligence. Here, there are number of base stations over the data network edges connected via blockchain linked to diverse sensors to examine them the data traffic emanated out of associated sensing devices.

The proposed blockchain-supported AI architecture is illustrated above in Fig. 5.

Further, the network data flows to fog intelligence comprised of different AI-based blockchain integrated fog nodes. All such blockchain linked fog nodes are connected to AI-based base stations and process the network data further to cloud intelligence.

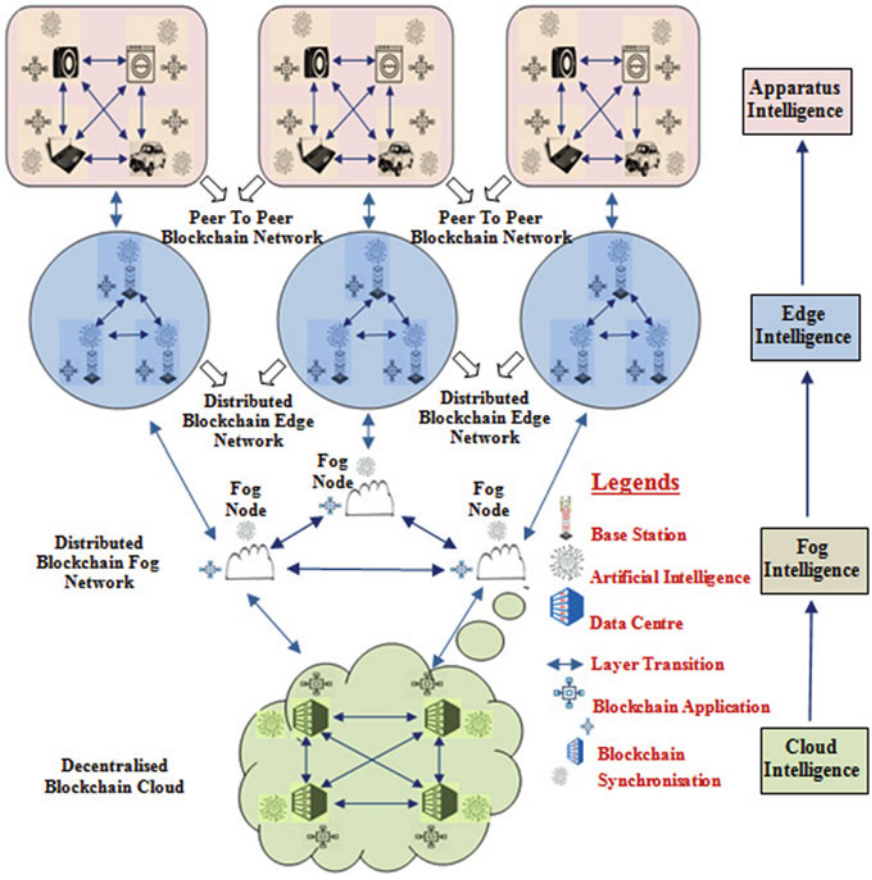


Fig. 5 Proposed blockchain supported AI Architecture

This layer is comprised of AI-based data centers having links to blockchain to facilitate secure and decentralized data for a range of smart applications. These data centers share their data processing results with cloud intelligence to enable them to exercise control over the massive data in the network [12]. Sharing of data by the data centers assists this layer finally against various privacy, latency and precision issues in the proposed blockchain integrated AI-based data network. Convergence of such blockchain enabled AI applications employ a number of consented network protocols.

5 Examination of Proposed AI Based Blockchain Intelligence

AI and blockchain are considered significant technologies for large number of prevalent smart applications. The proposed architecture is an endeavor to address many prominent challenges of AI like explicable AI, data distribution, artificial trust, etc., by integrating it with blockchain. Conversely, there is several scalability, power consumption and security-related challenges of blockchain which are effectively addressed through the proposed network architecture. Table 1 below examines different hierarchal network layers proposal against some specific parameters of associated application, assessment considerations, technological usage and the proposed methodologies [8, 12]. The proposed architecture is supposedly taking care of system

Table 1 Examination of the proposed architecture

Type of intelligence	Parameters			
	Associated application	Assessment considerations	Technology usage	Proposed methodology
Apparatus intelligence	Entity recognition	Precision, security and privacy, computational convolutions	Deep learning, blockchain	Deep learning based on secure blockchain
Fog intelligence	Attack recognition	Precision, computational sources	Machine learning, SDN, blockchain	Secure decentralization based on blockchain
Edge intelligence	Entity recognition	Precision, Latency impediments, privacy and security dimensions	Deep learning, blockchain	Deep learning through secure blockchain
Cloud intelligence	Weather prediction	Source administration, power utilization	Reinforcement learning, blockchain	Source administration based on blockchain

efficiency by boosting the blockchain in its decision driving and prediction analysis. Now, let us examine our above-mentioned AI architecture proposal integrated with blockchain both in qualitative as well as in quantitative terms of references.

5.1 Qualitative Examination

While doing the qualitative examination of the proposal, we find that incorporation of blockchain simplifies and facilitates in faster data processing speed. It also offers a distributed data sharing within the network to do away with the centralized system. The network architecture relates to number of evaluation parameters to include what type of AI and blockchain technique being used, software, cryptographic algorithm, etc. [13]. In apparatus intelligence layer, system performance is examined over precision, security, time lag and computational complications. While employing deep learning ML methods, supplementary memory and CPU may be required. At edge layer, as per the quantum of edge nodes in the network, object recognition time is measured. With shorter recognition time, system displays relatively better efficacy. In next fog layer, SDN is employed in examining input data which facilitates in brilliant attack recognition model [14]. Overall the network architecture is examined based on data recognition rate, precision, prognostic rate, computational sources, etc. It is found to be efficient decentralized network architecture with excellent computational efficacy, better data storage capacity and reduced power consumption.

5.2 Quantitative Examination

Quantitative examination of the proposed network architecture is done through the relative comparison of different intelligence layers in relation to their precision and latency [14]. Figure 6 above illustrates the subject comparison of apparatus, edge and fog intelligence layers both with and without inclusion of blockchain to present a fair idea about how effective it would be. While employing the blockchain, precision value varies from zero to 68%, 72% and 88%, respectively for apparatus, edge and fog intelligence layers. On the other hand, the same value without blockchain varies from zero to 58%, 62% and 78%, respectively. It signifies that the precision of data value would always be better while integrating blockchain with AI applications. Likewise, while examining the latency value with blockchain, we find it from zero to 58 ms, 63 ms and 78 ms, respectively for apparatus, edge and fog intelligence layers. At the same time these values without blockchain varies from zero to 42 ms, 48 ms and 8 ms, respectively [15]. It indicates that while using blockchain, latency is higher. Therefore, precision and security in apparatus intelligence is found relatively higher compared to other layers in hierarchy of the proposed data network architecture.

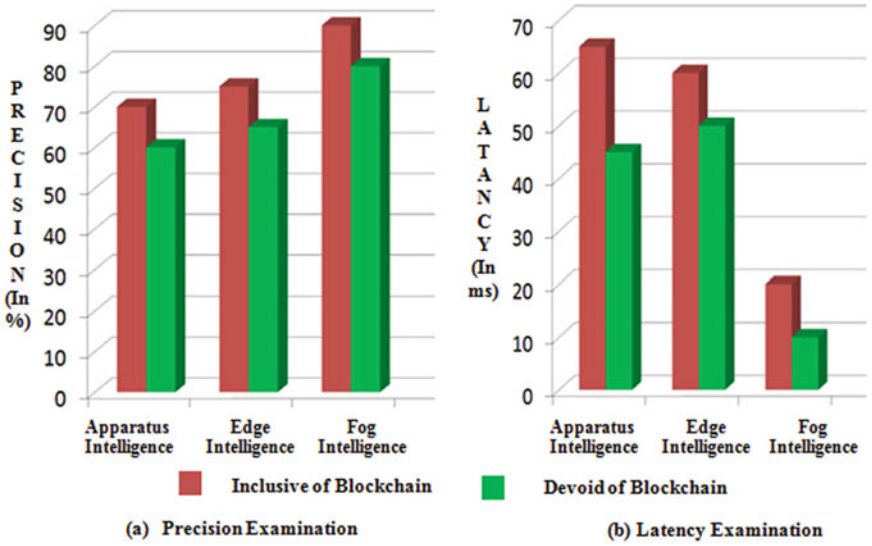


Fig. 6 Relative precision and latency comparison of proposed AI Architecture

6 Future Prospects and Challenges

The key dividend accrued out of the above proposed integration of AI with blockchain is the data storage in a decentralized distributed data network to establish a blockchain set up. This data network arrangement facilitates in transparency and real-time tracking in a data cloud of artificial intelligence against desired piece of information. However, still there are number of issues in AI and blockchain integration which will need deliberations by future research scholars to deal with. Different intrusion recognition systems have been facilitated by AI which assists in making precise distinction between the genuine and the malicious data traffic in IoT [16, 17]. Still, system security and user privacy remains a concern and development of better encryption algorithm is supposedly a key research area in such systems.

The user's trust is a key in any database mechanism where their personal credentials are accumulated. It must address their privacy concerns. Incorporating AI solutions too require regulatory and legal backing to protect user rights. Moreover, many prevalent applications like crypto-currency system, SON, bitcoin CNN, self-evolution networks, etc., may not be fully adaptable by AI and blockchain technologies primarily due to a very high order of security needs [18]. By and large, most of the AI-based solutions deal with a huge sum of data wherein there is always a probability that exists for data getting hacked or corrupted. In such cases, it would be pretty difficult for blockchain to evaluate them; however, it does offer the associated security cover in a decentralized database as well as scope to strengthen the system security in future.

There is a definite need of a comprehensive data analysis in the AI and blockchain integrated system [19]. In larger space applications, better security and scalability solutions will be needed for enhanced communication security requirements. It is likely to address the data mining and data conflict issues during the data communication thereby increasing the system efficacy. Since, in future the data traffic volume is likely to raise exponentially, therefore development of more effective computational resources will certainly be an inevitable need to resolve complicated computational issues for the future AI and blockchain integrated solutions [20]. As all AI-based apparatus consume quite a large energy to function, therefore an easier access of the desired applications through more energy efficient systems has to be a focus area for the future scholars.

7 Conclusion

With growing AI-based solutions demand and incorporating blockchain massively in a cyber domain, the integration of these two technologies showcase how effectively a decentralized data network mechanism can benefit prevalent smart applications. At the same time, there is an increase in likelihood of cyber attacks which will require a formidable system security support [21]. The distributed data storage in a data network offers an opportunity to network administrators to store a large quantum data in diverse applications by convergence of AI with blockchain in a cyber domain. However, a high quality taxonomy composition is required to be formed to alleviate the inadequacy and computational complexities in such systems. In this proposed distributed data cloud network architecture, each block employs smart contracts for peer-to-peer data transmission and does not need any centralized data hub. It leads to a faster and reliable data transaction due to the proposed AI and blockchain integrated network architecture [22]. However, such AI convergence with blockchain is likely to remain plagued with certain issues like interoperability, integrating heterogeneous data structures, energy efficacy, cost factor and many others which will through a good challenge to system developers to come to terms to prevailing requirements of such data network architectures.

References

1. Z. Zheng, S. Xie, H.N. Dai, X. Chen, H. Wang, Blockchain challenges and opportunities: a survey. Int. J. Web Grid Serv. **14**(4), 352–375 (2018)
2. A. Tiwari, U. Batra, Coalition of IoT and blockchain: rewards and challenges, in *6th International Conference on Computing for Sustainable Global Development (INDIACom)* (2019), pp. 327–332
3. R. Kumar, R. Tripathi, Scalable and secure access control policy for healthcare system using blockchain and enhanced Bell–LaPadula model. J. Ambient Intell. Human. **12**(2), 2321–2338. Springer, Berlin (2021)

4. P. Mamoshina, L. Ojomoko, Y. Yanovich, A. Ostrovski, A. Botezatu, P. Prikhodko, E. Izumchenko, A. Aliper, K. Romantsov, A. Zhebrak, Converging blockchain and next-generation artificial intelligence technologies to decentralize and accelerate biomedical research and healthcare. Onco Target **9**(5), 56–65 (2018)
5. W. Xiong, L. Xiong, Smart contract based data trading mode using blockchain and machine learning. IEEE Access **7**(10), 331–344 (2019)
6. R. Kefa, Convergence of AI, IoT, big data and blockchain: a review. Lake Inst. J. **1**(1), 1–18 (2018)
7. N. Kshetri, Can blockchain strengthen the internet of things? IT Profess. **19**(4), 68–72 (2017)
8. M. Banerjee, J. Lee, K.K.R. Choo, A Blockchain future for the internet of things security: a position paper. Digital Commun. Netw. **4**(3), 149–160 (2018)
9. C. Xu, K. Wang, M. Guo, Intelligent resource management in Blockchain based cloud data centers. IEEE Cloud Comput. **4**(6), 50–59 (2017)
10. A. Banafa, IoT, AI and blockchain: catalysts for digital transformation. IEEE Internet Things Mag. **1**(1) (2018)
11. T. Yang, F. Zhai, J. Liu, M. Wang, H. Pen, Self-organized cyber-physical power system Blockchain architecture and protocol. Int. J. Distrib. Sens. Netw. **14**(10), 69–74 (2018)
12. S. Rafati Niya, L. Pelloni, S. Wullschleger, A. Schaufelbühl, T. Bocek, L. Rajendran, A Blockchain-based scientific publishing platform, in *Proceedings of the IEEE International Conference on Blockchain and Cryptocurrency, ICBC19* (Seoul, South Korea, 2019)
13. P.K. Sharma, S. Rathore, Y.S. Jeong, J.H. Park, SoftEdgeNet: SDN based energy-efficient distributed network architecture for edge computing. IEEE Commun. Mag. **56**(12), 104–111 (2018)
14. A. Pieroni, N. Scarpato, L. Di Nunzio, F. Fallucchi, M. Rasho, Smarter city: smart energy grid based on Blockchain technology. Int. J. Adv. Sci. Eng. Inf. Technol. **8**(1), 298–306 (2018)
15. A. Tiwari, U. Batra, IPFS enabled blockchain for smart cities. Int. J. Inf. Technol. **13**(7), 201–211 (2021)
16. J. Mata, I. De Miguel, R.J. Duran, N. Merayo, S.K. Singh, A. Jukan, M. Chamania, Artificial intelligence (AI) methods in optical networks: a comprehensive survey. Opt Switch Netw. **28**(4), 43–57 (2018)
17. J. Hoey, T. Schröder, J. Morgan, K.B. Rogers, D. Rishi, M. Nagappan, Artificial intelligence and social simulation: studying group dynamics on a massive scale. Small-Group Res. **49**(6), 647–683 (2018)
18. M.A. Khan, K. Salah, IoT security: review, blockchain solutions, and open challenges, future generation computers. System **82**(19), 395–411 (2018)
19. D.R. Reyes, H.V. Heeren, S. Guha, L.H. Herbertson, A.P. Tzannis, J. Ducrée, Accelerating innovation and commercialization through standardization of microfluidic-based medical devices. Lab. Chip. **21**(1), 9–21 (2021)
20. A. Tiwari, C. Soni, Hardware trojans: an austere menace ahead, in *Cyber Security, Advances in Intelligent Systems and Computing*, vol. 729, ed. by M. Bokhari, N. Agrawal, D. Saini, vol. 729. (Springer, Singapore, 2018)
21. A. Kalla, T. Hewa, R.A. Mishra, M. Ylianttila, M. Liyanage, The role of Blockchain to fight against COVID-19. IEEE Eng. Manag. Rev. **48**(3), 85–96 (2020)
22. I. Sood, H. Pirkkalainen, A. Camilleri, Can Blockchain technologies facilitate the unbundling of higher education, in *Proceedings of the 12th International Conference on Computer Supported Education (CSEDU SCITEPRESS)*, vol. 2 (2020), pp. 228–235
23. T. Blakely, J. Lynch, K. Simons, R. Bentley, S. Rose, Reflection on modern methods: when worlds collide-prediction, machine learning and causal inference. Int. J. Epidemio **49**(6), 2058–2064 (2021)
24. S. Huang, J. Yang, S. Fong, Q. Zhao, Artificial intelligence in cancer diagnosis and prognosis: opportunities and challenges. Cancer Lett. **471**(3), 61–71 (2020)
25. T. Stevens, Knowledge in the grey zone: AI and cyber security. Digital War. **32**(1), 164–170 (2020)

26. J. Pawlick, E. Colbert, Q. Zhu, A game-theoretic taxonomy and survey of defensive deception for cyber security and privacy. ACM Comput. Surv. (CSUR), **52**(4), 82–93 (2019)
27. A.F. Hussein, N.A. Kumar, G.R. Gonzalez, E. Abdulhay, J.M.R. Tavares, V.H.C. Albuquerque, A medical records managing and securing Blockchain-based system supported by a genetic algorithm and discrete wavelet transform. Cogn. Syst. Res. **52**(7), 1–11 (2018)

Author Index

A
Abdul Basit, 315
Abhiram Anand Gulanikar, 741
Abhiti Sachdeva, 637
Adeniyi, Abidemi Emmanuel, 613
Afreen Hasan, 537
Ajay Kumar Garg, 753
Akshat Agrawal, 741
Amar Singh, 411
Amit Kumar Singh, 573
Anamol Verma, 489
Anil Kumar Dubey, 587
Anil Sharma, 27
Ankur Chaturvedi, 397
Ankur Omar, 537
Anupam Singh, 489
Anushka Sharma, 111
Apurva Agarwal, 439
Aruna Malik, 53, 139, 249
Ashwani Kumar, 547
Avtar Singh, 669

B
Bakonyi, Viktória, 625
Balwinder Singh, 795
Bhat, G. M., 315
Brahm Prakash Dahiya, 339

C
Charumathi, K. S., 3
Cinkler, Tibor, 277

D
Daanish Goyal, 637
Damasevicius, Robertas, 385, 741
Deepti Agrawal, 439
Devaashish Sharma, 357
Dhruv Shah, 453
Dilip Kumar Sharma, 439
Dilip Sharma, 425
Dipansh Mittal, 489
Diptendu Bhattacharya, 125
Domian, Balint, 149

F
Farou, Zakarya, 149

G
Gaurav Sharma, 263
Gayathri, K. M., 225

H
Hamza Abubakar Kheruwala, 67
Hanuman Agrawal Das, 397
Harsha Vardan Maddiboyina, 179
Himanshu Gupta, 385, 613
Himanshu Sharma, 573
Hitesh Vora, 371
Horváth, Tomáš, 149

I
Illés, Zoltán, 599, 625, 645, 659
Indranath Chatterjee, 125

J
Jai Prakash Verma, 67, 453

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

P. K. Singh et al. (eds.), *Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 855, https://doi.org/10.1007/978-981-16-8892-8

Jasleen Kaur, 339
Javaid A. Sheikh, 315
Jawthari, Moohanad, 683
Jinesh Melvin, Y. I., 3
Jyoti, 207
Jyoti Hatte, 225

K
Kamel, Mohammed B. M., 289, 725
Kapil Mehta, 695
Karan Bagoria, 695
Kartik Bhatia, 489
Kartik Gupta, 695
Karuna Sheel, 27
Kehinde, Adeniyi Jide, 613
Khandakar Faridar Rahman, 95
Kiran Srivastava, 13
Kirti Pal, 303, 327
Korom, Szilárd, 645

L
Lakshya Sharma, 695
Leena Sharma, 13
Lewu, Olatunde Petwilson, 753
Ligeti, Peter, 289, 725, 767
Lipika Gupta, 795

M
Mala Saraswat, 85, 587
Manav Chauhan, 695
Manu Sonia, 711
Manu Sood, 357
Marwah Yaseen, 725
Maskeliunas, Rytis, 385, 753
Meghansh Bansal, 397
Misra, Sanjay, 385, 613, 741, 753
Mohammed S. Ahmad, 67
Mohit Kumar Saini, 163
Monika Arora, 711
Mudasir Hamid Sheikh, 467
Muhammad-Bello, Bilkisu Larai, 753

N
Narbada Prasad Gupta, 37, 239
Naresh Kumar, A., 179
Neha Thakur, 669
Nidhi, 327
Nishtha, 357
Nitasha Bisht, 189

O
Ogundokun, Roseline Oluwaseun, 385, 613, 741
Oluranti, Jonathan, 753
Ouaari, Sofiane, 149

P
Pankaj Rahi, 501, 521
Partha Pratim Deb, 125
Parulpreet Singh, 239
Pawan Singh Mehra, 53
Pooja Dhiman, 563
Poornima Tyagi, 477
Pradeep Kumar Singh, 53, 67, 249
Pradeep Verma, 477
Pratyush, 111
Praveen Kumar Malik, 13, 189
Praveen Kumar Singh, 779, 827
Praveen Malik, 37
Preeti Kathiria, 371
Priya Shrivastava, 425
Punam Rattan, 339
Puneet Chandra Srivastava, 13
Punit Gupta, 637

R
Rahul Sharma, 411
Rajab, Husam, 277
Rakesh Kumar Saini, 163
Raman Kapoor, 587
Ranjan Kumar, 111
Reenu, 339
Reich, Christoph, 289
Renu Bahuguna, 85
Rintu Khanna, 573
Rishi Saraswat, 85
Rishu Gupta, 587
Ritu Dewan, 95
Rohit Bajaj, 501
Rumaan Bashir, 467
Rupali B. Patil, 225
Ruqsar Zaitoon, 547

S
Sadiku, Peter Ogirima, 385
Sajal Maheshwari, 111
Sajjan Singh, 207
Samayveer Singh, 53, 249
Sandeep Rathor, 537
Sangal, A. L., 669
Sanjay K. Sharma, 521

Sanjay Kumar Sahu, 239
Sanjay P. Sood, 501, 521
Sankar Ponnapalli, V. A., 179
Santosh Kumar Henge, 563
Saumya Borwankar, 453
Segun-Owolabi, Tobe, 741
Shaminder Kaur, 795
Shailender, 207
Shelej Khera, 207, 239
Shivleela Mudda, 225
Smita Agrawal, 371
Soha Maqbool Bhat, 809
Sonu Mittal, 467
Stoffa, Veronika, 683
Stoffová, Veronika, 659
Sudeep Tanwar, 67, 453
Suhaib Ahmed, 315, 809
Sushopti Gawade, 3, 263
Syed Saba Raoof, 547
Szabó, Dávid, 599
Szabó, Tibor, 625

T
Tiansheng, Xi, 277
Tripti Kunj, 303
Tushar Tyagi, 573

U
Umhara Rasool Khan, 315
Usha Patel, 371
Utkarsh Pandey, 37

V
Vandana Mohindru Sood, 695
Verma, Chaman, 659
Vipan Kakkar, 809
Vikram Rajpoot, 397
Vinayak Rai, 695
Vivechana Maan, 139

Y
Yagya Malik, 637
Yan, Yuping, 767