

A. Pasumpon Pandian  
Xavier Fernando  
Wang Haoxiang *Editors*



# Computer Networks, Big Data and IoT

Proceedings of ICCBI 2021

# **Lecture Notes on Data Engineering and Communications Technologies**

Volume 117

## **Series Editor**

Fatos Xhafa, Technical University of Catalonia, Barcelona, Spain

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

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

Indexed by SCOPUS, INSPEC, EI Compendex.

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

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

A. Pasumpon Pandian · Xavier Fernando ·  
Wang Haoxiang  
Editors

# Computer Networks, Big Data and IoT

Proceedings of ICCBI 2021



Springer

*Editors*

A. Pasumpon Pandian  
CARE College of Engineering  
Trichy, India

Wang Haoxiang  
Go Perception Laboratory  
Cornell University  
Ithaca, NY, USA

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

ISSN 2367-4512

ISSN 2367-4520 (electronic)

Lecture Notes on Data Engineering and Communications Technologies

ISBN 978-981-19-0897-2 ISBN 978-981-19-0898-9 (eBook)

<https://doi.org/10.1007/978-981-19-0898-9>

© 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

*We are honored to dedicate the proceedings  
of ICCBI 2021 to all the participants and  
editors of ICCBI 2021.*

# Preface

This conference proceedings volume contains the written versions of most of the contributions presented during the conference of ICCBI 2021. The conference provided a setting for discussing recent developments in a wide variety of topics including computer networks, big data, and Internet of things. The conference has been a good opportunity for participants coming from various destinations to present and discuss topics in their respective research areas.

This conference tends to collect the latest research results and applications on computer networks, big data, and Internet of things. It includes a selection of 73 papers from 286 papers submitted to the conference from universities and industries all over the world. All of the accepted papers were subjected to strict peer-reviewing by 2–4 expert referees. The papers have been selected for this volume because of quality and the relevance to the conference.

We would like to express our sincere appreciation to all authors for their contributions to this book. We would like to extend our thanks to all the referees for their

constructive comments on all papers; especially, we would like to thank the organizing committee for their hard work. Finally, we would like to thank the Springer publications for producing this volume.

Dr. A. Pasumpon Pandian  
Professor and Dean (R&D)  
CARE College of Engineering  
Trichy, India

Dr. Xavier Fernando  
Director, Ryerson Communications  
Lab, Professor, Department  
of Electrical and Computer  
Engineering  
Ryerson University  
Toronto, ON, Canada

Dr. Wang Hoaxing  
Go Perception Laboratory  
Cornell University  
Ithaca, NY, USA

# **Acknowledgments**

The ICCBI Conference Committee would like to thank all of the volunteers who have helped to organize this event. The organizers also wish to acknowledge publicly the valuable services provided by the reviewers.

On behalf of the editors, organizers, authors, and readers of this conference, we wish to thank the keynote speakers and the reviewers for their time, hard work, and dedication to this conference. The organizers wish to acknowledge Shri B. Prative Chend, Dr. S. Shanthi, Dr. S. Vijayarangan, and Dr. J. Jeyarani for the discussion, suggestion, and cooperation to organize the keynote speakers of this conference. The editors also wish to acknowledge the speakers Dr. Manu Malek, Alcatel-Lucent Bell Labs and Stevens Institute of Technology (ret.), New Jersey, USA, and Dr. Shajulin Benedict, Indian Institute of Information Technology, Kottayam, India, and participants who attended this conference. ICCBI 2021 would like to acknowledge the contribution made to the organization by its many volunteers. Members contribute their time, energy, and knowledge at a local, regional, and international level.

We also thank all the chairpersons and conference committee members for their support.

# Contents

<b>Rakshak—Prototype Application to Depict Crime Hotspot and Safest Path Between Locations .....</b>	1
Archana Naik, Aakanksha Gaur, Stuti Srivastava, Shashikant Saini, R. Gayathri, and Kavitha Sooda	
<b>Blockchain Technology in Application Development and Associated Challenges .....</b>	17
Vishal Polara, Pooja Bhatt, Dharmesh Patel, and Ketan Rathod	
<b>Hybrid Statistical and Deterministic-Based Pedestrian Tracking Algorithm for Location-Based IoT Applications .....</b>	29
J. P. D. Manoj Sithara and M. W. P. Maduranga	
<b>MAC-Based Secure Data Transmission in Vehicular Ad hoc Networks .....</b>	39
T. Kalaichelvi, L. Jabasheela, P. S. Ramaprabha, M. Shobana, G. Dhanalakshmi, W. Gracy Theresa, and H. Rashini	
<b>A Study on Challenges in Data Security During Data Transformation .....</b>	49
K. Devaki and L. Leena Jenifer	
<b>An Edge-Based Disjoint Path Selection Scheme for FANETs .....</b>	67
Orchu Aruna and Amit Sharma	
<b>Sign Language Interpreter .....</b>	83
Ramya Srikantheswara, C. B. Niveditha, A. Sindhu Sai, Reddigari Keerthi Reddy, and S. A. Akshayanjali	
<b>Architectural Insight of Neural Information Extraction, Retrieval, and Processing for Multimodal Neural Search .....</b>	93
Jainal S. Gosaliya, Adarsh K. Gupta, Akshay Ashok, and Swapnil M. Parikh	

<b>A Comparative Study Analysis on Air Monitoring and Purification Systems .....</b>	111
Nida Praveen, Lipika Goel, and Sonam Gupta	
<b>Malicious Node Detection and Prevention for Secured Communication in WSN .....</b>	121
Vaibhav Dabhade and A. S. Alvi	
<b>Performance Analysis and Assessment of Various Energy Efficient Clustering-Based Protocols in WSN .....</b>	137
Trupti Shripad Tagare and Rajashree Narendra	
<b>Academic Data Analysis and Projection Using Artificial Intelligence .....</b>	155
K. Kanagaraj, Joyce R. Amirtharaj, and K. Ramya Barathi	
<b>Analysis of Factor Verification Affecting Recruitment Process Through Social Dynamics .....</b>	171
Krishna Kumar Singh and Priyanka Srivastava	
<b>Routing Protocols in an Opportunistic Network: A Survey .....</b>	185
Sushil Kumar Mishra and Ruchika Gupta	
<b>K-Splits: Improved K-Means Clustering Algorithm to Automatically Detect the Number of Clusters .....</b>	197
Seyed Omid Mohammadi, Ahmad Kalhor, and Hossein Bodaghi	
<b>Implementation of IoT-Based Intelligent Patient Healthcare Monitoring System Using KNN Algorithm .....</b>	215
G. Sreenivasulu and T. P. Anithaashri	
<b>Comprehension of Ultra-Wideband Transceiver for Wireless Communication System based on Code-Shifted Reference .....</b>	223
R. Santosh Kumar, Rajashree Narendra, and R Devaraju	
<b>Optimisation of the Execution Time Using Hadoop-Based Parallel Machine Learning on Computing Clusters .....</b>	233
B. V. V. Siva Prasad, G. Sucharitha, K. G. S. Venkatesan, Tulasi Radhika Patnala, Thejovathi Murari, and Santoshachandra Rao Karanam	
<b>Medical Image Analysis Using Deep Learning Algorithm Convolutional Neural Networks .....</b>	245
D. Raghu and Hrudaya Kumar Tripathy	
<b>Data-Intensive Physics Analysis in Azure Cloud .....</b>	257
Igor Sfiligoi, Frank Würthwein, and Diego Davila	
<b>Identification of Assets in Industrial Control Systems Using Passive Scanning .....</b>	269
Aju Mathew Thomas, Mounesh Marali, and Lakshmikiran Reddy	

Contents	xiii
<b>Cybersecurity Governance in Information Technology: A Review of What Has Been Done, and What Is Next .....</b>	285
Yang Hoong and Davar Rezania	
<b>ML-Wasm Entropy and Plot: Dataframes and Plotting Powered by WebAssembly and Rust .....</b>	295
Dion Pinto, Arpit Bhat, Immanuel Gnanadurai, and Trupti Lotlikar	
<b>Multiclass Hierarchical Fuzzy Classification on Multi-labeled Data .....</b>	307
R. Kanagaraj, N. Rajkumar, K. Srinivasan, and E. Elakiya	
<b>An Efficient Approach for Identification of Multi-plant Disease Using Image Processing Technique .....</b>	317
K. Kranthi Kumar, Jyothi Goddu, P. Siva Prasad, A. Senthilrajan, and Likki Venkata Krishna Rao	
<b>Polling Cycle Analysis Using Different Modulation Types for IoT-Based Health Control in a Smart City .....</b>	327
Imane Benchaib, Salma Rattal, Kamal Ghoumid, and El Miloud Ar-Reyouchi	
<b>Wireless Sensor Network Lifetime Improving Based on Routing Protocols .....</b>	339
Bilal Saoud, Mounia Boucif, and Mourad Daas	
<b>Matching Forensic Composite Sketches with Digital Face Photos: A Bidirectional Local Binary Pattern-Based Approach .....</b>	349
H. T. Chethana and Trisiladevi C. Nagavi	
<b>Reduced Complexity of LDPC Codes using Hard Decision Decoder .....</b>	367
Allu Swamy Naidu, Appala Naidu Tentu, and Ajeet Singh	
<b>A Review on Video Sharing over Content Centric Networks .....</b>	383
C. Victoria Priscilla and A. R. Charulatha	
<b>An Overview of Augmenting AI Application in Healthcare .....</b>	397
Aarthy Chellasamy and Aishwarya Nagarathinam	
<b>Development of an Android Application to Carry Out Tourist Visits in Madrid as a Value-Added Service .....</b>	409
Antonio Sarasa-Cabezuelo	
<b>MFES Framework for Efficient Feature Selection Among Subsystems in Intelligent Building .....</b>	421
Abba Babakura, Abubakar Roko, Aminu Bui, Ibrahim Saidu, and Mahmud Ahmad Yusuf	

<b>QOS Management Protocol for Mobile Ad Hoc Networks Using Mobile Agents .....</b>	437
Mallikarjun B.C. and H. S. Phalanetra	
<b>Design of IoT Platform for Monitoring and Control of Variables of Industrial Processes .....</b>	451
Hernando González, Azarquiel Diaz, Luis Jaimes, and Carlos Meza	
<b>A Study on Identification of Plant Diseases Using Image Processing .....</b>	463
Disha Sushant Wankhede, Amit Gamot, Kashish Motwani, Shaunak Kayande, Vidhi Agrawal, and Chetan Chinchulkar	
<b>A Robust and Accurate IoT-Based Fire Alarm System for Residential Use .....</b>	479
Tanjil Hossain, Md. Ariful Islam, Alif Bin Rahman Khan, and Md. Sadekur Rahman	
<b>Kernel Feature Variant-Based Gaussian Process Regression for Prediction of Snail Rings .....</b>	493
M. Shyamala Devi, N. Abhishek Rao, S. G. Kushal Kumar, G. Dheeraj, and K. Govinda	
<b>Mutual Information Score-Based Clustering for Evaluation of Image Dominant Color .....</b>	505
M. Shyamala Devi, N. K. Manikandan, D. Manivannan, Y. Lakshmi Akshitha, G. Chandana, K. Lasya Priya, and G. Vijayalakshmi	
<b>Hybrid Deep Learning-Based Music Recommendation System .....</b>	517
M. Sunitha, T. Adilakshmi, and Mehar Unissa	
<b>CNN-Based Deep Learning Network for Human Activity Recognition During Physical Exercise from Accelerometer and Photoplethysmographic Sensors .....</b>	531
Sakorn Mekruksavanich and Anuchit Jitpattanakul	
<b>A Review Towards Research in Multi-robot Coordination System .....</b>	543
M. Pavithra and T. Kavitha	
<b>Feature Extraction and Representation Learning via Deep Neural Network .....</b>	551
T. Anuradha, Arun Tigadi, M. Ravikumar, Paparao Nalajala, S. Hemavathi, and Manoranjan Dash	
<b>Drowsiness Detection Using Facial Features, Image Processing and Machine Learning .....</b>	565
S. Nandhini, Vaishnavi Venkatasubramanian, and C. Aparna	

Contents	xv
<b>Linear Separability as a Condition for Solving Multiple Problems by a Single Threshold Neuron .....</b>	<b>575</b>
Kostadin Yотов, Emil Hadzhikolev, and Stanka Hadzhikoleva	
<b>Chemicals Informatics: Search Structural Factors and Optimal Composites .....</b>	<b>593</b>
Takashi Isobe and Yoshihiro Okada	
<b>Analysis of Temperature Impacts on Material-Dependent Thermoelastic Damping in Simply Supported Rectangular Microplate Resonators Applying Size Effects .....</b>	<b>609</b>
R. Resmi, V. Suresh Babu, and M. R. Baiju	
<b>Template Protection in Multimodal Biometric System Using Watermarking Approach .....</b>	<b>617</b>
C. Vensila and A. Boyed Wesley	
<b>A Big Data Deep Learning Approach for Credit Card Fraud Detection .....</b>	<b>633</b>
Kandasamy Illanko, Raha Soleymanzadeh, and Xavier Fernando	
<b>A Study of VLC Between Vehicles and Traffic Signal Lights .....</b>	<b>643</b>
Jonathan Diller and Xavier Fernando	
<b>A Comprehensive Study of Various DC Faults and Detection Methods in Photovoltaic System .....</b>	<b>657</b>
Alaa Hamza Omran, Dalila Mat Said, Siti Maherah Hussin, and Sadiq H. Abdulhussain	
<b>Forecasting and Seasonal Analysis of Air Quality Index using Machine Learning Models during COVID-19 Pandemic .....</b>	<b>677</b>
Priyanka Harjule, Basant Agarwal, Ashish Burdak, Satvik Gupta, Saurav Singh, and Shivdeep Singh	
<b>Accurate Segmentation of Lung Nodule using Adaptive Weights as Feature for Recurrent Neural Network .....</b>	<b>699</b>
R. Janefer Beula and A. Boyed Wesley	
<b>Review on Segmentation of Facial Bone Surface from Craniofacial CT Images .....</b>	<b>717</b>
Jithy Varghese and J. S. Saleema	
<b>Development of a Fully Convolutional Network for the Segmentation of Adipose Tissues on Abdominal MRI .....</b>	<b>739</b>
B. Sudha Devi and D. S. Misbha	

<b>Improvement of Clarity for Foggy/Dusty Weather Images Using Triple Threshold Method .....</b>	<b>753</b>
Minu Inba Shanthini Watson Benjamin, N. S. Kalyan Chakravthy, Baddeti Syam, R. Navaneethakrishnan, Jee Joe Michael, and J. N. Swaminathan	
<b>The Olympic Gold Medalists on Instagram: A Data Mining Approach to Study User Characteristics .....</b>	<b>761</b>
Amirhosein Bodaghi and Jonathan J. H. Zhu	
<b>Performance Analysis of KNN Algorithm to Improve the Process of Hemodialysis on Post-Covid-19 Patients .....</b>	<b>775</b>
N. Vijaya, G. Revathy, D. Sivanandakumar, C. Sasikala, and B. Sreedevi	
<b>Secure Mobile Internet Banking System Using QR Code and Biometric Authentication .....</b>	<b>791</b>
S. Ajish and K. S. Anil Kumar	
<b>A Sparrow Search Algorithm for Detecting the Cross-layer Packet Drop Attack in Mobile Ad Hoc Network (MANET) Environment .....</b>	<b>809</b>
S. Venkatasubramanian, A. Suhasini, and N. Lakshmi Kanthan	
<b>Blockchain-Based Internet of Vehicles for Intelligent Transportation System Using Fog Computing .....</b>	<b>827</b>
U. Sakthi, J. Dafni Rose, Dahlia Sam, and M. K. Kirubakaran	
<b>Smart, Safe, and Secure Shopping Experience Using Beacons .....</b>	<b>837</b>
J. K. Lakshmi Divya, R. Iswarya, and V. S. Felix Enigo	
<b>Android Game for Amblyopia Treatment: A Prospective Study .....</b>	<b>853</b>
Sarah AlGhamdi, Sadiqa Alghawas, and Nazeeruddin Mohammad	
<b>A Hybrid Intrusion Detection Approach Based on Deep Learning Techniques .....</b>	<b>863</b>
Diego F. Rueda, Juan C. Caviedes, and Wilmar Yesid Campo Muñoz	
<b>Develop a Smart Data Warehouse for Auto Spare Parts Autonomous Dispensing and Rack Restoration by Using IoT with DDS Protocol .....</b>	<b>879</b>
R. Shiva Shankar, Ravibabu Devareddi, Gadiraju Mahesh, and V. MNSSVKR Gupta	
<b>Link Prediction in Paper Citation Network based on Deep Graph Convolutional Neural Network .....</b>	<b>897</b>
Bui Thanh Hung	

Contents	xvii
<b>Traffic Event Reporting Framework Using Mobile Crowdsourcing and Blockchain .....</b>	909
Abin Oommen Philip, RA. K. Saravanaguru, and P. A. Abhay	
<b>A Flexible Protocol for a Robust Hospitals Network Based on IoT .....</b>	931
Salma Rattal, Kamal Ghoumid, and El Miloud Ar-Reyouchi	
<b>Toward Data Visualization and Data Forecasting with COVID-19 Vaccination Statistics .....</b>	945
Vaishnavi Kulkarni, Jay Kulkarni, and Anurag Kolhe	
<b>Network Physical Layout-Based Reliable Routing in Vehicular Ad Hoc Networks .....</b>	961
S. Padmakala, A. Akilandeswari, G. Gugapriya, and Himanshu Shekhar	
<b>A Hybrid Split and Merge (HSM) Technique for Rapid Video Compression in Cloud Environment .....</b>	969
R. Hannah Lalitha, D. Weslin, D. Abisha, and V. R. Prakash	
<b>A New Intrusion Detection and Prevention System Using a Hybrid Deep Neural Network in Cloud Environment .....</b>	981
Subalakshmi Mani, Bose Sundan, Anitha Thangasamy, and Logeswari Govindaraj	
<b>Smart Energy Metre Based on IoT .....</b>	995
K. Rubitha, J. Jecintha, K. Shifana Begum, and S. Suriya	
<b>Author Index .....</b>	1007

## About the Editors

**A. Pasumpon Pandian** received his Ph.D. degree in the Faculty of Information and Communication Engineering under Anna University, Chennai, TN, India, in 2013. He received his graduation and postgraduate degree in Computer Science and Engineering from PSG College of Technology, Coimbatore, TN, India, in the year 1993 and 2006, respectively. He is currently working as Dean (R&D) in CARE College of Engineering, Trichy, TN, India. He has 26 years of experience in teaching, research and IT industry. He has published more than 20 research articles in refereed journals. He acted as a conference chair in IEEE and Springer conferences and guest editor in *Computers and Electrical Engineering* (Elsevier), *Soft Computing* (Springer) and *International Journal of Intelligent Enterprise* (Inderscience) journals. His research interests include image processing and coding, image fusion, soft computing and swarm intelligence.

**Xavier Fernando** is Professor at the Department of Electrical and Computer Engineering, Ryerson University, Toronto, Canada. He has (co-)authored over 200 research articles, two books (one translated to Mandarin) and holds few patents and non-disclosure agreements. He was IEEE Communications Society Distinguished Lecturer and delivered close over 50 invited talks and keynote presentations all over the world. He was a member in the IEEE Communications Society (COMSOC) Education Board Working Group on Wireless Communications. He was Chair IEEE Canada Humanitarian Initiatives Committee 2017–2018. He was also Chair of the IEEE Toronto Section and IEEE Canada Central Area. He is a program evaluator for ABET (USA). He was a visiting scholar at the Institute of Advanced Telecommunications (IAT), UK, in 2008 and MAPNET fellow visiting Aston University, UK, in 2014. Ryerson University nominated him for the Top 25 Canadian Immigrants Award in 2012 in which was a finalist. His research interests are in signal processing for optical/wireless communication systems. He mainly focuses on physical and MAC layer issues. He has special interest in underground communications systems, of cognitive radio systems, visible light communications and wireless positioning systems.

**Dr. Wang Haoxiang** is currently the director and lead executive faculty member of Go Perception Laboratory, NY, USA. His research interests include multimedia information processing, pattern recognition and machine learning, remote sensing image processing and data-driven business intelligence. He has co-authored over 60 journal and conference papers in these fields on journals such as Springer *MTAP*, *Cluster Computing*, *SIVP*; *IEEE TII*, *Communications Magazine*; Elsevier *Computers & Electrical Engineering*, *Computers, Environment and Urban Systems*, *Optik*, *Sustainable Computing: Informatics and Systems*, *Journal of Computational Science*, *Pattern Recognition Letters*, *Information Sciences*, *Computers in Industry*, *Future Generation Computer Systems*; Taylor & Francis *International Journal of Computers and Applications* and conference such as IEEE SMC, ICPR, ICTAI, ICICI, CCIS, ICACI.

# Rakshak—Prototype Application to Depict Crime Hotspot and Safest Path Between Locations



Archana Naik, Aakanksha Gaur, Stuti Srivastava, Shashikant Saini,  
R. Gayathri, and Kavitha Sooda

**Abstract** With every passing day, there is a rise in crime that adversely affects the growth of society. The density and intensity of crimes vary with the region in any city. Technological advancement can be made useful to make people aware of dangerous locations. In this work, we have built a model to visualize the crime-prone regions between the two locations provided by the user and plot the safe route between them. An algorithm to identify the crime hotspots and the safe route is proposed. The real-world datasets of Toronto in Canada and Indore in India are considered. The proposed model comprises data collection of the crime of these cities, data analysis to delineate the crime hotspots in the region, and the safest route between places. Visualizing the outcomes using a heatmap makes people aware of risky locations and further helps in reducing the crime count of the city.

**Keywords** Crime data analytics · Crime reporting · Hotspot · K-means clustering · Safest route guide

---

A. Naik (✉) · A. Gaur · S. Srivastava · S. Saini · R. Gayathri

Department of Computer Science and Engineering, Nitte Meenakshi Institute of Technology, Yelahanka, Bangalore 560064, India  
e-mail: [archana.naik@nmit.ac.in](mailto:archana.naik@nmit.ac.in)

A. Gaur  
e-mail: [1nt17cs002.aakanksha@nmit.ac.in](mailto:1nt17cs002.aakanksha@nmit.ac.in)

S. Srivastava  
e-mail: [1nt17cs189.stuti@nmit.ac.in](mailto:1nt17cs189.stuti@nmit.ac.in)

S. Saini  
e-mail: [1nt17cs174.shashikant@nmit.ac.in](mailto:1nt17cs174.shashikant@nmit.ac.in)

R. Gayathri  
e-mail: [1nt17cs060.gayathri@nmit.ac.in](mailto:1nt17cs060.gayathri@nmit.ac.in)

K. Sooda  
Department of Computer Science and Engineering, B.M.S. College of Engineering, Basavanagudi, Bangalore 560019, India  
e-mail: [kavithas.cse@bmscse.ac.in](mailto:kavithas.cse@bmscse.ac.in)

## 1 Introduction

As the population rises, criminal activities escalate, and the risk factor for the safety of citizens increases. Crime adversely affects society's growth and the quality of life of citizens. Criminal activities are dispersed over space [1]. Local criminal activities intend to cause harm, loss of property, and unfair gain in wealth or resources through corrupt or illegal means [2]. Reports of criminal activity are on the rise every day, and a large amount of crime-related data is being generated. To control the conditions, it is necessary to obtain the relevant and timely data [3]. With the help of various data mining techniques, this data can be analyzed using the records stored by the government of any country and can be used to predict the possibility of crime occurring.

Data mining is the process of analyzing large volumes of data to extract and analyze the various patterns hidden in the data. By analyzing the data, the crime hotspot can be found with the help of data mining techniques. If any region is a crime hotspot, it will be alerted. By identifying the frequent patterns between crime and location, it can be found which location has a high rate of crime.

In this paper, an approach to identify the crime-prone regions in a city is presented. The application has been given the name "Rakshak" which means protector as it protects the user from dangerous locations. Rakshak is a mobile application that helps the user locate the crime hotspots in a city and can find the safest route between two locations that the user enters. The various coordinates of the city are clustered using K-means clustering. From these clusters, the average crime in each region of the city is calculated, which helps in calculating a threshold value to identify the crime hotspots. The regions having a crime rate higher than the threshold are considered as hotspots and are visualized on Google Maps using the Google Maps API. Each region is given a safety index based on the number, type, and severity of the crime. For the calculation of the safety index, a risk factor is assigned to each type of crime. Then, to calculate the safest path, all the possible routes between the two locations entered by the user are extracted and analyzed. The average safety index of each route is calculated, and the route with the lowest safety index is given as the safest route. The coordinates of this route are sent to the application which then visualizes the route. Coming to the user interface, the application is made using Flutter, which is an open-source UI development software. It is divided into two parts. One being the portal that shows the crime hotspots of a given city and the safest route between two locations. The other is the crime reporting portal that allows the user to enter crime type, crime time, and location, which will keep the dataset updated for the analysis.

The organization of the rest of the paper is as follows: Section 2 discusses about the literature reviewed for crime analysis and crime mapping techniques. Section 3 illustrates the methods and functions used to depict the hotspot locations in the city and safest path between two locations. Section 4 is about the result followed by conclusion.

## 2 Related Work

In today's world, increasing crime rates have become a growing concern for government officials. Therefore, it has become a necessity to use technology in this area so that it will be helpful for officials to find culprits. There has been innumerable work done in this field for the purpose of reducing crime. Different types of datasets have been analyzed and meaningful data has been extracted. Extracted data consists of crime type, location, date, day, and time. Researchers have been working on various techniques and methods to find the best way to prevent criminal activity.

Authors Tahani et al. [4] presented their work on crime prediction based on crime type. Their objective was to find spatial and temporal criminal hotspots and predict the type of crime that might occur next at a specific location. An Apriori algorithm is used to find the criminal hotspots in the areas. Naïve Bayesian classifier and decision tree are used to predict an expected crime. Because the decision tree results revealed complex patterns, the Naïve Bayesian classifier was chosen as the preferred methodology. Mohammed et al. [5] proposed an approach in which with the help of crime attributes, DBSCAN technique is used to detect the patterns of crime in a particular city. A geographic information system (GIS) was used for the temporal analysis. Aarthi et al. [6] proposed a system to extract data from crime data records and perform clustering. They used the K-means algorithm for clustering of the data. For the live data, a streaming algorithm was used. Affinity propagation and radial basis (RBF) network was to select the data required for the clustering. Deepika et al. [7] proposed an approach to detect crime in India using data mining techniques for which they used crime data from each state. They have used the K-means algorithm for clustering and the random forest algorithm and neural networks for classification. The results were visualized on Google Maps. Sonawane et al. [8] proposed a crime analysis tool to predict crime that uses the K-means algorithm for data classification. They grouped the crimes using K-means. The crimes are correlated using Pearson's correlation coefficient. The prediction was done using simple linear regression. The authors depicted the correlation between various locations and crimes.

Gera et al. [9] provide information about crime in Delhi. Crime profiling was done in three different tiers—based on crime types where the crimes were divided into three categories, based on the type of area, and based on both the type of crime and type of area. For each analysis, the results are formulated in a table that tells us the percentage of crime. Benjamin Fredrick David et al. [10] did a survey on the data mining techniques used to predict and analyze crime. They segregated the techniques into text, content, and natural language processing-based methods which included decision tree classifier to detect suspicious emails about criminal activities, clustering and classification to identify the patterns in crime, spatial and geolocation-based methods, prisoner-based methods, and communication-based methods. Feng et al. [11] utilized big data analytics and visualization techniques for their exploratory analysis to identify crime patterns and trends in three cities of the USA—San Francisco, Chicago, and Philadelphia. They combined and compared different algorithms for the analysis. Patil et al. [12] proposed a criminal prediction model to predict crime.

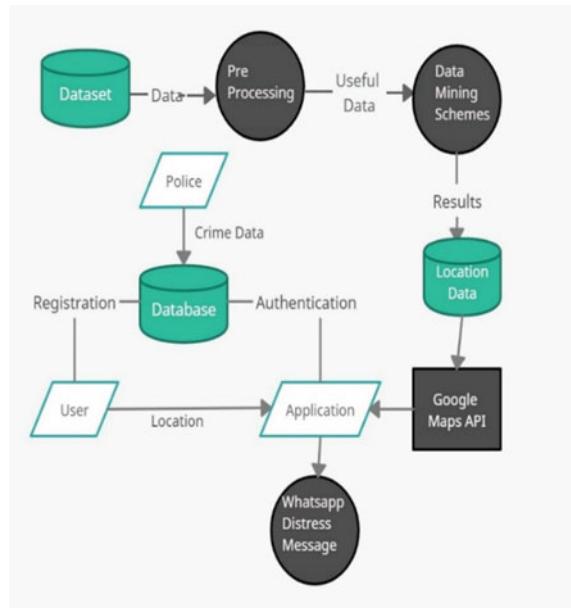
A real-life authentic dataset was implemented for the analysis. K-means clustering technique was used to cluster the data according to high and low values; this clustered data was given as input to the Apriori algorithm. The classification was done using the Naïve Bayes algorithm. The goal of the paper was to develop a system that builds risk surfaces with keeping time efficiency in mind. Ramasubbareddy et al. [13] implemented a web application to predict crime. Their application identified the frequent patterns in the crime data using the Apriori algorithm. They used a decision tree classifier and Naïve Bayes as searching algorithms by giving locations as input to identify the crime that had a high possibility to occur.

Tayal et al. [14] proposed an approach to detect the crime and identify the criminals. This was done using K-means clustering to cluster the crimes, and based on the magnitude of crime, these clusters were plotted on Google Maps. They used K-nearest neighbor for the classification to identify the criminals. Yadav et al. [15] discussed crime detection and prediction using supervised, semi-supervised, and unsupervised techniques. They created a regression model, and using this, the crime rate for different states is predicted. The approach focused on suppressing the crime rate and helping the local police. Kumar et al. [16] have proposed a method using KNN machine learning algorithm to predict crime in a region. For feature selection, the extra tree classifier is used. The Seaborn heatmap visualizes the variation of crime in a month. Thota et al. [17] have analyzed risk zones in India with respect to female and male crime count in the states through K-means clustering using WEKA software. Reddy et al. [18] have proposed a model that predicts crime at a particular location using K-nearest neighbor and the Naïve Bayes algorithm. A dataset of UK was taken for analysis. Visualization of crime hotspots is done on Google Maps using Ggplot, Ggmap, and GoogleVis.

Different techniques and algorithms have been applied for the analysis of crime dataset of various cities. Prediction of the crime type is one of the prime concerns in today's world to make people aware of the criminal activity that can occur at a location. Widely used algorithms for crime prediction are Apriori, K-nearest neighbors, long short-term memory (LSTM), Naïve Bayes classifier, and decision tree. Before prediction, clustering of crime data is crucial for pinpointing the crime hotspots of the region and is done using the K-means algorithm.

### 3 Methodology

This section outlines the workflow and methodology of our proposed approach. The block diagram shown in Fig. 1 gives a detailed description of the system design that has been used, mentioning the different processes implemented in the system. The first step is data preprocessing. Null values from the dataset are removed, and the features necessary for the analysis are selected. Data mining techniques and algorithms are applied to the preprocessed dataset. From the results of the analysis, the location data is extracted and visualized in the application using the Google Maps API.

**Fig. 1** Data flow diagram

In order to use the application, the user must first register by providing information such as an email address, an emergency phone number, and a password. The data provided is stored in the database. In case of any danger, the user can send an emergency message to the emergency number provided. The user can also report any crime by clicking on the “record crime” option and submitting the crime details. These details get stored in the database.

### **3.1 Data Collection and Preprocessing**

Two real-world datasets have been taken for the analysis. The datasets of cities that have been chosen are Toronto, Canada, and Indore, India. Both the dataset used are taken from Kaggle that is an open-source platform. It provides numerous datasets for users to build models that solves real-world problems and also allow them to publish their own datasets. Initially, analysis was done on Toronto’s dataset, and using the same functions and formulas, it has been implemented on an Indian city.

#### **3.1.1 Toronto Crime Dataset**

Toronto city crime dataset [19], initially, has 29 columns and 206,376 rows. Cleaning of dataset is done, and null values from it have been removed. After cleaning, dataset contained 173,683 records of crimes at each region in the city from the year 2014 to

2019. Each record contains the features of the crime such as the type of crime (the crimes are divided into five categories based on major crime indicator (MCI)—auto theft, assault, break and enter, robbery, and theft over), neighborhood (locality where crime occurred), latitude and longitude of the place where crime occurred, the time and date when the crime occurred, and reporting time and date of crime and premise type (crime occurred in indoor place or outdoor place). Out of the 29 attributes in the original dataset, 15 attributes such as the type of crime (MCI), neighborhood, time of occurrence, date of occurrence, latitude and longitude, and hood\_ID were selected for the analysis.

### 3.1.2 Indore Crime Dataset

Indore city dataset [20] has 2139 rows and 9 columns; all the columns and rows have been taken into consideration while doing the analysis. Dataset has different types of crime in the form of acts (act379—theft, act302—murder), latitude and longitude of place of crime and the time, and date when the crime occurred. For identification of safest path between two locations, two columns named neighborhood and hood\_ID have been added to this dataset which makes it exactly same as that of Toronto city. This helps in applying the algorithms created for the analysis on the dataset easier.

## 3.2 Clustering and Hotspot Detection

Clustering is done on the latitude and longitude features to cluster the data points based on location. A small data frame containing only the latitude and longitude columns of the refined dataset is created, and the clustering algorithm is applied to it. The clustering of the dataset is done using K-means clustering algorithm to create the clusters, and these clusters are utilized to identify the hotspots. K-means for clustering technique is the best algorithm for working with large datasets, and it also provides the flexibility to choose the number of clusters. The value of K is taken as 35 to create the clusters, and then, K-means is applied to the new data frame created.

After assigning clusters to each data point, these clusters are fed to an algorithm that calculates the average crime count of each cluster. This is done by calculating the total number of criminal cases for each cluster and dividing it by the number of data points in that cluster as shown in Eq. (1).

$$\text{Average Crime Count} = \frac{\text{Sum of magnitude of crime in the cluster}}{\text{Number of locations in the cluster}} \quad (1)$$

Then the mean of the average crime counts of each cluster is calculated. This mean is the threshold value that helps in determining the crime hotspots in the city. After this, Algorithm 1 is used to compare the number of crimes at each location to

**Table 1** Representation of safety level at location

Color assigned	Danger level	Magnitude of crime
Green	Safe	Less than threshold/2
Yellow	Moderately Safe	Between threshold/2 and threshold
Red	Dangerous	Greater than threshold

the threshold. Using the threshold, the locations are divided into three zones—safe, moderately safe, and dangerous as shown in Table 1. If the number of crimes at a location is greater than the threshold, that location is considered as a crime hotspot and is visualized in Google Maps. The visualization is done by creating a heatmap over the Google Maps interface using the locations—latitude and longitude and the magnitude of crime in that location.

---

**Algorithm 1** Hotspot Detection Algorithm

---

```

1: procedure HOTSPOTDETECTION(data)
2:   for  $i = 0, 1, \dots, \text{len}(\text{data})$  do
3:     count  $\leftarrow \text{data}[i][2]$ 
4:     if  $\text{count} < \text{threshold}/2$  then
5:       safe_location.append(data[i])
6:     else if  $\text{count} \geq \text{threshold}/2$  and  $\text{count} < \text{threshold}$  then
7:       moderately_safe_location.append(data[i])
8:     else if  $\text{count} \geq \text{threshold}$  then
9:       dangerous_location.append(data[i])
10:    end if
11:  end for
12:  return dangerous_location
13: end procedure

```

---

The hotspot detection Algorithm is used to detect the crime hotspots, that is, the locations with high crime rates with the help of a threshold value. The locations are classified into three categories—safe, moderately safe, and dangerous. The locations classified as dangerous are the crime hotspots.

Step 1: Iterate through the location data which contains latitude, longitude, and the magnitude of crime in that location.

Step 2: For each latitude and longitude, store its respective count of the number of crimes in a variable “count” and compare its value to the threshold value.

Step 3: If the magnitude of crime in a location is less than threshold/2, that location is classified as “safe” and is stored in a list called “safe\_location.” If the magnitude of crime in that location lies between threshold/2 and threshold values, then that location is classified as “moderately safe” and is added to a list named “moderately\_safe\_location.” Else, if the location has a magnitude of crime

above the threshold, it is classified as “dangerous” and is stored in a list named “dangerous\_location.”

Step 4: Go back to step 2 and continue with all the location data.

Step 5: The function returns the list “dangerous\_location” which contains all the locations whose magnitude of the crime is above the threshold. This list is then used for visualizing the crime hotspots.

### 3.3 Safest Path Identification

The objective is to determine the safest route between two regions in the city. For the identification of the safest route, a factor called the safety index is introduced. The safety index is calculated with the help of magnitude and severity of crimes in each location.

As the user enters the “origin” and the “destination,” all the possible routes are extracted, and their average safety index is calculated. The route with the least safety index is returned as the safest route. Algorithm 2 is used to identify the safest path. The input given to the algorithm is a list P which contains all the possible routes extracted from the origin and destination and the safety index of each location. The route with the least average safety index is returned as the safest route.

---

#### Algorithm 2 Safest Path Algorithm

---

```

1: procedure SAFESTPATH(P,safety_index)
2:   for i = 0,1,...len(P) do
3:     for j = 0,1,...,len(P[i]) do
4:       route_safetyindex ← 0
5:       index ← indexof(P[i][j])
6:       route_safetyindex ← route_safetyindex + safety_index[index-1]
7:     end for
8:     route_safetyindex ← route_safetyindex/len(P[i])
9:     avg_safety_index.append(route_safetyindex)
10:    end for
11:    min ← avg_safety_index[0]
12:    min_index ← 0
13:    for i = 1,2,...len(avg_safety_index) do
14:      if avg_safety_index[i] < min then
15:        min ← avg_safety_index[i]
16:        min_index ← i
17:      end if
18:    end for
19:    return P[min_index]
20: end procedure

```

---

### ***3.4 Google Maps Implementation***

For the implementation and visualization on Google Maps, a new Google Cloud Platform (GCP) project is created and set up. Google Maps SDK and Directions API are added to the project. An API key is created, and the Flutter project is set up to use Google Maps using the API key. The hotspots are visualized by drawing the heatmap on Google Maps. Once the user enters the two locations, which are the origin and destination, Google Maps returns all the routes between those two locations. These routes are analyzed, and the safest route is drawn on the map using the polyline algorithm.

### ***3.5 Crime Reporting***

New crime records can also be added through the application. The addition of new crime data can be done by anyone using the application. Here, they will be asked to enter the type of crime, the time of the occurrence, and the location of the crime. This data gets stored in the firebase.

## **4 Results and Discussion**

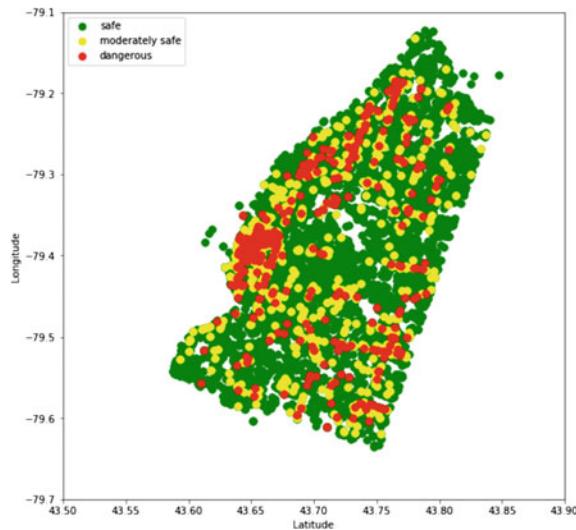
The application contains user registration, crime reporting, and a menu on the top-left in the form of a left arrow icon, which contains four options—heatmap, safe path, record crime, and log out. The user interacts with the application using a map interface. Once the user registers and logs in to the application, the Google Maps interface is visible. The hotspot detection is based on the points (latitude and longitude) that are present in the dataset, and the crime count for these points is calculated.

Each location has been assigned a safety index based on the magnitude and severity of each crime committed at that location. This helps in identifying the safest route between the locations entered by the user.

The subsection provides the results of evaluation on the Toronto and Indore city datasets.

### ***4.1 Toronto City***

Figure 2 shows the plotting of different locations (latitude and longitude) on the basis of the magnitude of crime at each location. The locations are divided into three levels:

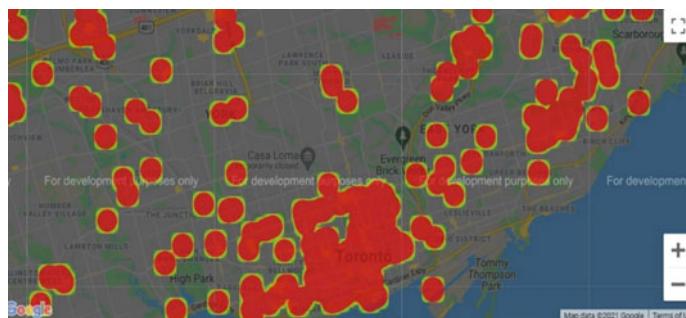


**Fig. 2** Marking of location coordinates based on crime intensity

safe, moderately safe, and dangerous on the basis of threshold values as shown in Table 1.

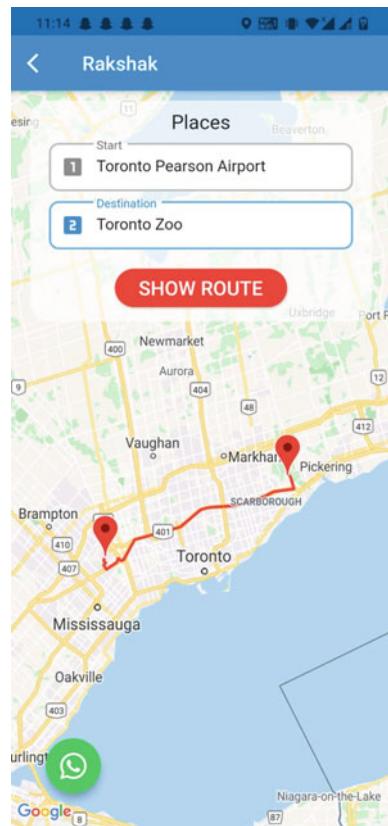
Figure 3 depicts the locations that have a magnitude of crime higher than the threshold visualized on the Google Maps as the crime hotspots. The user is able to locate these hotspots through the application. Once the user enters origin and destination, the safest path between the locations is visualized.

Figure 4 depicts the safest path between the locations selected as “Toronto Pearson Airport” and “Toronto Zoo.”



**Fig. 3** Heatmap of major crime hotspots in Toronto

**Fig. 4** Safest path between Toronto Pearson Airport and Toronto Zoo



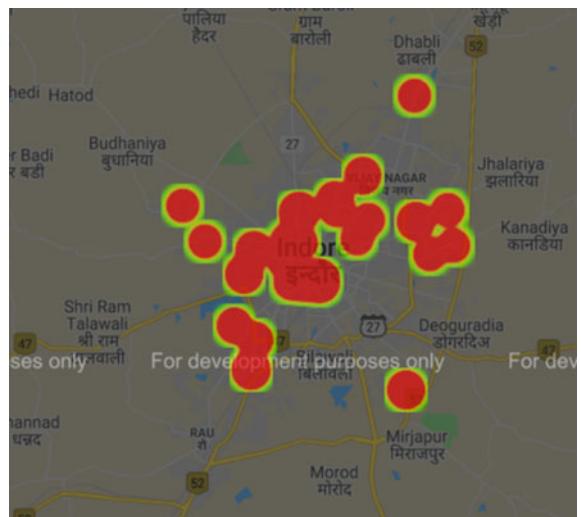
## 4.2 Indore City

The analysis of hotspot detection and safest path identification was also done on Indore crime data. Due to the small size of the dataset, only a few locations have been identified as crime hotspots. Figure 5 depicts the hotspots identified in Indore. It can be seen that locations such as Vijay Nagar and Dhabli have been identified as crime hotspots. This helps the user be aware while visiting these locations. Figure 6 shows the safest path between the locations “Rani Bagh Main” and “Bhanvarkuan.”

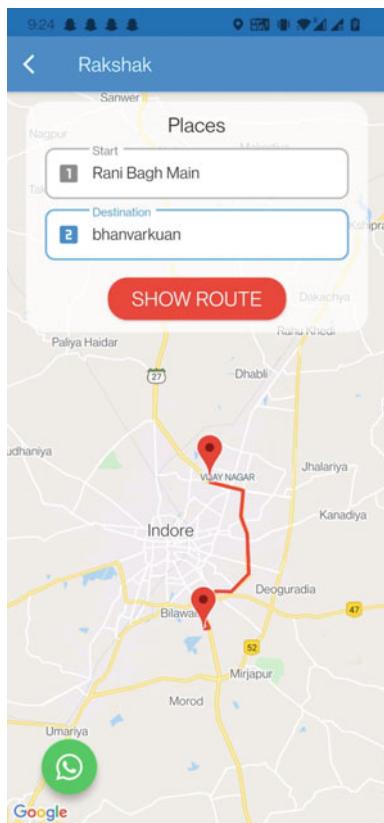
## 4.3 Crime Reporting

When the user clicks on the record crime option from the menu, it takes them to the crime reporting page. Here, the user is asked to enter the type of crime that occurred,

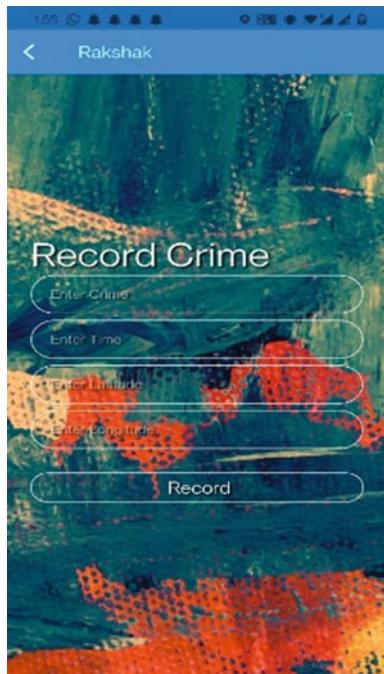
**Fig. 5** Heatmap of major crime hotspots in Indore



**Fig. 6** Safest path between Rani Bagh Main and Bhanvarkuan



**Fig. 7** Crime reporting window



time of occurrence, and the location as shown in Fig. 7. On clicking the “record” button, this data gets stored in the database. The data can also be added by the police officials.

## 5 Conclusion

Crime rates have been increasing at a faster rate in the last few years. Crimes are now not only committed in sparsely populated areas but also in areas that have high population density. They occur during the day as well as the night. Hence, it is important for people to be aware of the areas with high crime concentrations and take necessary precautions.

In this paper, we proposed and developed a model to identify the crime-prone regions and provide a safe route to avoid the crime-prone locations. For the purpose demonstration, the dataset from Toronto, Canada, and Indore, India is used. Algorithm to identify the crime hotspots within the city has been developed considering the crime dataset. The result of this is represented as a heatmap. An algorithm to find the safest route between the selected source and destination has been developed. The work presents the visualization of crime rates and the safest route between two locations entered by the user. It is an attempt to help users identify the safest path and the hotspots for crime in the city. The K-means algorithm is used for the clustering of

the datasets. The magnitude of the crime is identified in each cluster, and a threshold value is calculated. These threshold values are used to determine the hotspots. For identifying the safest path, a safety index is calculated, and the path with minimum safety index is returned. Furthermore, the application also has a facility to add a crime.

The further work can be done to enhance the crime reporting feature by including authentication of the data that will be updated by the authorized personnel, who are responsible for maintaining the law and order. This approach is useful to identify the crime hotspots in a city and safest path between two location which makes the user aware of the locations and path which are unsafe to travel.

## References

1. Wang, D., Ding, W., Lo, H., Stepinski, T., Salazar, J., Morabito, M.: Crime Hotspot Mapping Using the Crime Related Factors—A Spatial Data Mining Approach. Springer, New York (2012)
2. Kasamani, B.S., Alaka, B., Chepnetich, M.: A real-time location based prototype for notification of crime hotspots using crowdsourcing. In: 2020 IST-Africa (2020). ISBN: 978-1-905824-65-6
3. Kaur, B., Ahuja, L., Kumar, V.: Crime against women: analysis and prediction using data mining techniques. In: Com-IT-Con, India (2019)
4. Almanie, T., Mirza, R., Lor, E.: Crime prediction based on crime types and using spatial and temporal criminal hotspots. In: IJDKP (2015)
5. Mohammed, A.F., Bailee, W.R.: The GIS based criminal hotspot analysis using DBSCAN technique. In: ISCAU (2020)
6. Aarthi, S., Samyuktha, M., Sahana, M.: Crime hotspot detection with clustering algorithm using data mining. In: ICOEI (2019). ISBN: 978-1-5386-9439-8
7. Deepika, K.K., Vinod, S.: Crime analysis in India using data mining techniques. In: IJET, pp. 253–258 (2018)
8. Sonawane, T., Shaikh, S., Shaikh, S., Shinde, R., Sayyad, A.: Crime pattern analysis, visualization and prediction using data mining. In: IJARIIE (2015). ISSN(O) 2395-4396
9. Gera, P., Vohra, R.: City crime profiling using cluster analysis. In: IJCSIT (2014). 0975-9646
10. Benjamin Fredrick David, H., Suruliandi, A.: Survey on crime analysis and prediction using data mining techniques. ICTACT J. Soft Comput. (2017)
11. Feng, M., Zheng, J., Ren, J., Hussain, A., Li, X., Xi, Y., Liu, Q.: Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data. IEEE (2019)
12. Patil, R., Kacchi, M., Gavali, P., Pimpuria, K.: Crime pattern detection, analysis and prediction using machine learning. In: IRJET (2020). ISSN: 2395-0072
13. RamasubbaReddy, S., Aditya Sai Srinivas, T., Govinda, K., Manivannan, S.S.: Crime Prediction System. Springer (2020)
14. Tayal, D.K., Jain, A., Arora, S., Agarwal, S., Gupta, T., Tyagi, N.: Crime Detection and Criminal Identification in India Using Data Mining Techniques. Springer-Verlag, London (2014)
15. Yadav, S., Timbedia, M., Vishwakarma, R., Yadav, N.: Crime pattern detection, analysis and prediction. Presented at the International Conference on Electronics, Communication and Aerospace Technology (2018)
16. Kumar, A., Sukhdev, Y., Verma, A., Shinde, G., Lal, N.: Crime prediction using K-nearest neighboring algorithm. Presented at the International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) (2020)

17. Thota, L.S., Alalyan, M., Khalid, A.-O.A., Fathima, F., Changalasetty, S.B., Shiblee, M.: Cluster based zoning of crime info. In: ICACC (2017)
18. Reddy, H.K., Saini, B., Mahajan, G.: Crime prediction & monitoring framework based on spatial analysis. In: ICCIDS (2018)
19. Pastor, K.: Toronto Police Data (Version 3). Kaggle. Available: <https://www.kaggle.com/kapastor/toronto-police-data-crime-rates-by-neighbourhood> (2020)
20. Raut, Y.: Indore Police Crime Dataset (Version 1). Kaggle. Available: <https://www.kaggle.com/yashraut/indore-police-crime-datasetia> (2020)

# Blockchain Technology in Application Development and Associated Challenges



Vishal Polara, Pooja Bhatt, Dharmesh Patel, and Ketan Rathod

**Abstract** Blockchain, the word itself, suggests a block of chain most useful for secure peer-to-peer transactions. It is most popular and in current trends to develop various security-related applications. Blockchain technology is developed using various advance technique of cryptography and economic models. It is also a public ledger of all cryptographic transaction done online. Blockchain network also provides various types of implementations like permission less or permission blockchain network. It also provides facility of making smart contract to collect information digitally using cryptographic function. It also provides facility to create framework which is backward application compatible. Applications developed using blockchain technology are tamper-resistant and evident of digital ledgers implemented in a distributed manner without having a central administrator that makes it more reliable. In this manuscript, introduction of the working of blockchain technology for application development and various advantages and challenges of blockchain technology is discussed.

**Keywords** Algorithm · Cryptography · Decentralization · Hard fork · Ledger · Security · Smart contract · Soft fork · Symmetric

---

V. Polara (✉) · D. Patel  
BVM Engineering College, Anand, Gujarat, India  
e-mail: [vishalpolara@gmail.com](mailto:vishalpolara@gmail.com)

P. Bhatt  
MBIT Engineering College, Anand, Gujarat, India

K. Rathod  
Babariya Institute of Engineering and Technology, Vadodara, Gujarat, India

## 1 Introduction

### 1.1 Overview of Blockchain Technology

It was introduced by Satoshi Nakamoto in 2008. This idea was combined with other computing technologies to create recent cryptocurrencies. It is necessary to implement a mechanism without having a central authority to protect electronic cash. Bitcoin is the first blockchain-based cryptocurrency network.

It protects digital transaction as it is implemented without using central repository or database and without any central administrator (e.g., government, bank, or organization). Initially, all users can do transactions, and the transaction cannot be changed in the blockchain network once published.

It has made significant growth in the field of decentralized autonomous organization typically implement decision making systems to make it possible for their online community to reach agreements [1]. It has also played an important role for consumer industry to maintain supply chain of tea like goods [2]. One of the great applications of blockchain technology is in recent advanced construction to identify process, policy, and society perspective needs [3]. It also helps in making payment to construction contractor without bothering about their location [4]. Blockchain also help in making sharing-based payment system in agricultural sector, where producer and consumer can share payment information related to the crops which simplify payment system. [5].

There is more use of cryptographic functions; here, users mainly use public and private keys to sign digitally and do transactions securely within the system [6]. The industry which would like to implement blockchain technology needs to understand fundamental aspects of the implementation mechanism of the blockchain network. It does not require a higher level of software abstraction to modify the working of data.

One of the most significant issues in blockchain networks is validating how users or participants agree that the transaction performed on the network is valid. There are many models available for validation transactions with advantages and disadvantages based on the type of business.

### 1.2 Category of Blockchain

#### 1.2.1 Permissionless

It is a blockchain network implemented using a decentralized ledger platform that is open for everyone to publish their blocks without permission from any authority. It usually is open-source software that is available for anyone to download freely.

Here, malicious users can also attempt to publish blocks to alter the code of the system. This network often uses a multipart agreement which requires the users to expend or maintain the resources when it tries to publish blocks to avoid this type of activity.

### 1.2.2 Permissioned

It is a type of blockchain network where some authority (centralized or decentralized) verifies the block, which is published by the user. It allows only authorized users to enter or maintain the network; it allows us to put specific restrictions on reading access and make transactions.

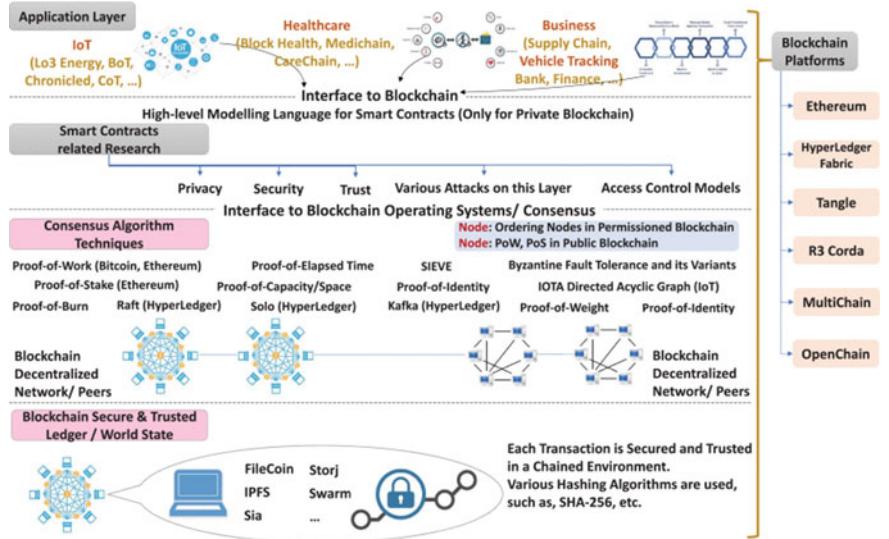
Each user has its own identity, which maintains the level of trust with each other as they all are authorized users to publish the blocks; if any misbehavior observer, it will be revoked later on. Consensus models usually faster and require spending less amount on commutation in case of a permission blockchain network.

## 2 Literature Review

It is helpful to change the face of the digital transaction the way money transfer from one person to another for the business transaction by omitting the third party; it allows us to make the system efficient [6]. It is also helpful in designing a system for procurement, distribution, handling, and procession materials for the halal supply [7]. It can help for securing medical data and understand medical treatment with Hadoop combining blockchain [8]. Nowadays, various social media frameworks are available; so to securely post messages on social media, blockchain plays an important role [9]. It is helpful to reduce the overall energy consumption required to flood data over the Internet. It can create an alternative payment system for survival of many country [10]. Now, it is a time of automation various IOT-based system is developed using blockchain technology [11]. Blockchain system functions as a proof-of-concept prototype, showing the feasibility of applying blockchain in smart homes with IoT functionalities [12]. Blockchain is also used to provide vaccination status where proof of concept is used [13]. Figure 1 shows the various application and working of blockchain ecosystem.

## 3 Blockchain Components

It looks complex, but it is prepared in a simple way by observing all the components of the blockchain network individually. In this section, few components of blockchain are explained.



**Fig. 1** Summary of blockchain ecosystem [13]

### **3.1 Cryptographic Hash Functions**

One of the most critical components of this technology is the cryptographic hash function used to perform many operations. It uses the method of the hash function, which applies to data to generate unique output like message digest, which is an input of any size like text or image. Even change in the tiny bit creates a different message digest. It generates three types of data:

- They are preimage resistant.
  - They are second preimage resistant.
  - They are collision resistant.

### 3.2 *Transactions*

It represents communication between two users. In cryptocurrencies, it is known as the transfer of bitcoin-like currency between two blockchain network users. In the case of business-related transactions, it could be a way of recording activities occurring on digital or physical assets. Every block contains zero or one transaction.

Each blockchain network has its type of data, so when data contain some transaction-related information, it depends on the type of blockchain implementation.

### ***3.3 Asymmetric-Key Cryptography***

In this technology, public-key cryptography is also popular as an asymmetric-key algorithm used. It uses a pair of public and private keys, which are related to each other. Each user has a public key for encryption and decryption. A private key is a valid approach to protect data [13].

It is difficult to identify the private key even though both the critical public and private are related. The private key is utilized here for encryption, and decryption is happening with the help of a public key. It provides trust between two unknown parties by providing integrity and authenticity of transactions by keeping transactions public. It proves that authenticate user has sent data.

### ***3.4 Ledgers***

It is helpful to keep history and track of the exchange of goods and services, and it is usually a collection of transactions recorded on the blockchain network. It is recorded digitally more often in an extensive database operated and owned by a centralized authority. It is implemented in a distributed manner with one server's help or distributed over a cluster of servers [14].

It is more in demand to have distributed ledger ownership. Blockchain networks provide both the type of facilities like distributed physical architecture and ownership. It usually contains many computers when distributed ownership is created to compare with centrally managed architecture. Interest increases just because of various features like trust, security, and reliability provided in distributed ledgers containing central ownership.

### ***3.5 Blocks***

In this type of network, users submit transactions using software like desktop applications or other related software. It sends this type of transaction to the nodes in the blockchain network. This node is of two types publishing or non-publishing.

Once added by the publishing node, the block typically contains a block header and block data. A block header usually contains data about the data or metadata for a particular block.

### **3.6 Chaining Blocks**

Block forms a chained block by connecting where each block contains the hash digest of the previous block's header. If the header of the previous block changed hash and gets changed in the next block, this results in a change of hash in all the blocks followed by the first block. This feature provides an easy way to detect and reject altered blocks. Next section discusses about the network block of a nodes and implementation mechanism.

## **4 Article Taxonomy and Fork-Based Blockchain Network**

In this article, taxonomy of blockchain technology and its application are covered. The article focusses on literature review of blockchain type, various types of blockchain implementation, smart contracts, component of blockchain, and advantages and application of blockchain with future work. Section 1 deals about introduction of article followed by Sect. 3 which covers about the blockchain components used to create blockchain network. In Sect. 4, comparative analysis is made based on previous study, and Sect. 5 gives introduction about smart contract made using blockchain. Section 6 discusses about limitations, Sect. 7 discusses various applications of blockchain, and Sect. 8 summarizes the future work of blockchain technologies.

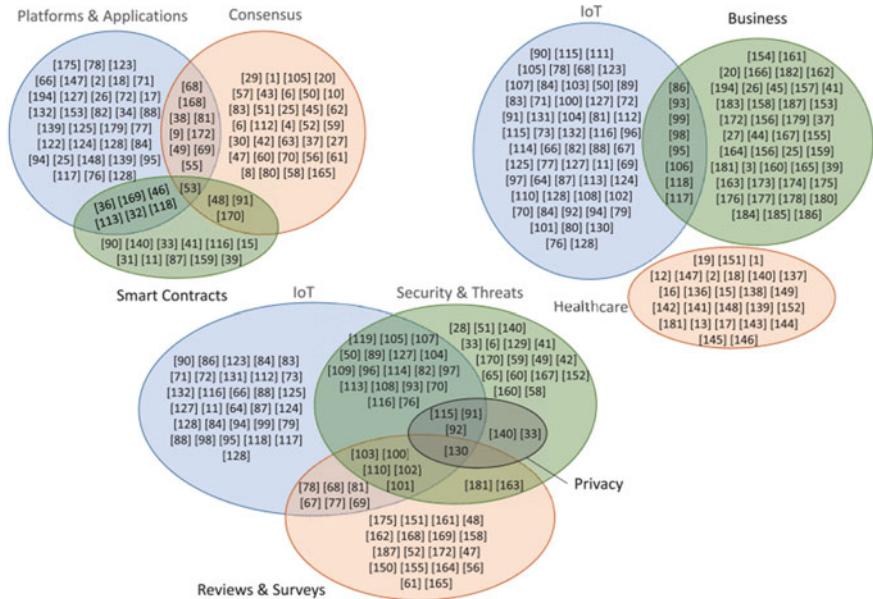
The use of blockchain technology is divided into multiple categories like consensus techniques, smart contracts, the IoT, healthcare, business, and various platforms that are related to blockchain. It is also had intersections between different area as shown in Venn diagram of Fig. 2 in which summary of articles related to platform and application is given. There are many review papers published related to IOT, business, and healthcare-related area. There are few papers that also cover interdiscipline application like IOT and business. Figure 2 also summarizes security, threats, and privacy in IOT (Internet of things).

Table 1 gives implementation of blockchain network for various business.

The following section discusses about types of blockchain implementation. There are two types of forks to implement blockchain network.

### **4.1 Soft Forks**

It is a particular type of implementation of a blockchain network, capable of creating a backward-compatible network. Here, the node which is not updated can also communicate with the updated node. If the node does not require to be updated, then rules of updation will not be followed.



**Fig. 2** Venn diagram of existing research related to various platform

**Table 1** Implementation of blockchain for business

	Public blockchains	Consortium blockchains	Private blockchains
Credit mechanism	Proof of work	Collective endorsement	Self-endorsement
Bookkeeper	All participants	Participants decide in negotiation	Self-determined
Incentive mechanism	Needed	Optional	Not needed
Typical application scenario	Bitcoin	Clearing	Audits

Reducing the size of a block from 1 MB to 0.5 MB in Blockchain is an example of a fictional soft fork. The block which is updated adjusts their block size and continues their transaction operation in normal mode. The nodes which do not update can see the status of this block as valid as the change made is not violating the rules. If the new node size will be greater than 0.5 MB, then the updated node gets rejected and considered an invalid node.

## 4.2 Hard Forks

It is the type of blockchain network implementation that is not backward compatible. Here, all publishing nodes require to switch to a specific block number at a given point of time using an updated protocol. All the existing nodes require to update as per the new protocol so that the newly formatted blocks do not have rejection. In the new implementation, the node which is not updated cannot continue to transact as they are designed to follow a specific version of the block and reject the block which does not follow that version.

## 4.3 Cryptographic Changes and Forks

In use of cryptographic technologies in a blockchain network, it is advisable to use the hard fork to implement the network depending on the type of flaw; if a flaw exists between underlying algorithms, it requires having a robust algorithm for all future clients [15]. The switching block must require locking all existing blocks into SHA-256 (secure hash algorithm), and the all-new blocks will require to use a new hashing algorithm. Blockchain network will select a cryptographic algorithm based on its need from the available list of an algorithm.

Now in the next section, discussion about the smart contract of the blockchain network is made.

## 5 Smart Contracts

It is a term defined by Nick Szabo in 1994 to specify the transaction which is executed with the help of computers. Smart contracts are helpful for a specific purpose like to satisfy contractual conditions, for example, payment terms and to minimize both types of exception that occurs accidentally or by a malicious user and reduce the need for trusted intermediates.

It enhances the power of blockchain technology. It collects code and data deployed on the blockchain network, and all the transactions signed cryptographically, e.g., Ethereum's smart contracts, Hyperledger.

When nodes on the blockchain network execute the smart contract, they will get the same result, and the result will be stored on the blockchain network. In a blockchain network, the transaction which users perform sends data using the public function of intelligent contract [16].

## 6 Limitation of Blockchain

In the previous section, features of blockchain were discussed. Now in this section, various limitations of blockchain network implementation are discussed.

### 6.1 *Immutability*

There are rumors that blockchain ledgers are not changeable, but they are not entirely accurate. As blockchain ledgers are tamper-evident and tamper-resistant, they are helpful for financial transactions. There is a possibility of changing the Blockchain in a specific case, so it is not considered entirely immutable. There is a specific condition in which it violates the condition of immutability [17].

### 6.2 *User Involvement in Blockchain*

Blockchain network follows specific rules, practices, and processes through which it gets direction and control. There are rumors that no one controls the blockchain network, and no one has ownership of it, but it is not entirely accurate. When permission blockchain networks are implemented, they are typically set up and run by some owner who governs the network. In the permission less blockchain network, there are many users, publishing nodes, and software developers who control the blockchain network. Here, each group of developers have the right to control the network, which will give the direction to the blockchain network.

### 6.3 *Network-Based Attack*

In a blockchain network when transactions are done, they are all immutable and secure that cannot be changed, but it is partially true because it happens only when the transaction gets published or included in the publishing block. Those transactions available on blockchain networks but not included in publishing blocks are still vulnerable to various attacks.

Blockchain networks and their applications are not safe from malicious users who perform network scanning operations to discover vulnerabilities. A new blockchain network also contains some vulnerabilities and weaknesses discovered and attacked by this malicious user.

## 6.4 Malicious Users

In permissionless blockchain networks, users are not allowed to execute their code on the blockchain network as it is not possible to do one-to-one mapping of the users of the different blockchain networks.

- It ignores transactions from specific types of nodes or even of the country.

They will create the most prolonged change to confuse users to migrate on a longer chain, which will result from most of the work done on the existing chain.

- It also refuses to transmit blocks to the other nodes, which disrupts the distribution of information and creates an issue of shortage of information.

## 6.5 No Trust

There is a belief that there is no trusted third party available in blockchain network; usually, blockchain networks are trustless, but it is not the case with all type of blockchain network [17].

In a permissionless blockchain network, there is no trusted party certifying the transaction, but there is still a requirement to implement some mechanism to trust the network, especially when the network is permissionless.

## 6.6 Resource Usage

Blockchain technology is nowadays implemented worldwide, and all transactions are also verified; it is also synchronized among multiple users. Many nodes on a blockchain network require more time to process the transaction that consumes much electricity [18].

## 6.7 Identity and Infrastructure of Public Key

Blockchain technology comes with critical public infrastructure; it does not mean it supports the concept of identity. In this type of network, there is no one-to-one relationship existing between private critical pairs for users; usually, the user has multiple private keys, and also, there is a relationship between blockchain address and public keys; here, multiple addresses are derived from a single public key [19].

## 7 Application of Blockchain

Blockchain technology is new, so it is difficult for business entities to decide how to use it efficiently. They have a fear of losing data which creates frustration to implement it universally. Blockchain technology usually is suitable for the following types of applications [20].

- Application which include a large number of participants
- Application having participants in a distributed manner
- Application requiring less use of trusted third party
- When there is a requirement of transfer of information between two parties
- Application like the digital land record, digital property identification
- Application requiring decentralized naming service
- Application requiring cryptography-based security.

## 8 Conclusion

In this article, introduction of blockchain, its component, and various types of blockchain implementation are covered. This article also covered the previous works done in the field of blockchain network. It also covers various advantages and limitations of blockchain as there is much application of blockchain. User can create a network using Blockchain as per their application and need; as discussed, it has many advantages. It is applicable in almost all fields for distributed, reliable and secure transactions of data or money. In the future, blockchain network can be explored to be used for the medical field using deep learning. It can also be useful for the current COVID situation to track total amount of person vaccinated.

## References

1. Faqir-Rhazoui, Y., Arroyo, J., Hassan, S.: A comparative analysis of the platforms for decentralized autonomous organizations in the Ethereum blockchain. *J. Internet Serv. Appl.* (2021). <https://doi.org/10.1186/s13174-021-00139-6>
2. Li, J., Kassem, M.: Applications of distributed ledger technology (DLT) and Blockchain-enabled smart contracts in construction. *Autom. Constr.* (2021). <https://doi.org/10.1016/j.autcon.2021.103955>
3. Paul, T., Mondal, S., Islam, N., Rakshit, S.: The impact of blockchain technology on the tea supply chain and its sustainable performance. *Technol. Forecast. Soc. Change* (2021). <https://doi.org/10.1016/j.techfore.2021.121163>
4. Hamledari, H., Fischer, M.: Construction payment automation using blockchain-enabled smart contracts and robotic reality capture technologies. *Autom. Constr.* (2021). <https://doi.org/10.1016/j.autcon.2021.103926>
5. Elubebek kyzy, I., Song, H., Vajdi, A., Wang, Y., Zhou, J.: Blockchain for consortium: a practical paradigm in agricultural supply chain system. *Expert Syst. Appl.* (2021). <https://doi.org/10.1016/j.eswa.2021.115425>

6. Surjandari, I., Yusuf, H., Laoh, E., Maulida, R.: Designing a Permissioned Blockchain Network for the Halal Industry using Hyperledger Fabric with Multiple Channels and the Raft Consensus Mechanism (2021). ISSN: 2196-1115. <https://doi.org/10.1186/s40537-020-00405-7>
7. Zhang, X., Wang, Y.: Research on intelligent medical big data system based on Hadoop and blockchain (2021). <https://doi.org/10.1186/s13638-020-01858-3>
8. Arquam, M., Singh, A., Sharma, R.: A blockchain-based secured and trusted framework for information propagation on online social networks (2021)
9. Xu, S., Liao, B., Yang, C., Guo, S., Hu, B., Zhao, J., Jin, L.: Deep reinforcement learning assisted edge-terminal collaborative offloading algorithm of blockchain computing tasks for energy internet. <https://doi.org/10.1016/j.ijepes.2021.107022,2021>
10. Aysan, A.F., Sadriu, B., Topuz, H.: Blockchain Futures in Cryptocurrencies, Trade and Finance: A Preliminary Assessment. BEMP (2020). <https://doi.org/10.21098/BEMP.V23I4.1240>
11. Fat, J., Candra, H.: Blockchain application in internet of things for securing transaction in ethereum TestNet. IOP Conference (2020). <https://doi.org/10.1088/1757-899X/1007/1/012194>
12. Ma, M., He, Z., Xu, Q., Li, X.J.: Design and Development of Smart Home Sensing Supported by Blockchain Technology. ACM (2019). <https://doi.org/10.1145/3377170.3377281>
13. Odoom, J., Soglo, R., Danso, S.A., Xiaofang, H.: A Privacy-Preserving Covid-19 Updatable Test Result and Vaccination Provenance Based on Blockchain and Smart Contract. ICMRSiSIT (2019). <https://doi.org/10.1109/ICMRSISIIT46373.2020.9405872>
14. Hochstein, M.: Don't Use a Blockchain Unless You Really Need One. CoinDesk (2018)
15. Mell, P., Kelsey, J., Shook, J.: Cryptocurrency Smart Contracts for Distributed Consensus of Public Randomness (2017). [https://doi.org/10.1007/978-3-319-69084-1\\_31](https://doi.org/10.1007/978-3-319-69084-1_31)
16. Greenspan, G.: The Blockchain Immutability Myth. CoinDesk (9 May 2017)
17. Peck, M.: Reinforcing the links of the blockchain. IEEE (2017)
18. Peck, M.E.: Do you need a blockchain? IEEE Spectrum: Technology, Engineering, and Science News. IEEE Spectrum, <https://spectrum.ieee.org/computing/networks/do-you-need-a-blockchain> (2017)
19. Narayanan, A.: Analyzing the 2013 Bitcoin Fork: Centralized Decision-Making Saved the Day. MultiChain (2015)
20. Nakamoto, S.: Bitcoin: A Peer-to-Peer Electronic Cash System. <https://bitcoin.org/bitcoin.pdf> (2008)

# Hybrid Statistical and Deterministic-Based Pedestrian Tracking Algorithm for Location-Based IoT Applications



J. P. D. Manoj Sithara and M. W. P. Maduranga

**Abstract** Location-based services are considered one of the primary applications in the Internet of Things (IoT), where the geographical location of a moving object is involved. This paper presents an improved statistical–deterministic algorithm using information received from cellular base stations in a non-line-of-sight (NLOS) environment. This approach can also be utilized in an indoor positioning system. This novel algorithm is developed as a hybrid model using deterministic modeling for real-time dynamic and statistical techniques. Signals receiving at the cellular base stations from the user’s mobile phone are used to estimate the coordinates of a pedestrian. Simulations were conducted in both 2D and 3D environments. Meanwhile, the algorithm was tested to understand the impact of the number of base stations in the cellular network, using the error probability of the coordinates triangulation.

**Keywords** Pedestrian tracking algorithms · IoT · Outdoor localization

## 1 Introduction

Recent advances in IoT development enable state-of-the-art location-based services in smart cities. Pedestrian tracking could consider as one of the exciting and essential applications. Moreover, advancements in dynamic mobile localization technologies have constantly been exploited by the militaries, smart cities as mainstream vehicle and pedestrian tracking applications. With the advent of smartphone technology, positioning methods based on cellular networks in pedestrian tracking have rapidly evolved. Yet, cellular network-based pedestrian tracking still faces major

---

J. P. D. Manoj Sithara (✉)

Department of Electrical and Computer Engineering, The Open University of Sri Lanka, Nawala, Nugegoda, Sri Lanka

e-mail: [manojjsithara1984@gmail.com](mailto:manojjsithara1984@gmail.com)

M. W. P. Maduranga

Department of Computer Engineering, General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka

e-mail: [pasanwellalage@kdu.ac.lk](mailto:pasanwellalage@kdu.ac.lk)

accuracy issues due to multipath and non-line-of-sight (NLOS) errors. These factors, coupled with the limited number of cellular base stations, create reliability issues. In urban contexts, the severity of NLOS is exacerbated because high-rise buildings and structures obstruct signal propagation. These impediments subject the signal to reflections and refractions, preventing the signal from flowing directly from the base station to the mobile devices. In NLOS conditions, the more dependable Global Positioning System (GPS) suffers from these signal propagation aberrations as well. Most cutting-edge solutions for solving this problem [1–3] are based on the assumption that mobile devices have inertial sensors, gyroscopes, or pedometers. However, size and cost limits prevent inertial sensors from being used in most mobile devices, limiting their practical application. A GPS tracking unit is generally carried by a moving vehicle or person who uses the global positioning system to determine and track its precise location using only four satellites. Nevertheless, the system's accuracy will vary considerably at certain conditions [4, 5]. Therefore, properties of the non-line-of-sight (NLOS) cellular environment have been heavily investigated to develop algorithms for pedestrian tracking proposed [3–9]. Those statistical methods explain in works of literature [6, 7], are have been improving accuracy than standard GPS systems, and can reduce the cost of mobile equipment's [1, 2] to add sensors to pedestrian tracking. But, if we reduce the fact of the statistical interface to using actual human walking base half deterministic and the half statistical way, we can obtain an astonishing degree of accuracy to pedestrian tracking in (NLOS) cellular environment and reduce the computational complexity of algorithms. Here, all the relevant literature in Refs. [6, 7, 10] propose statistical interfacing of developing algorithms. Moreover, works are existing on possible pedestrian tracking in indoor environments using received signal strength (RSS) and machine learning [9, 11, 12].

The recent work available related to this algorithm [6] has few limitations. In their algorithm [6], it is required to provide the initial coordination of the pedestrian to perform the algorithm. The environment has been considered as without real obstacles such as buildings, trees, and vehicles. Further, the number of base stations used for simulation is restricted to a fixed value, where the impact of the number of the base station on the algorithm's accuracy is not investigated. Also, the simulation scenario is limited to the 2D environment [13–16].

In existing algorithms, it considers the entire statistical approach for pedestrian tracking. However, these methods are unable to identify the deterministic component of the scenario. Moreover, these algorithms only consider the path of walking as the only object in the scenario, and other obstacles such as buildings and trees are not considered. Due to the reasons mentioned above, existing algorithms for pedestrian tracking are not realistic in real scenarios. For example, if pedestrians walk through indoor environments such as buildings and other open areas which did not have proper roads. In this paper, we propose a scheme to improve the accuracy of pedestrian position fine-tuning and predict the location in the 3D environment. This algorithm will be applicable for all pedestrian tracking scenarios with no predefined map, including unknown territory, buildings, parks, etc. Our strategy consists of three techniques. In the first, we propose using weighted prior position estimates instead of using raw preliminary position estimates during the fine-tuning process in the

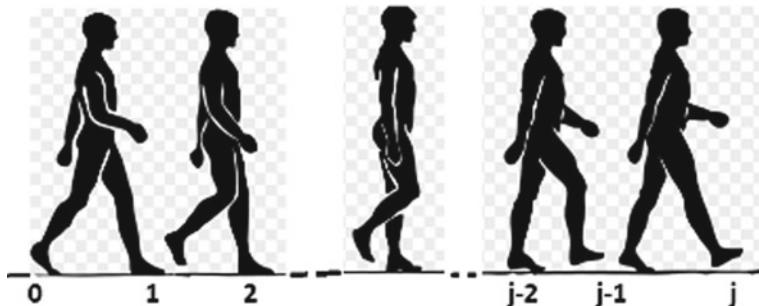
scheme in Ref. [5]. This weighting reduces the impact of irrelevant past information. Second, we propose repeating the fine-tuning iteratively, with the previously fine-tuned location serving as the input to the current iteration. As a result, the final estimate is intended to converge to a position with the least amount of inaccuracy. Finally, we present a barrier-based mapping approach that minimizes error even further by approximating final estimates to a preset set of correct places. The rest of the paper is organized as follows. Section 2 discusses the existing low-complex position fine-tuning scheme in Ref. [6], which is the basis of our work. In Sect. 3, we introduce the proposed three techniques for estimated position fine-tuning followed by simulation results in Sect. 4 to verify the accuracy improvement by the proposed scheme under different conditions. Finally, Sect. 5 concludes the overall performance of the algorithm.

## 2 System Model

An approach on statistical and deterministic is taken place to track accurate portions of a mobile pedestrian in an (NLOS) environment. We improve the mathematical model given in [6]. As per Fig. 1, the last two steps for three different displacements in a direction  $x^\mu$  (i.e.,  $(j-2, j-1, j)$  point in space to  $x, y, z$ -directions to ( $\mu$ : 1, 2, 3). Hence, derive the discrete velocity component and the discrete  $v_{j-1}^\mu, v_j^\mu$  and acceleration  $a_j^\mu$ .

$$v_{j-1}^\mu = \frac{x_{j-1}^\mu - x_{j-2}^\mu}{t_{j-1}^\mu - t_{j-2}^\mu} = \frac{\Delta x_{(j-1, j-2)}^\mu}{\Delta t_{(j-1, j-2)}} \quad (1)$$

$$v_j^\mu = \frac{x_j^\mu - x_{j-1}^\mu}{t_j^\mu - t_{j-1}^\mu} = \frac{\Delta x_{(j, j-1)}^\mu}{\Delta t_{(j, j-1)}} \quad (2)$$



**Fig. 1** Actual human walking with different time cuts

$$a_j^\mu = \frac{v_j^\mu - v_{j-1}^\mu}{t_j^\mu - t_{j-1}^\mu} = \frac{\Delta v_{(j,j-1)}^\mu}{\Delta t_{(j,j-1)}} \quad (3)$$

According to the previous session, the discreet version of Newton's equation for displacement in direction can quantify the below [17].

$$\Delta x_{(j,j-2)}^\mu = v_{j-1}^\mu \Delta t_{(j,j-2)} + 1/2 a_{(j-1)}^\mu \Delta t_{(j,j-2)}^2 \quad (4)$$

### 3 Proposed Improvements

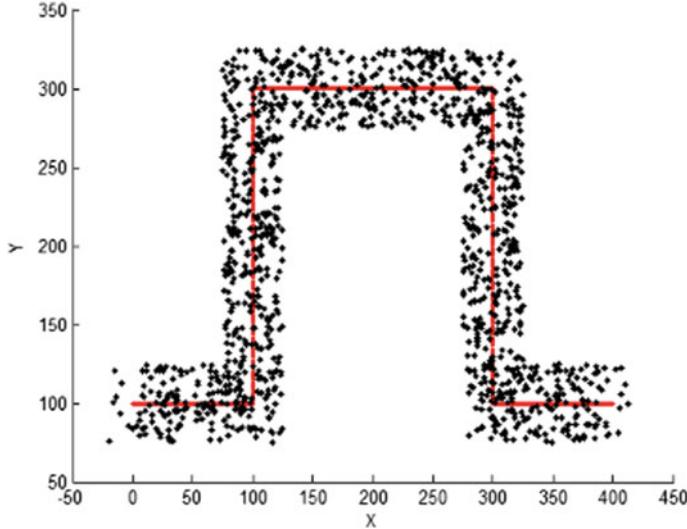
#### 3.1 Standard Error Realization for Base Stations

From the above model, to extend the proposed error localization of base stations into several dimensions, explain intenser  $x^\mu$ . Therefore, demand the statistical error optimization relative to  $n$  number of base stations for  $\mu$  dimensions within  $R^\mu$  the radius to the actual pedestrian position  $x_{true,j-2}^\mu$  at the time  $t_{j-2}$ . Therefore, it can be a demand error with  $i^{th}$  base station and has the relation of  $\sum_\mu (x_{i,j-2}^\mu - x_{true,j-2}^\mu)^2 \leq R_i^2$ . Consequently, it can compute the average value of the square sum as shown in Eq. 5.

$$\sqrt{\frac{\sum_{i=1}^n \sum_\mu (x_{i,j-2}^\mu - x_{true,j-2}^\mu)^2}{n}} \leq R \quad (5)$$

By using terms standard divination terms  $\sigma^\mu$  in each dimension as equation shown below for time-invariant standard deviation of R, the fact can be used as  $\sigma_{j-2}^\mu = \sigma^\mu$  That, generalize the errors will only depend on directions, not by time accounting to the above factors present following equation to base station error of estimation  $\sum_\mu (\sigma^\mu)^2 \leq R^2$ .

To reduce the statistical interface of the algorithm to a half statistics and half deterministic model, we have to introduce an entirely new statistical interface compared to the statistical method proposed in papers [6, 7]. Here the nature of the equation completely changes to statistical into deterministic manner because of any average weighting to find velocity and acceleration components and not need any average weighting because we can measure actual components of velocities as real human walk. We avoided those statistical weighting methods to actual walking, a massive improvement to predicted locations, and reduced the algorithm's computational complexity. Here, the modified equation set compares to the proposed equation in Refs. [6, 7]. Thus, using the weighted average of pre-estimate to the average pedestrian position relative to base stations at the time  $t_{j-2}$  is a different method mentioned in the literature [6, 7]. Hence that propose, the equation shows below.



**Fig. 2** Actual path and predicted path

$$X_{pre,j}^{\mu} = \overline{X_{est,j-2}^{\mu}} + \Delta X_{(j,j-2)}^{\mu} \quad (6)$$

By modifying the estimated average tensor,  $\overline{X_{est,j-2}^{\mu}}$  relative to the  $n$  number of base stations as compared to literature mentioned in [6, 7] at the time  $t_{j-2}$  can have  $X_{est,(j-2)}^{\mu} = \sum_{i=1}^n X_{i,(j-2)}^{\mu}/n$ . Typical base station localization and actual path of pedestrian walking are shown in Fig. 2. Furthermore, the line shows the true path, and the dotted indicates the position of localized position of a pedestrian using data of base stations.

Then the weighted average of base station localization can be computed using the following tensor in Eq. 7 [18].

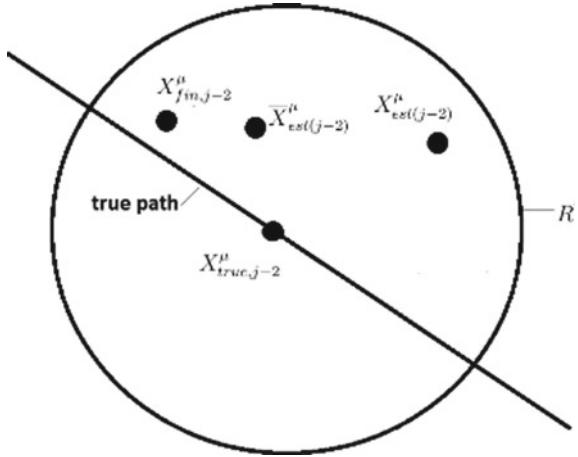
$$\overline{X_{est,j-2}^{\mu}} = \frac{W_1 X_{est,j-2}^{\mu} + W_2 X_{fine,j-2}^{\mu}}{W_1 + W_2} \quad (7)$$

If  $W_1$  and  $W_2$  are predefined weighted values, then  $\overline{X_{est,j-2}^{\mu}}$  is a placed value between  $X_{est,j-2}^{\mu}$  and  $X_{fine,j-2}^{\mu}$ , where  $W_1$  and  $W_2$  can be obtained via trial and error for the best-optimized point.

Here  $X_{fine,j-2}^{\mu}$  is the fine-tune value of the algorithm at the time  $t_{j-2}$ . However, propose equation tensor  $X_{pre,j}^{\mu}$  gives the infinity advance equation to the predicted pedestrian position than in [6, 7]. Thus,  $W_1, W_2$  equations are weighting at can be selected by trial and error.

Figure 3 shows that the weighted average can be close to the actual pedestrian path for some tune value  $W_1, W_2$ . From that predicted estimate value in the literature's [6, 7] less efficient than the newly proposed method in a weighted average of  $\overline{X_{est,j-2}^{\mu}}$ .

**Fig. 3** Geometric view of localization

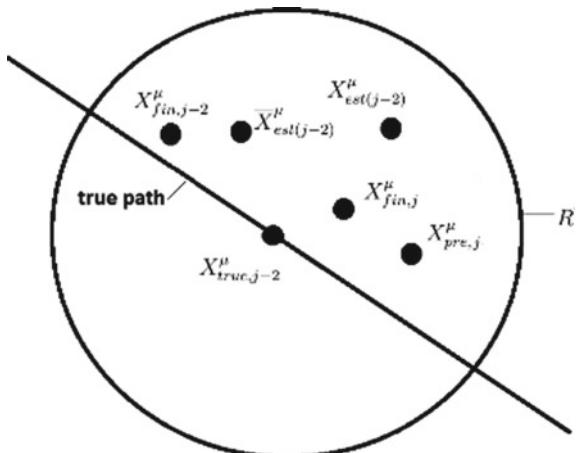


Computing the predicted value at  $j^{\text{th}}$  time as we add deterministic part of human walking to the  $X_{\text{est},j-2}^\mu$ . We can develop the relation between predicted position  $X_{\text{pre},j}^\mu$  and fine-tune at the time  $X_{\text{fine},j-2}^\mu$  and thus build up a relationship using a weighted average method  $W_4 \leq W_3$  [18]. The final fine-tune value  $X_{\text{fine},j}^\mu$  at the time  $t_j$  can be proposed by the following equation.

$$X_{\text{fine},j}^\mu = \frac{W_3 X_{\text{pre},j}^\mu + W_4 X_{\text{fine},j-2}^\mu}{W_3 + W_4} \quad (8)$$

Figure 4 shows the final geometric representation of the component of coordination suggested by the algorithm.

**Fig. 4** Final view of localization



### 3.2 Standard Time Realization for Fine-Tuning the Value

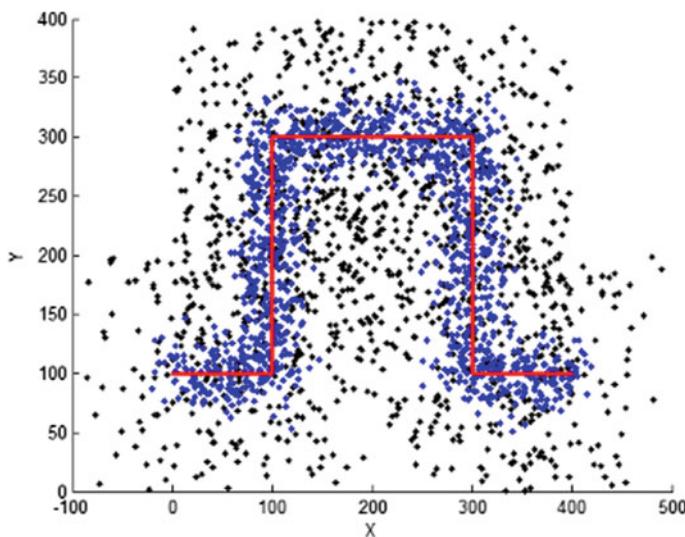
In conclusion, the final error of fine-tune value of the time point  $j$  is the error of the actual path value of the pedestrian. Therefore, we introduce error at the time for mutually orthogonal dimensions, as shown below.

$$e_j = \sqrt{\frac{\sum_{\mu} (x_{true,j}^{\mu} - \hat{x}_{true,j}^{\mu})^2}{3}} \quad (9)$$

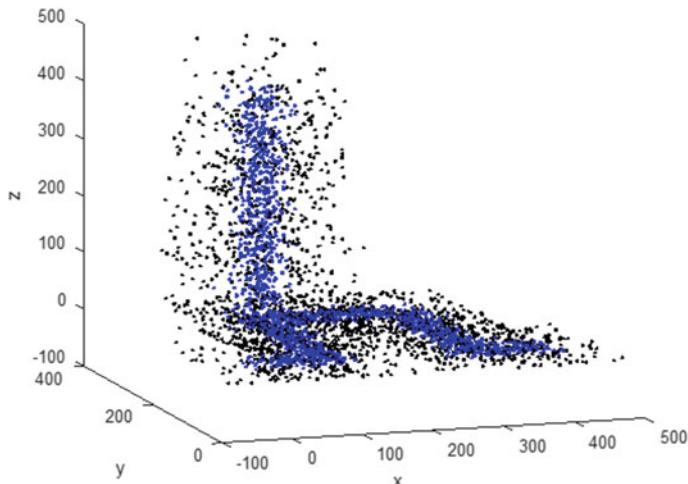
## 4 Simulation of Algorithm

The localized distance was kept during the simulations and assumed that the walking pattern is average [6, 7]. During testing, it was observed that the following weighting is appropriate for more accurate algorithm simulations.  $\{W_1 = 40, W_2 = 60, W_3 = 60, W_4 = 40\}$ .

The simulation area was restricted to size  $350 \times 450 \text{ m}^2$ . The distances between two steps human walk are assumed as the maximum and minimum values as,  $L_{\max} = 0.64 \text{ m}$  and  $L_{\min} = 0.26 \text{ m}$ . Simulations were done using MATLAB 2018. Figure 5 shows the simulation of the algorithm in a 2D environment where x- and



**Fig. 5** 2D simulation for 100 m localized radius



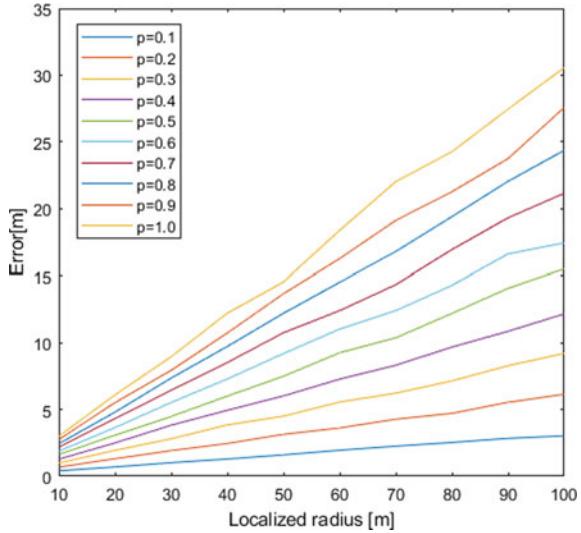
**Fig. 6** 3D simulation for 100 m localized radius

y-coordinates are considered, where scattered dots in black are the triangulated coordinates received from base stations. Blue color dots represent the predicted coordinates given by the algorithm—meanwhile, the actual path representing in red color in the figure. Figure 6 shows the simulation results in a 3D environment, where the algorithm predicts the walking pedestrian's x-, y-, and z-coordinates. The representation of colors of the graph is the same as Fig. 5. However, in 3D environments, it is not convenient to show the actual path simulation environment due to the natural 3D view.

Furthermore, simulations were conducted to study the relationship between the number of the base station and the algorithm's accuracy. We assumed that the number base station and error probability of triangulating coordinates are inversely proportional. With that assumption, we account for binomial distribution for a localized distance of error. The binomial distribution is assumed so that one try of the base-stations network has the same distance of  $\sqrt{2}$  meter error from actual path for 100% error probability. Therefore, we assumed the number of trials from base stations has total tries the same as the rounded number for a distance of localized radius. Thus with 100%, error probability gives base triangulation coordinate of a point on localized radius and error probability < 100%, indicating coordinate in points on the area of localized radius. When localized distances are increased, the error of the predicted coordinates gives a linear relationship, as shown in Fig. 7.

Moreover, when the error probability,  $P$ , increases with the localized radius, the prediction error will increase. This occurs due to the reduction of accuracy of the triangulation. Figure 7 shows the error of predicted coordination against the localized radius of the mobile base station.

**Fig. 7** Error versus localized radius distance



## 5 Conclusions

In this work, we proposed a novel algorithm to predict the location of a moving pedestrian. Our algorithm is based on using deterministic dynamics and statistical optimization for mobile base station triangulation. In this algorithm, we have been modeled statements to predict the geographical location of a moving human in a real-time environment. The proposed algorithm was simulated different conditions in 2D and 3D environments. Where it has used best fine-tune values of weightings via trial and error. Also, simulated the impact of the number of base stations against the accuracy of the algorithm. When the number of base stations is increased, the accuracy of the algorithm is increased significantly.

## References

1. Foxlin, E.: Pedestrian tracking with shoe-mounted internal sensors. *Comput. Graph. Appl.* **25**(6), 38–46
2. Jin, Y., Toh, H.-S., Soh, W.-S., Wong, W.-C.: A robust pedestrian tracking systems with low cost sensors, PerCom 2010 pp. 222–230, 21–25 March 2011
3. Nnenna, E.J., Onyekachi, O.H.: Mobile positioning technique in GSM cellular Networks. *Int. J. Comput. Technol. Electr. Eng.* **2**(6), 21–29 (2012)
4. Kegen, Y., Dutkiewicz, E.: Geometry and motion base positioning algorithms for mobile tracking in NLOS environment, GLOBECOM 2010, pp. 1–5
5. Gustafsson, F., Gunnarsson, F.: Mobile positioning using wireless network:possibilities and fundamental limitations based on available wireless network measurements. *IEEE Signal Process. Mag.*

6. Lumbini, C., Balasuriya, N.: An improved pedestrian tracking algorithm in NLOS environment. Paper publish in 2015 10th Asia Pacific Symposium on Information and Telecommunication Technologies (APSITT)
7. Guptha, C., Biswas, D.: Pedestrian tracking algorithm in NLOS environment. Proc. IEEE Conf. Adv. Netw. Telecommun. Syst. 73–75 (2012)
8. Vijesh, J.C., Raj, J.S.: Location-based orientation context dependent recommender system for users. J. Trends Comput. Sci. Smart Technol. (TCSST) **3**(1), 14–23 (2021)
9. Maduranga, M.W.P., Abeysekara, R.: Supervised machine learning for RSSI based indoor localization in IoT applications. Int. J. Comput. Appl. **183**(3), 26–32 (2021)
10. Wu, S., Li, J., Liu, S.: Improved positioning algorithm based on two step least square in NLOS environments. J. China Univ. Post Telecommun. **18**(5), 58–63 (2011)
11. Maduranga, M.W.P., Ganepola, D., Kathriarachchi, R.P.S.: Comparison of trilateration and supervised learning techniques for BLE based indoor localization. In: 14th KDU International Research Conference(KDU-IRC), September 2021
12. Maduranga, M.W.P., Abeysekara, R.: Comparison of supervised learning-based indoor localization techniques for smart building applications. In: International Research Conference on Smart Computing and Systems Engineering (SCSE), September 2021s
13. May, P., Ehrlich, H.C., Steinke, T.: ZIB structure prediction pipeline: composing a complex biological workflow through web services. In: Nagel, W.E., Walter, W.V., Lehner, W. (eds.) Euro-Par 2006. LNCS, vol. 4128, pp. 1148–1158. Springer, Heidelberg (2006)
14. Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco (1999)
15. Czajkowski, K., Fitzgerald, S., Foster, I., Kesselman, C.: Grid Information Services for Distributed Resource Sharing. In: 10th IEEE International Symposium on High Performance Distributed Computing, pp. 181–184. IEEE Press, New York (2001)
16. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. Technical Report, Global Grid Forum (2002)
17. Nelkon, M., Parker, P.: Mechanics and properties of matter, 5th edn, pp. 1–47. Heinemann Educational Books Ltd., London
18. Walpole, R.E., Myers, S.L., Ye, K.: Probability ans Statistics for Engineers and Scientist, 8th edn, pp. 25–85. Dorling Kindersely Pvt. Ltd., No. 11, Community center, Punchsheel Park, New Delhi, India (2013)

# MAC-Based Secure Data Transmission in Vehicular Ad hoc Networks



**T. Kalaichelvi, L. Jabasheela, P. S. Ramaprabha, M. Shobana, G. Dhanalakshmi, W. Gracy Theresa, and H. Rashini**

**Abstract** The transmission of safety-related signals is crucial for effectively implementing applications in vehicle ad hoc networks (VANET). A body MAC technique is offered for high-density applications. The feature enables cars to converse with someone in a side impact manner prior to data transfer. Through the body mechanism, the cars generate a logistic queue, and the queue might reach the stream once its length reaches a particular threshold. When the buffer consumes the bandwidth, cars in the lineup will receive it using time-division multiple access (TDMA). A buffer shall compete for channel allocation on interests of all locations in the line, significantly minimising conflict from a single node. The space a queued selects to contact the stream is decided by the arrival rate of the associated procedure, as opposed to fully random access. Because other waits have been in the soul method in

---

T. Kalaichelvi (✉)

Artifical Intelligence and Data Science, Panimalar Institute of Technology, Chennai, India  
e-mail: [ai\\_dshod@pit.ac.in](mailto:ai_dshod@pit.ac.in)

L. Jabasheela · W. G. Theresa

Department of Computer Science and Engineering, Panimalar Engineering College, Chennai, India  
e-mail: [gracytheresaw@pit.ac.in](mailto:gracytheresaw@pit.ac.in)

P. S. Ramaprabha

Department of Electrical and Electronics Engineering, Panimalar Institute of Technology, Chennai, India  
e-mail: [ramaprabaps@pit.ac.in](mailto:ramaprabaps@pit.ac.in)

M. Shobana

Department of Computer Science Engineering, SRM Institute of Science and Technology, Kattankulathur, India  
e-mail: [shobanam@srmist.edu.in](mailto:shobanam@srmist.edu.in)

G. Dhanalakshmi

Department of Information Technology, Panimalar Institute of Technology, Chennai, India  
e-mail: [dhanalakshmig@pit.ac.in](mailto:dhanalakshmig@pit.ac.in)

H. Rashini

Department of Artificial Intelligence and Data Science, Panimalar Institute of Technology, Chennai, India

this case, a queue that completes their self-sorting process first can avoid collisions. The performance of the proposed protocol is compared to that of another common spectrum sensing in vehicular ad hoc networks (VANET). The results of the study and modelling in expressway and metropolis environments demonstrated that the new methodology can significantly decrease bandwidth usage and latency, especially in congested areas.

**Keywords** Vehicular ad hoc networks medium access control · Time-division multiple access · Sensing · Network performance

## 1 Introduction

The mobile ad hoc infrastructure would be a conscientious platform that intends to optimise transit quality and reliability by delivering a variety of security applications using vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) interactions. The primary reason for implementing vehicular communications is the safety component, which relies also on sending of wall posts providing acceleration, situation and so on. To avoid collisions, automobiles send status notifications per 100 ms. The spectrum sensing transmission signals are often used to find automobiles in an accident hazard or particular issue, as well as to deliver alarms when a collision or unexpected brake is detected [1–3].

In high-density environments, the fundamental problem in VANET is circuit obstruction in the MAC layer. There are many devices transmitting status signals in the routing path, causing major bandwidth consumption and increasing the time delay. A suitable maximum delay demand for time-critical routing protocols is 100 m/s. This one has been mandated that the packet delivery ratio (PDR) not be less than 90%. A proof-of-concept study reveals the areas of rising frequency on delays and PDR.

Concerns of intrusion detection were already widely investigated; therefore, numerous MAC protocols are already developed to assess VANET effectiveness. Because survival signals are predominantly transmitted, no response signal is sent after successful authentication. As a result, the sender is unable to modify the contention window because it is oblivious from any bandwidth utilisation. The authors describe a technique for dynamic available bandwidth and packet size adaptation based on projected local road capacity and impact rates. The authors study the potential of carrier sense multiple access (CSMA)-based transmission channels in VANET using stochastic geometry methods. The theoretical effectiveness of such distributed coordination function (DCF) network layer in IEEE 802.11p, that is predicated on CSMA, is examined, and a retransmission approach is proposed to improve system reliability. However, more research finds that the restoration strategy is only effective in low-density scenarios and is not ideal for large density circumstances since the extra restoration accumulate by a faulty transport worsen full loaded.

## 2 Related Works

To improve VANET speed, a slew of constellation strategies have been developed. In a constellation architecture, the presumption MAC inside a group and the central argument IEEE 802.11 MAC among pattern automobiles are used. Every automobile has two antennas that operate on scientific organisations. The scientists propose a distributed multichannel and mobility-aware cluster-based MAC (DMMAC) system which allows automobiles to send status updates to the base station each 100 ms. As a compensation to assure connectivity among all parties present, the coverage area is lowered, which reduces the functionality of security applications. It includes a nomenclature and statistical comparison of the existent constellation techniques in VANET, as well as an explanation of the grouping situation's fundamental components [4–6]. The majority of constellation techniques try to keep the cluster structure intact over a long period of time by employing complicated control mechanisms, where performance may be affected by vehicle motion.

- The massive unstructured random access is organised by the self-sorting method. Unlike CSMA-based methods, workstations in almost the same queue can compete for communication links overall, reducing problems induced by huge memory controller.
- When compared to previous slotted approaches, the proposed MAC protocol lacks tight and universal synchronisation, as well as a set frame structure. Nodes that successfully build a queue during the conscience operation have had the ability to manage the route and send TDMA datagrams directly [7, 8].
- In contradiction to the stacked approaches previously outlined, awareness of the capacity constraints is not required from each frame. As a result, the self-sorting MAC protocol does not require the frequent transfer of slot allocation signals to sustain the time-based, implying that the inefficiency in the proposed protocol is significantly lower [9, 10]. Because the medium access encryption time is substantially smaller than the forecasting demand arrival rate, the control messages were split from both the digital signatures and the spatial precision is great.
- An access procedure of the proposed protocol is flexible. In substring cycles, vertices really aren't required to stay in almost the same position that may result in accidents due to vehicular movement. And after time transfer, the buffer is no longer existent within proposed protocol, and a current self-mechanism is launched when another station receives transmitted data to broadcast. The procedure is irrespective of its earlier propagation, exhibiting its movement resistance [11–13].

The soul-based sorting algorithm for cars with packets to send consists of three stages: consciousness, lane reserving and file transfer. There is no tight and universal synchronisation in the proposed protocol, nor is there a defined three-phase structure [14]. The transition from one phase to the next is all prompted by a unique situation. For example, if ranked matches duration reaches to create steam within standard parts, the line can rapidly begin stream allocation [15], and if the line properly fills

the network, users in the line will transmit the input signals in the sequence of entering on list. There is no need to preserve the architecture because the line will indeed be dynamically terminated [16, 17] as soon as the last member completes transmitting data.

#### A. Unique Sorting

Automobiles with a quasi-barrier seek to take part into and launch a peer procedure. When a queue is recognised while soul, the automobile can fight to enter it. Otherwise, the automobile may trigger an identity process and be a transient head [18–20] of queue (QH), also with potential to become a permanent QH.

#### B. Channel Reservation [21, 22]

If ranked matches time exceeds its minimum  $t$ , then bandwidth booking process starts immediately. In such cases, the channel reserve scheme is needed to limit clashes induced by hidden ports: when they are quenched inside the somebody else's channel access range complete the soul method simultaneously time frame and start wireless communication quickly; otherwise, datagram in image gradient might collide in the optical layer.

#### C. Transmission of Information

When the QH transmits the third allocation notification note, the data transmission will begin. First, the QH shall directly transfer bundle that includes relevant details as well as the buffer node order. This is used to alert queue networks [23–25] that didn't receive the Arp request during the soul procedure. All stations in the stack transmit its wireless signals progressively from its head to tail in TDMA. Each channel is again reopened, and stations with packets to deliver start a new peer technique and rerun the previous procedure. Unlike established notched mobility models [26, 27], just participant for collision avoidance is conveyed even during communication session, and thus no additional amount static routes, like the redistribution knowledge of each hole in the aforementioned adventure and TDMA-based protocols, will be included in the packets.

### 3 Proposed Methodology

A dynamic multimedia and movement constellation MAC (DMMAC) feature supports automobiles to deliver revised event logs to a network per 100 ms. As a concession to assure interaction among parties present, the coverage area is lowered, which reduces the effectiveness of collision avoidance. The main challenge in VANET in high-density environments is circuit bottleneck in the application layer. Many devices emit status signals in the transmission range, producing severe packet collisions and increasing time delay (Table 1).

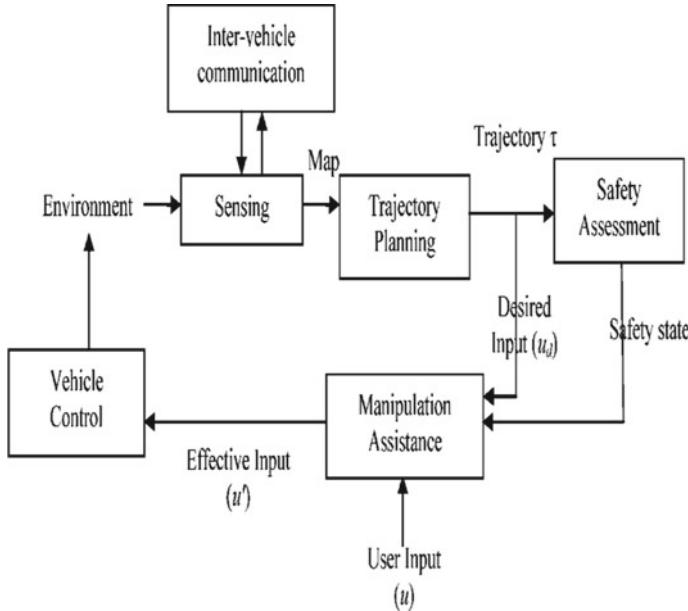
Figure 1 depicts the edges and vectors of city blocks in a sample infrastructure. The edges are considered as destination spots and the vehicles travel based on the requested targets as per the data communication protocol standards.

**Table 1** Value of parameters of simulation

Parameter	Value
Frequency ( $f$ )	0.1–0.4 automobiles/m
Velocity in highway	80–120 km/h
Velocity in city	50–60 km/h
Highway length	1 km
City street length	1 km
Width of junction area	100 m
Packet dimensions	200 Bytes
The frequency range	24 Mbps
The range of data transmission ( $R$ )	300 m
Transmission range with self-sorting ( $R_l$ )	100 m
Slot ( $\sigma$ )	10 $\mu$ s
Contention window size ( $W$ )	10
Arrival rate of packets ( $\lambda$ )	10 pkts/s
The likelihood of becoming a QH ( $P_1$ )	0.5
Length of queue ( $t$ )	5
Maximum attempt number ( $m$ )	40

**Fig. 1** Scenario based on city blocks [28]

The proposed architecture is shown in Fig. 2 the working flow of MAC-based sorting. The first step is to sense the data, and the second the trajectory location is assessed with safety state. The effective control is given from the manipulation assistant based on MAC protocol.



**Fig. 2** Artefacts architecture of proposed system

## 4 Analytical Model

### A. Model of the System

They compare the results of the designed MAC protocol to that of existing slotted topologies of mean latency and PDR, as well as its superiority in overhead.

### B. Queuing Success Probability

Throughout this part, we will design a Markov chain to determine the probability of successfully establishing a buffer. The self-sorting technique employs a restricted transmission range RI to eliminate neighbouring buffer duplication, and the carrier frequency of the proclamation packet is 2RI. If a required to transmit a passenger, it will briefly get to be a QH with the likelihood of P1 by delivering three brief QH designation signals in the region [2RI, 2RI].

### C. Service Possibility

Throughout this section, researchers measure the total gallery wall for transmitting that a network can acquire after the autonomy and lane allocation operations. If a node is in a buffer that has properly constructed a buffer of length t, this can send datagrams during the transmitted data. Instead, it will have to contend also for offering chance in later stages.

#### D. Implementation Overhead

The operational complexity of the personality protocol proposed in this study is analysed and compared to the conventional stacked leach protocol: TDMA-based and ALOHA-based procedures that are most significant. To ensure the impartiality and objectivity of the analysis, these following principles have been developed:

- i. This provides good are the identical, except perhaps the technical cost defined by various protocols.
- ii To compare the overhead performance of various protocols, we devised a statistic known as implementation efficiency (IE). The IE is calculated by dividing the quantity of a packet's payload by an overall number of bits necessary to deliver the transmission.

## 5 Performance Evaluation

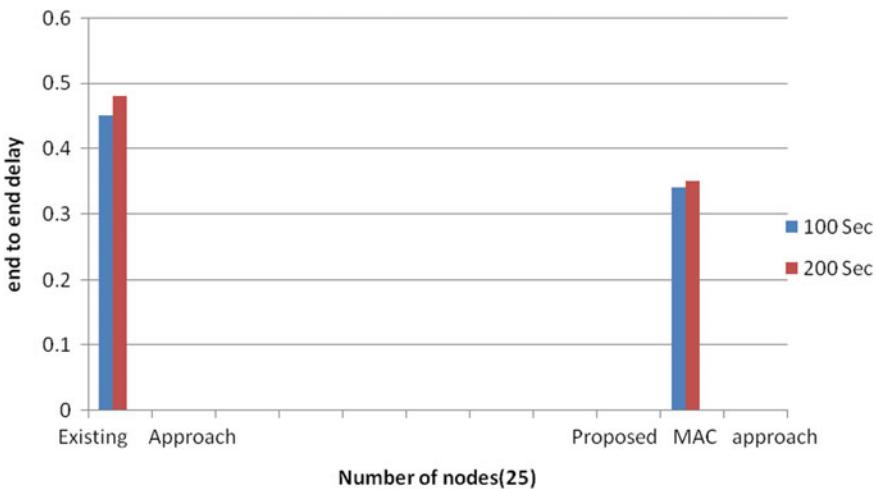
We perform a number of analyses and simulations to assess the proposed protocol in latency and PDR, and the results are reported in this section. The models make use of freeway and urban environments. The results of the system protocol are examined using various parameter values.

The freeway category is based on a one-way freeway segment at vehicle speeds ranging from 80 to 120 km/h, which again is standard for highways. The city scenario consists of a simple street, a perpendicular street, and four squares. The intersection of two roads is referred to as a junction area. Vehicles approaching the intersection will have an equal opportunity to choose any of the possible directions. Vehicles at the intersection [29] can interact with automobiles on both highways that are within transmission range. Due to the presence of cities, a vehicle that is not in the detection zone can only communicate with automobiles on the very same street and are within coverage area.

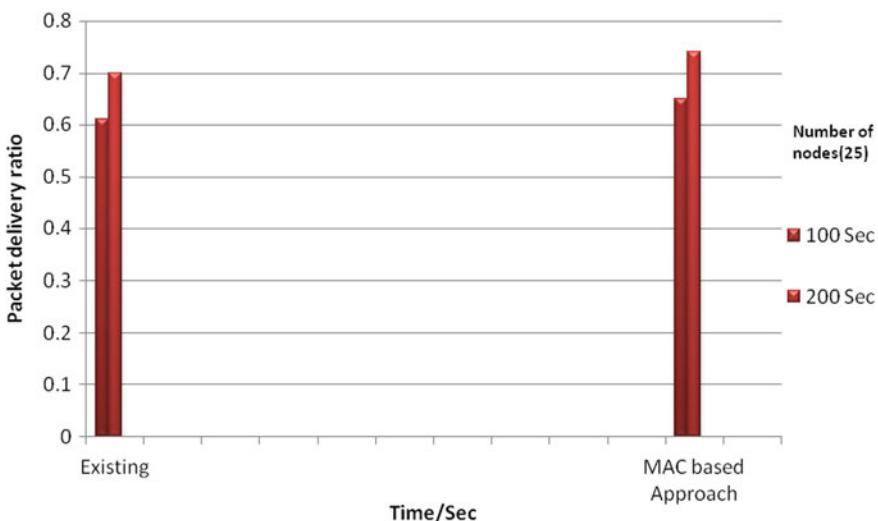
Nevertheless, the proposed protocol's performance will degrade in hyper environments. Collisions become more dangerous as a result of this during the peer process in hyper-conditions. Furthermore, if a line effectively fills the route, the computation time in the line capable of transmitting information packets stays unchanged, whether in moderate or medium-density scenarios, suggesting therefore routers will take more time to secure a position in a queue. As a result, our suggested protocol's condition increases at first and then deteriorates with density. Despite its packet loss in severely congested environments, the ego protocol outperforms other MAC approaches in VANET.

Evaluate the results of the ego approach on highways and in cities. At the same vehicle density, the efficiency of the control signals in the interstate situation outperforms the efficiency in providing a clear direction. Because of city block barrier, cars not positioned at the fusion zone could not get route occupied notifications from several other routes in a city environment. If two lines are on opposing sides of a street and are both within transmission range of an automobile in the edge switch,

transmissions from such two queues will clash if they broadcast at almost the same time. Crashes have a negative effect on productivity in a city setting. Nevertheless, in both cases, the suggested methodology exceeds the competition. Figure 3 depicts the end-to-end delay, and Fig. 4 implicates the efficient packet delivery ratio with respect to proposed system.



**Fig. 3** End-to-end delay



**Fig. 4** Packet delivery ratio

## 6 Conclusion

Throughout this study, an innovative simulative MAC protocol is created that makes use of elevated scenario characteristics to boost the effectiveness of elevated VANET, as the primary challenge with VANET is meeting needs in dense circumstances. Automobiles stand in line by ego in the manner of a "counting off", and indeed the line that reaches the specified length is granted access to the content. To significantly reduce communication conflicts, controlling signals for ego are intended to be crash. The findings of the study and simulation evaluate the protocol's achievement of latency and PDR. Furthermore, the latency is smaller than that of standard stacked mobility models that have a specific format and necessitate assignment information communication. The future efforts will be on customising the protocol to various circumstances, such as more structural response minimisation and tolerance for heterogeneous latency in cognitive radio networks.

## References

1. UFC Commission et al.: R&ofcc 03-324. Dedicated Short Range Communications Report and Order (2003)
2. Mak, T.K., Laberteaux, K.P., Sengupta, R.: A multi-channel vanet providing concurrent safety and commercial services. In: Proceedings of the 2nd ACM International Workshop on Vehicular Ad Hoc Networks, pp. 1–9. ACM (2005)
3. Eichler, S.: Performance evaluation of the IEEE 802.11 p wave communication standard. In: 2007 IEEE 66th Vehicular Technology Conference, pp. 2199–2203. IEEE (2007)
4. Graffling, S., Mähönen, P., Riihijärvi, J.: Performance evaluation of IEEE 1609 wave and IEEE 802.11 p for vehicular communications. In: 2010 Second International Conference on Ubiquitous and Future Networks (ICUFN), pp. 344–348. IEEE (2010)
5. A Intl: Standard specification for telecommunications and information exchange between road-side and vehicle systems-5 Ghz band dedicated short range communications (DSRC). Medium Access Control and Physical Layer Specifications, pp. E2213–03 (2003)
6. Hafeez, K.A., Zhao, L., Ma, B., Mark, J.W.: Performance analysis and enhancement of the DSRC for vanet's safety applications. *IEEE Trans. Veh. Technol.* **62**(7), 3069–3083 (2013)
7. Khabazian, M., A'issa, S., Mehmet-Ali, M.: Performance modeling of safety messages broadcast in vehicular ad hoc networks. *IEEE Trans. Intell. Transp. Syst.* **14**(1), 380–387 (2013)
8. Rawat, D.B., Popescu, D.C., Yan, G., Olariu, S.: Enhancing vanet performance by joint adaptation of transmission power and contention window size. *IEEE Trans. Parallel Distrib. Syst.* **22**(9), 1528–1535 (2011)
9. Nguyen, T.V., Baccelli, F., Zhu, K., Subramanian, S., Wu, X.: A performance analysis of csma based broadcast protocol in vanets. In: INFOCOM, 2013 Proceedings IEEE, pp. 2805–2813. IEEE (2013)
10. Hassan, M.I., Vu, H.L., Sakurai, T.: Performance analysis of the IEEE 802.11 MAC protocol for DSRC safety applications. *IEEE Trans. Veh. Technol.* **60**(8), 3882–3896 (2011)
11. IW Group et al.: IEEE standard for information technology–telecommunications and information exchange between systems–local and metropolitan area networks–specific requirements–part 11: Wireless lan medium access control (mac) and physical layer (phy) specifications amendment 6: Wireless access in vehicular environments, vol. 802, p. 11. IEEE (2010)

12. Su, H., Zhang, X.: Clustering-based multichannel mac protocols for qosprovisionings over vehicular ad hoc networks. *IEEE Trans. Veh. Technol.* **56**(6), 3309–3323 (2007)
13. Hafeez, K.A., Zhao, L., Mark, J.W., Shen, X., Niu, Z.: Distributed multichannel and mobility-aware cluster-based mac protocol for vehicular ad hoc networks. *IEEE Trans. Veh. Technol.* **62**(8), 3886–3902 (2013)
14. Yao, Y., Zhou, X., Zhang, K.: Density-aware rate adaptation for vehicle safety communications in the highway environment. *IEEE Commun. Lett.* **18**(7), 1167–1170 (2014)
15. Cooper, C., Franklin, D., Ros, M., Safaei, F., Abolhasan, M.: A comparative survey of vanet clustering techniques. *IEEE Commun. Surv. Tutorials* (2016)
16. Borgonovo, F., Capone, A., Cesana, M., Fratta, L.: Rr-aloha, a reliable r-aloha broadcast channel for ad-hoc inter-vehicle communication networks. In: Proceedings of Med-Hoc-Net, vol. 2002 (2002)
17. Scopigno, R., Cozzetti, H.A.: Mobile slotted aloha for vanets. In: IEEE 70th Vehicular Technology Conference Fall (VTC 2009-Fall), pp. 1–5. IEEE (2009)
18. Han, F., Miyamoto, D., Wakahara, Y.: Rrob: a TDMA-based MAC protocol to achieve high reliability of one-hop broadcast in vanet. In: 2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops), pp. 87–92. IEEE (2015)
19. Haddad, M., Muhlethaler, P., Laouiti, A., Zagrouba, R., Saidane, L.A.: TDMA-based MAC protocols for vehicular ad hoc networks: a survey, qualitative analysis, and open research issues. *IEEE Commun. Surv. Tutorials* **17**(4), 2461–2492 (2015)
20. Omar, H.A., Zhuang, W., Li, L.: Vemac: a TDMA-based MAC protocol for reliable broadcast in vanets. *IEEE Trans. Mob. Comput.* **12**(9), 1724–1736 (2013)
21. Mugunthan, S.R.: Wireless rechargeable sensor network fault modeling and stability analysis. *J. Soft Comput. Paradigm (JSCP)* **3**(1), 47–54 (2021)
22. Chen, J.I.Z., Yeh, L.T.: Data forwarding in wireless body area networks. *J. Electron.* **2**(2), 80–87 (2020)
23. Jiang, X., Du, D.H.: Ptmac: a prediction-based TDMA MAC protocol for reducing packet collisions in vanet. *IEEE Trans. Veh. Technol.* **65**(11), 9209–9223 (2016)
24. Zhang, R., Cheng, X., Yang, L., Shen, X., Jiao, B.: A novel centralized TDMA-based scheduling protocol for vehicular networks. *IEEE Trans. Intell. Transp. Syst.* **16**(1), 411–416 (2015)
25. Yao, Y., Rao, L., Liu, X.: Performance and reliability analysis of IEEE 802.11 p safety communication in a highway environment. *IEEE Trans. Veh. Technol.* **62**(9), 4198–4212 (2013)
26. Suma, V.: Automatic spotting of sceptical activity with visualization using elastic cluster for network traffic in educational campus. *J. Ubiquit. Comput. Commun. Technol.* **2**, 88–97 (2020)
27. Dhaya, R., Kanthavel, R.: Bus-based VANET using ACO multipath routing algorithm. *J. Trends Comput. Sci. Smart Technol. (TCSST)* **3**(1), 40–48 (2021)
28. Balaji, S., Sasilatha, T.: Detection of denial of service attacks by domination graph application in wireless sensor networks. *Cluster Comput. J. Netw. Softw. Tools Appl.* **22**(6), 15121–15126 (2019). <https://doi.org/10.1007/s10586-018-2504-5>
29. Chen, J.I.Z., Hengjinda, P.: Enhanced dragonfly algorithm based K-medoid clustering model for VANET. *J. ISMAC* **3**(1), 50–59 (2021)

# A Study on Challenges in Data Security During Data Transformation



K. Devaki and L. Leena Jenifer

**Abstract** Big data analytics has become an essential study due to its significant technological advancement in data storage, processing capability, analytics methods, and its applications. Nowadays, handling massive data and securing valuable data from loss and leakage have become a challenging task. Data is also stored and processed as semi-structured and unstructured with high volume and velocity. High programming overheads in coarse-grained nature of scientific workflow cause internal data loss. Therefore, data security is implemented to safeguard the information during data transformation. Data security helps in protecting customer data, which maintains privacy and prevents data loss. Unauthorized user access, abnormal operations, leakage, inaccuracy, and loss during data transformation cause loss in internal information, which leads to incomplete data dependencies and differentiation problems. The need for data security supervision is necessary due to the increase in volume and velocity of data. This paper reveals the study related to big data characteristics and security challenges faced during data transformation.

**Keywords** Big data security · Data transformation · Data leakage · Anomaly detection · Data accuracy

## 1 Introduction

Big data is a field that deals with methods for analyzing, methodically extracting information from, or otherwise dealing with data volumes that are too large or complicated for typical data-processing application software to handle. Data capture,

---

K. Devaki

Professor, Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai, India

e-mail: [devaki.k@rajalakshmi.edu.in](mailto:devaki.k@rajalakshmi.edu.in)

L. Leena Jenifer (✉)

Research Scholar, Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai, India

e-mail: [leenajeniferc87@gmail.com](mailto:leenajeniferc87@gmail.com)

storage, analysis, search, sharing, transfer, visualization, querying, updating, information privacy, and data source are all issues in big data analysis. The ability to show patterns, trends, and relationships, particularly those affecting individuals and enterprises, is one of the most critical expectations from big data analytics, with the purpose of guiding meaningful decisions [1]. The huge volume of data produced by various fields such as engineering, social networking, health care, and business is highly volatile. Several challenging problems like data storage, organization and representation, processing and security prevail in big data technological field. Its characteristics play a vital role in storage and manipulation process where the quality of data has to be maintained by imposing security. Previously, the characterization of big data insisted on 3Vs (volume, velocity, and variety), but now it is developed into 10Vs (additionally, veracity, validity, value, visualization, volatility, vulnerability, and variability). Big data also deals with vast collections of datasets and tedious processing issues during data transformations.

Organizations can use data transformation to change the structure and format of raw data as needed. The process of modifying the format, structure, or values of data is known as data transformation. The process of extracting solid, trustworthy data from different sources is known as data transformation. This entails transforming data from one structure (or none) to another in order to link it with a data warehouse or other applications. It enables you to use modern business intelligence tools to build useful performance reports and estimate future trends using the data. Data transformation can be used in a variety of processes, including data integration, data migration, data warehousing, and data wrangling. The unsolved problems in the existing methodologies pose a significant impediment. The existing methodologies in the fields of data security, health care, cloud computing, and IoT experience the following as limitations, such as network jamming, tampering, unfairness, collisions, misdirection, flooding, reprogram, overwhelm, during data transformation. In IoT systems, wireless connections such as Bluetooth, Wi-Fi, and other server modules are available for transferring measured data. For analytical purposes, data acquired by IoT sensors can be embedded on any type of material. Data mining methods based on neural network architectures have been developed that can analyze such readings in a cloud architecture [2]. It is vital to preserve the accuracy of big data to avoid data fabrication. The development in information processing methods resulted in data security issues. Due to the expansion of information processing methods, some of the big data analytics challenges are identified (i) Generation of duplicate data, (ii) Lack of security focus in databases, (iii) Sharing of confidential information, (iv) Lack of access control mechanism, (v) Challenges of data provenance, (vi) Dilemmas of cryptographic algorithms, etc.

These challenges also create problems in decision-making strategies. To overcome the difficulties in decision-making process, machine learning algorithms and tangible queries are used to analyze the key elements in big data. Data is prone to be modified and leaked by unauthorized users, and as a consequence, there may be a loss in original data. The most difficult part of the transformation is to determine the quality and trustworthiness of data. The characteristics of big data (10Vs) are associated with large-scale cloud infrastructures, where the traditional security

mechanisms are not enough to handle corrupted data in complex distributed environments. The main characteristics of big data are referred to as 10Vs—volume, velocity, variety, veracity, value, visualization, volatility, validity, vulnerability, and variability [3]. Huge volume of data storage and velocity of processing provides a way to unauthorized user modifications. Uncertainty perspectives can be caused by a multitude of factors, including veracity (the ability to measure truth), variability, and variety. Data validity, variability, and volatility are the major issues identified during interoperability challenges. Value and vulnerability issues lead to data leakage and loss in data security. Data visualization and vulnerability can lead to accuracy issues. Thus, the characterization of big data defines its impact on the challenges of data security during storage, processing, transformation, and manipulation. Thus, the 10Vs identified as big data dimensions contribute as the key factors that have an impact on data security issues.

The challenges of big data security are discussed in Sect. 2. The research need is highlighted in Sect. 3, whereas Sect. 4 outlines the proposed line of research, and Sect. 5 states the conclusion and its direction for future work.

## 2 Challenges of Big Data Security

Big data consists of a large amount of personal identity information used by the banking industry, IT companies, social media networks, health care, and other industries. Whenever a huge amount of information is generated, a proper balance between the utility of the data and its privacy should be maintained. Data duplication and anonymous data occupy a huge volume of storage space, which leads to stability problems. Identification of vulnerabilities and risk assessments plays a crucial role in the fields of big data security. Data provenance plays a vital role in validity, authenticity, and integrity in reproducing the results consistently. Scientific workflow systems employ one of the approaches to establish provenance dependencies: (1) they depend on workflow computations to declare dependency relationships at runtime; (2) they execute implicit assumptions concerning dependency patterns from where they are derived; (3) they assert no dependency information at all; or (4) they infer dependency information automatically [4].

This study assesses data security issues such as uncertainty perspectives, unauthorized user modification, interoperability, data leakage, and data accuracy issues that arise when storing and transforming data using provenance information. A discussion of effective approaches for dealing with such challenges, as well as their limitation, is also recommended by the study. Figure 1 represents a visual illustration about the challenges of data security and the research problem identified.



**Fig. 1** Illustration of challenges in data security

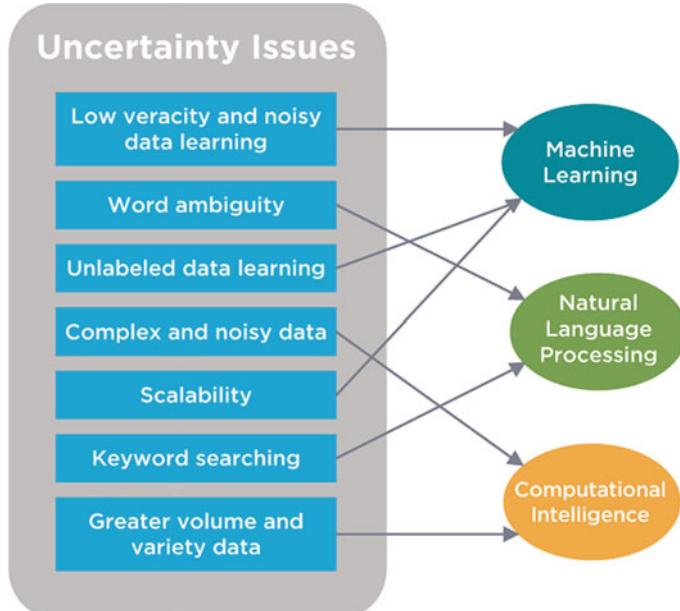
## 2.1 Uncertainty Perspective

Uncertainty in big data can refer to “doubt” or “suspicion” or “dilemma” prevailing in the data truthfulness. Something that lacks certainty and seems to be imperfect during analysis may lead to uncertainty. The issues related to veracity, variety, and variability may contribute to big data uncertainty perspectives. According to a survey conducted in 2018, the challenges emerging from uncertainty issues are increasing everyday because the amount of data produced each day is around 2.5 quintillion bytes [5]. Social networks such as Google process beyond 40,000 searches every second and Facebook account holders upload nearly 300 million snaps and share nearly 51,000 comments every single day. Uncertainty accounts for inaccurate and missing data due to noise and incompleteness [6]. Uncertainty has been noted as one of the primary challenges identified by Wang et al. [7], which involve unknown or imperfect information. The author focuses on the veracity characteristics to determine the impacts based on uncertainty learning performance. Uncertainty problems can be addressed using the following types, such as Shannon entropy, classification entropy, fuzziness, non-specificity, and rough degree [6, 7]. Fuzziness-based semi-supervised learning and ambiguity-based model tree (AMT) are two researches that have been introduced to learn big data uncertainty, in addition to the above-mentioned types. Uncertainty focuses on the veracity features which indicate nearly 80% of the data remains missing during the transformation process, and they are possibly tackled

using some of the above-mentioned techniques defined with mathematical definitions [7].

Many distinct types of uncertainty exist in big data analytics, all of which can have a negative impact on the effectiveness and accuracy of outcomes. The accuracy and trustworthiness of analysis get affected if uncertainty prevails in data, due to oversight in social sensing networks. Chao Huang et al. discussed solving the truth discovery problem in social sensing applications using the Scalable Uncertainty-Aware Truth Discovery (SUTD) scheme which is compared to state-of-the-art solutions as a result of the evaluation. Real-world datasets gathered from Twitter in 2015 and used for demonstration. True claims and undecided claims rubrics are used for evaluating the truth table. It also addresses the problem related to big data variability issues and provides a better suggestion for handling real-time datasets in social networks [5]. Uncertainty can also be measured as (i) inaccurate or incomplete data and (ii) ambiguous data. In social sensing, the uncertainty of reported records and their scalability contribute to address veracity and variability problems [6]. Advanced analytical techniques are used for predicting high precision course of action and in decision-making strategies. Artificial intelligence techniques such as machine learning (ML), natural language processing (NLP), and computational intelligence (CI) [6, 7] are used to address various uncertainty issues as shown in Fig. 2.

The data obtained from wearable devices and social media [5, 6] leads to veracity and a variety of data accumulation and data conflicts, which makes managing uncertainty issues in health care a difficult task again. Uncertainty embedded in data



**Fig. 2** Classification of data uncertainty handling strategies

processing has a major impact on learning performance and data stored on the web. TRUTHFINDER algorithm was invented by Yin et al. [3] to find the true facts among conflicting information. The TRUTHFINDER algorithm is used to solve the inter-dependency between trustworthiness and identifying secure websites. The author has used real datasets such as book authors and movie runtime to categorize the conflicting information and true information stored on websites. TRUTHFINDER solemnly addresses the veracity issues and contributes toward overcoming the uncertainty challenges in data security. Sources are assumed to be independent in the SUTD scheme and dependency exists when connected through social networks. Correlations are not assumed between the claims [5]. The Internet of things (IoT) [8, 9] uses a global positioning system, radio frequency identification technology, sensors, and a variety of other devices to collect data. IoT technologies are used in a variety of sectors, including environmental monitoring, recommendation systems, forest protection, smart transportation, health care, and smart cities. Hyperspectral sensors are used in the earth observation area to image specific regions utilizing hyperspectral remote sensing technologies. Clustering [9] is a frequently utilized and well-studied technique in unsupervised classification.

A huge amount of un-factual interpretations, fabrication of data, and spam are noticed in social events, which cause a high risk of uncertainty. Strong social dependency and rumor spreading in social events are high compared to physical event sources. Complications in dynamic fluctuations of uncertainty information are not well understood or justified. When the data sizes diverge, aggregation of multidimensional uncertainty information fails. Using more comprehensive methodologies, it is possible to improve natural language processing and text mining. More scalable and accurate social sensing applications need to be developed to address big data uncertainty issues. Thus, the information stored on web often provides conflicting, duplicate, and doubtful information about the same product with different specifications. These cause veracity, variability, and variety issues, which lead to uncertainty challenges.

## 2.2 Unauthorized User Modification

The term anomaly refers to the existence of an unusual data or abnormality. The unusual or abnormal behaviors observed in the data streams are called anomaly detection. Anomaly detection remains as one of the vast areas to be addressed in data security challenges. Anomaly detection refers to the identification of unexpected behavior patterns in data. Anomalies operations can be categorized as data fabrication, sensitive data deletion, and anonymous data insertion during data transformation, which leads to incomplete or inaccurate data dependency. Anomalies operations are detected in the distributed and parallel processing systems. A computation provenance system was developed to capture provenance data with MapReduce in Hadoop. Provenance data [10] is a lineage or pedigree which describes the origin of data and the process by which it arrived in its current state. The anomalies in tampered and

forged data are recovered using a set of variants. Anomaly detection done using this system remains limited to handling compromised computation or user data only [11]. Due to the large accumulation of data, anomaly detection has become a challenging task in the field of data security. High volume may lead to high dimensionality of data. Traditional methods [12] as such as cluster-based technique, distance-based technique, density-based technique, and classification-based technique are used to handle dimensionality problems and process a huge amount of high-dimensional data in a manageable time span. By expressing a triangular model of vertices for data dimensionality, Thudumu et al. [12] have addressed the constraints of classic anomaly detection algorithms. The model discusses the taxonomy for anomaly identification in high-dimensional large data, problem, algorithms and tools used to overcome the challenges of anomaly detection. The anomaly detection from logs was discovered using LogLens [13], an unsupervised machine learning approach. LogLens was a real-time log analysis system which was used to diagnose the root cause of complex problems. It automates anomaly detection processes from logs without target system knowledge and user specification. On power grids, anonymous variables and trash collectors, also known as the “curse of dimensionality,” are accessible. A random matrix theory-based algorithm was used to detect anomaly asymptotic empirical spectral distribution for spatial-temporal datasets of power consumption. It has limitations over speed, sensitivity, and reliability in practice [14].

Context anomalies are data points that are discovered as aberrant when compared to metadata. Big Data Provenance Model (BDPM) introduced by Gao et al. was an extension of the provenance model in order to address data security issues based on context anomalies. Layers of data are used to express provenance information. The dependencies between entities, agents, and activities are identified by generating a dependency log. To find abnormalities, the author used coexistence and inference rules. The provenance graph is generated based on a real-time log to detect the presence of abnormal operations [15]. Security is a critical concern for cloud users as they deal with vast amounts of data. The proposed method [16] combines finger-print as a biometric trait with an N-stage Arnold transform to verify the legitimacy of the user. Due to the limitations of the former methods, many people prefer to use biometric methods to authenticate and manage their accounts. In this paper, a method that overcomes these disadvantages by performing various phases of user enrollment, user verification, and processing and verification scheme was proposed. The complexity of the system and its template matching techniques have been used to minimize the chances of unauthorized entry. A novel based-arc hidden semi-Markov model and state summarization were utilized to derive behavior usage from raw sequences. This model is proposed for detecting anomalies accurately in cloud environments by observing the usage behaviors created by servers, i.e., system call identifier sequences. The representation of these behaviors is related to volume, velocity, and variety of characteristics [17].

Deep learning is a technique that uses neural network to learn and recognize images with normal and abnormal conditions. It achieves a good accuracy after being trained by classifiers. In order to distinguish visual activities/events, the raw input data can be retrieved using a single classifier CNN. For training and classification

purposes, the proposed method has combined two algorithms such as CNN and SVM. Many motion characteristics have been included to the frame in order to incorporate strong graphical features and improve alliteration accuracy. This is the flaws in the present framework proposal [18]. Different types of deep anomaly detection methods [19] are used to address the following challenges such as CH 1: Achievement of high anomaly detection recall rate is addressed, CH 2: Anomaly detection in independent and high-dimensional data is addressed, CH 3: Efficient learning of normal or abnormal data is done, CH 4: Noise-resilient anomaly detection is addressed, CH 5: Complex anomalies detection is done, CH 6: The problem of deriving anomalous explanations from specific detection methods remains largely unsolved is addressed. Table 1 summarizes the methods used for detecting the unauthorized user access and the challenges addressed by them.

The anomaly detection procedure becomes more sophisticated as the volume of streaming data grows and cannot be stored statically. Anomalous data points are not identified among a huge volume of data with high-dimensionality issues. A better balance of performance and accuracy in anomaly detection is needed [12]. It is necessary to refine algorithms based on the capture of temporal and spatial properties of anomalies for examination. To overcome the incomplete data dependency problem,

**Table 1** Different types of deep anomaly detection methods

Learning types	Methods	Challenges addressed
Anomaly measure dependent learning	Auto encoder	• CH 1, CH 2, CH 4, CH 5
	Generative adversarial networks	• CH 1, CH 2
	Predictability modeling	• CH 1, CH 2, CH 5
	Self-supervised classification	• CH 1, CH 2, CH 4
Generic normality feature learning	Distance-based measures	• CH 1, CH 2, CH 3, CH 4
	One class classification measures	• CH 1, CH 2
	Clustering-based measures	• CH 1, CH 2, CH 4
Various end to end anomaly score learning	Ranking models	• CH 1, CH 2, CH 3, CH 4, CH 6
	Prior-driven models	• CH 1, CH 2, CH 3, CH 4
	Softmax models	• CH 1, CH 2, CH 5
	End to end one class classification	• CH 1, CH 2

advanced strategies must be provided to address features such as velocity, volume, and a variety aspect.

## 2.3 *Interoperability*

Interoperability refers to a system's and services' capacity to interact, exchange, share, and consume resources or data. Due to the lack of compatibility between diverse platforms, more obstacles are linked to data exchange and repeatability. Validity, variability, and volatility characteristics of big data are addressed during sharing, exchange, and consumption of resources. Data stored on the web remains inconsistent due to interoperable processing. The PROV set model was designed in 2013 by Missier [20] to promote the spread of information on the web which instantiates interoperability among different management systems, provenance producers, and consumers that address the validity and variability characteristics. The PROB tool was used to track provenance, which is based on the volatility characteristics, allowing researchers to interact by exchanging information without sharing the real dataset. Interoperability plays a vital role in the field of health care where clinical information is exchanged across many divisions, which leads to variability and validity issues. The Semantic Interoperability Model for Big Information in IoT [SIM-IOT] was introduced to achieve interoperability between data generated by various information systems in healthcare industry. Interoperability across heterogeneous datasets is done by a core ontology constructed, where the clinical information is securely used, and modification of data is restricted during data transformation to ensure validity of data [21]. It is a challenging task to find websites with useful and compatible data. To overcome this issue, a web-based prototype was implemented by Cheung et al. [22] which allows interoperability among disparate types of yeast genome data in tabular and RDF formats.

The emerging data economy, where systems are able to transform business and enhance human experience, needs direct collaboration to perform interoperability. CWL Prov (Common workflow Language) is an interoperable workflow and structured representation model using the W3C Prov defined by Farahzaib Khan et al. CWL Prov enables portability, interoperability, and reproducibility among the workflow platforms. Computing interoperability is mainly focused on syntactic, semantic, and programmatic information [23]. Information can contain numerous datasets, where data is required to be grouped and merged based on some common fields. The PROV data model was developed by W3C to extend interoperable representation and explore information on the web. It encapsulates key characteristics of provenance that enable the integration of agents, entities, and activities, while also identifying their interdependence. The CWL tool enables sharing of computational analysis, which provides preservation of data and methods through interoperability [20]. To overcome the difficulties of integrating data from disparate sources, storing data in suitable repositories on the web, some of the above-mentioned tools and techniques can be used to promote interoperability with data security measures.

There is no established interoperability over huge sensing data streams in the existing system. To address this problem, the efficiency of symmetric key encryption can be enhanced. During interoperability, the uniform resource locator (URL) may be unavailable, or the syntax may be changed as the data retrieval process is done on the web server [22]. To overcome this, schema representation needs to be performed. Resources shared during data transformations are sometimes rendered ineffective due to incomplete provenance data and limited access to data [23].

## 2.4 Data Leaks and Data Loss

Unauthorized sharing or disclosure of sensitive information to unauthorized recipients is known as data leakage. Data leakage is identified as a major issue in the fields of cloud computing, health care, and the banking sector. It is observed that nearly 73% of customer data, computer system, and network information are leaked, which accounts for the value and vulnerability issues. Big data and data science applications consider data leakage as an unlawful action. Physical data leakages like dumpster diving, shoulder surfing, and photocopies also happen during data access and transformation. Organization survival depends on securing sensitive data from falling into the wrong hands. Zhang et al. [24] proposed a rule-based data provenance tracing algorithm to detect various data leakage threats in provenance data, which identifies file stealing, renaming, and file movements in cloud directories. It also detects files sent and received across disparate machines in the identical cloud and email client file leakages. It fails to detect the overridden files during data integrity and transformation, which leads to vulnerability issues. Park et al. [25] suggested an access evaluation algorithm to protect the provenance data, which is more sensitive, and a dependency-based policy is used which provides access control administration. Abstracted dependency names and matching dependency path patterns are employed by using regular expressions. The suggested architecture provides for extremely expressive policy expression and also allows for easy and effective access control administration. digital signature algorithm (DSA) was proposed by Bates et al. [26] to implement a file transfer application to block the derived files of transmission which prevents data leakages and scattering of sensitive data. The storage overhead incurred by automated provenance collection is a serious challenge. Linux provenance model (LPM), a framework for building trusted provenance-aware execution was demonstrated. LPM is used as the cornerstone of a provenance-based data loss prevention system that can scan file transmissions in tenths of a second to detect the presence of sensitive lineages. LPM does not address the issue of provenance confidentiality. It is, nonetheless, an important challenge that has been considered.

Suen et al. [27] used S2Logger, a data event logging mechanism, to capture, analyze, and visualize data events in a cloud environment to avoid data leakages. It detects data leakages and policy violations by analyzing the data provenance. Non-critical paths in the data transfer graph can be pruned by using simplified approaches. Data leakage restrictions are enforced at control points after data event notifications

from the corresponding hosts are received. Alabi et al. [28] had used Hadoop MapReduce algorithm to develop a framework for gathering provenance data and examining data spillage within the Hadoop cluster. It detects the system vulnerability related to data leakage events within Hadoop systems. It is not easy to collect data provenance in distributed or parallel processing system that addresses system security and data accountability. To collect information in a synchronized manner, care must be taken to determine the extraction points within a distributed system; maintain the authenticity and integrity of provenance logs; and develop visualization tools to help analysts understand system security threat levels at glance and detect when problem arise.

Alneyadi et al. [29] have proposed Data Leakage Prevention Systems (DLPS) to detect and prevent confidential data leakages which are in use, in transit, and at rest. In the world of information security, data leaking is a persistent issue. Academics and practitioners are constantly attempting to develop data leakage prevention and detection technologies to address this issue. DLPSs are becoming more widely recognized as preferred solutions for locating, monitoring, and safeguarding personal information. Furthermore, the majority of current approaches have significant flaws, particularly when dealing with personal data that is constantly changing. This is due to the fact that they rely on rigid approaches. Zhang et al. [30] introduced lossy trapdoor functions (LTFs) to solve the sensitive trap door leakage problem in the LTF system. A secure application deployment in sensitive data-revealing environments is used, in which a side-channel analyzer monitors the secret channel, watches the private memory, and detects the algorithm operating to extract some sensitive information. Shi et al. [31] introduced matrix factorization to predict the missing data in the time series from multiple sources. It is used to find the missing data in multivariable time series by engaging an improved matrix factorization technique. To improve the accuracy of missing data prediction in multivariable time series, unique strategies to constrain the matrix factorization are done by merging both the temporal smoothness of each time series data and information from many sources. The proposed technique focuses primarily on incorporating data from social networks, which varies from sensor networks significantly.

Information saved in databases can be accessed by unauthorized people through various data leakage channels like tablets, smartphones, emails, and social media, which are used during data exchange and transformation. The assets of the organization are to protect their own sensitive data. DLPs are used to protect sensitive data and are classified according to their approaches rather than their application. DLPs [29] do not provide classification of confidential data semantically. Privileged malicious insiders (humans) are considered as the most harmful threat to leaking confidential data. Data mining tools and big data technologies like Hadoop can be used to store and to strengthen the security of data, which prevents data leakage. Provenance data is captured, and access control mechanisms and provenance graph analysis are used to minimize unauthorized user modifications and misuse of data by social network platforms. Vulnerability in enormous data must be avoided to increase the value of accountability in revenue at the right time. Hence, the revenue and reputation of the

organization were investing a significant number of resources and time-dependent on the preservation of sensitive data.

## 2.5 Data Accuracy

Error-free or quality data, which maintains consistency, is known as data accuracy. Maintaining data accuracy is a challenging task due to huge chunks of data being generated in this fast-paced digital world. To maintain data accuracy, quality goals must be established, processed, and data overloading should be prevented at particular phases. The common attributes of accuracy are identified as timeliness, completeness, consistency, validity, and uniqueness. To maintain the system data safe, it is necessary to check and update it on a regular basis. In order to determine the performance efficiency of a micro-grid solar power system, a big data mining technique was used. The created method was tested on a variety of datasets in order to determine its prediction accuracy and assault resistance for data transmission. The approach's drawback is related to network limitations; if communication is delayed, the possibility of protecting individuals at a vital time gets reduced which causes minor and severe problems in IoT architecture's connected systems [8].

During data transformation, a continuous stream of data from numerous sources is generated, and the accuracy of the data is violated, resulting in vulnerability issues. Salman Sultana et al. have proposed a novel approach known as the Spread Spectrum Watermarking-based solution. The provenance, which was the key factor in assessing the trustworthiness of data, was embedded into the interpacket delays to securely transmit the provenance data packets during data transmission. The confidentiality and visualization during transformation of data over the network were accomplished through watermarking technique. The approach has the limitation of handling large provenance data [32]. Ikbal Taleb et al. have considered accuracy as one of the intrinsic data quality dimensions which contextually exhibits reputation, accessibility, and relevance during data preprocessing [1]. A Quality of Big Data (QBD) model was proposed to support data quality profile selection and adaptation in the preprocessing phase. The limitation of this model was that it lacks quality rules diversity, i.e., a consensual agreement that defines the structure that establishes high-quality data, value, etc.

The access control mechanism was also a significant method used to maintain proper visualization and to avoid vulnerability in data. The layer-based provenance data architecture suggested by Rajeev Agarwal et al. [33] includes a security access control mechanism. A provenance data and visualization layer design that is simple to use has been described. The model's drawback is that it necessitates a benchmark for evaluating provenance information performance. Extension of Provenance-Based Access Control (PBAC) has been introduced by Park et al. [25] which handles the challenges of traditional access control mechanisms. A PROB toolkit was proposed to track the provenance of data in which raw data was collected, processed and preprocessed data was loaded into the cluster file system which produces the derived

data. These derived data are visualized, shared, and distributed to maintain the quality of the data [34]. Electronic control units (ECUs) are designed Smys et al. [35] to provide secure and limited access to specific users while still allowing third-party requests to be restricted. To ensure authenticity and validity, vehicles with a proper record will not be allowed to use the blockchain network to communicate with each other. This architecture outlines the various steps involved in a vehicle's lifecycle and evaluation against a similar database was performed.

Sun et al. proposed an Aware Access Control Framework with typed provenance which applies access control mechanisms to improve the accuracy of data. A layered architecture was built within the framework which comprises a typed provenance model (TPM) and a collection of TPM interpreters. The TPM bridges the gap between provenance questions and complex provenance graphs in solution space. A homework grading system was implemented with a proper access control mechanism to give proper access limitations to the students and mentors. It also defines the existence of primitive and composite dependency. The limitation of this framework was that it fails to explore and optimize the performance overhead generated by provenance query engines [36]. Hu et al. [37] proposed an improved NTRU cryptosystem to present a secure and verifiable access control scheme to protect the outsourced big data stored in the cloud. Data owners are allowed to update the access policy dynamically. The outsourced ciphertext was also updated by the cloud server, which enables efficient access control over the data stored in the cloud. A verification process to validate the user and the data owner is provided to recover the plain text. By establishing an access control mechanism, only authorized users are allowed to access, modify, and retrieve data.

The framework used is considered for data provenance captured only by the application system and not by the operating system. The access control method is developed using a typed provenance model (TPM) interpreter that is solely meant to work with RDF-based provenance stores. In real-time settings, performance overhead is not addressed [36]. Improved NTRU cryptosystem used to analyze the correctness and computational complexity of the threshold secret sharing with attribute-based access control needs to be developed to decrypt outsourced cipher data in the cloud [37]. Provenance security and time-based flow watermarking were used for secured transmission of data streams, and it fails to address scalability and data degradation issues during data transmission [32]. Thus, some of the above-proposed frameworks are designed to secure the accuracy of data by open visualization and reduction of vulnerability in data. As the data grows, the techniques for securing the accuracy of data are not adequate and improvisation of mechanisms is essential.

### 3 Research Gap

Big data analytics is a form of advanced technology which involves complex applications such as predicting market trends, future needs, and improving efficiencies in the company's supply chain. Data is prone to modifications and leakages during

data transformation by unauthorized users. Data provenance is one of the effective approaches used to overcome this issue. Data provenance is used during debugging data, data transformation, evaluating the quality, trust in data, and implementing access control for derived data. Vulnerability to fake data generation, data leakages, and data inaccuracy during transformation leads to the problem of incomplete data dependencies. Handling huge volumes of data might lead to un-factual interpretations, spam, and duplication of data. The uncertainty perspective may end up with data security risks like data masking, phishing, and accidental exposure of sensitive data during transformation. Data accessed through networks uses large stores of data which can easily learn the normal behavior of networks. Removal of anomalous data in supervised learning may result in significant improvement of accuracy. The presence of anomalous data points may lead to higher dimensionality issues. The high volume of growing data and the velocity of accessing the data lead to vulnerability and fabrication of data in an unstructured environment.

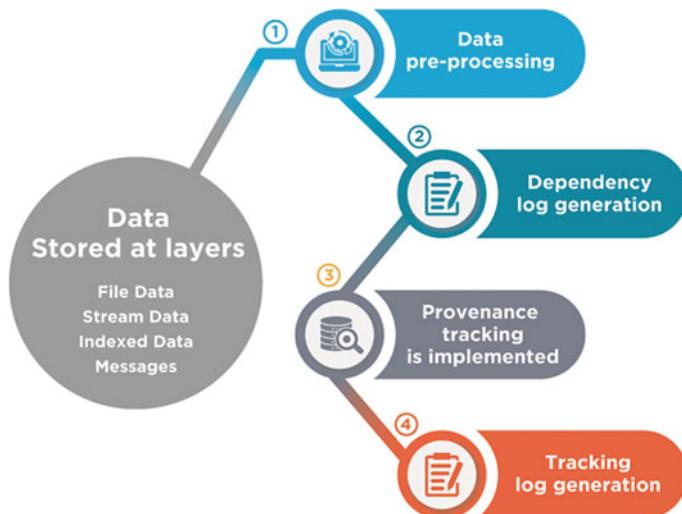
During interoperability, there is no automatic optimization and discovery of quality proposals which lack quality rules diversity. To refine and optimize the new investigations, an analytical model is required. Existing mechanisms are unable to find the overridden files when a variety of data is handled. It fails to detect the causes of data leakages, and provenance confidence is not addressed properly. Data leakage prevention systems are based on techniques rather than applications. Lossy trapdoor functions are used to address only limited and restricted information. Disclosure of sensitive and confidential information leads to data leakage issues which are more expensive to protect. Data breaches occur by intruders who are against the organization. They can exploit any loopholes and pinpoint weak security to gain access to private and government company secrets. All the techniques and methods discussed as existing systems have both advantages and limitations. Research challenges in big data analytics begin at the time of data creation, data storing, data processing, data manipulation, and data transformation due to its huge volume and high velocity. The adoption of implicit rules can often lead to incomplete and inaccurate data dependency during computation. It automatically affects the value characteristics of big data. Even though researchers have provided valuable techniques to handle provenance data, performance overheads issues during data transformation are not addressed.

Internal data loss happens due to high programming overheads and calculation processes in the coarse-grained nature of scientific workflow. All possible data dependencies are not maintained during data transformation. Unauthorized users and abnormal operations are identified during transformation. Data inaccuracy and data loss may occur as a result of data leakage, resulting in incomplete or inaccurate data dependencies and data integrity issues. Thus, the scientific workflow requires proper recording of dependencies and significant overheads observed during data transformation.

## 4 Proposed Line of Research

Data security supervision protects data from being accessed, modified, or stolen. Data security supervision also helps to prevent electronic data from being hacked. The classification of security controls includes management security, operational security, and physical security controls. Failure of security supervision leads to theft of intellectual property, identity theft and information extortion, and programming overheads which lead to the problem of inaccurate or incomplete data dependencies. To overcome the problem of inaccurate or incomplete data dependencies, provenance ontology and provenance tracking are to be adopted in the proposed methodology. Unstructured data, such as webpages, documents, videos, and satellite images, which are not organized and stored properly, are prone to data leakages and modifications. Modification and fabrication of data in the field of supply chain affect the company's efficiency. Hence, a clear, fine-grained ontology is required to categorize and store the data.

Figure 3 explains the proposed methodology where data is stored in layers such as file data, stream data, indexed data, and messages, which categories the unstructured data from disparate sources. MapReduce is used for data preparation. Data preparation operations include cleaning data, extracting relevant features from data, eliminating duplicate items from datasets, and changing data formats. MapReduce provides an excellent framework for conducting multiple operations in parallel while processing huge datasets. Dependency among the data and entities occurs when a task, milestone, or activity is dependent on the completion of another job or milestone before it can begin or finish. When an output from one job or project is required as a mandatory input for another activity, a dependency exists. Dependency must be



**Fig. 3** Proposed system

recognized and tracked since it affects project success. A dependency log is generated to identify the dependencies between the entities and the data which has been stored. Provenance tracking is a technique to be adopted in the proposed system to maintain user access to information. Provenance tracking enables full transparency and accountability to the corresponding environments. It helps in solving the problem of inaccurate and incomplete data dependencies which occur due to uncertainty perspectives, unauthorized user access, interoperability, and data leakage challenges.

## 5 Conclusion and Future Work

This paper discusses the 10V characteristics of big data and the various challenges faced during data transformation. The study briefly discusses the uncertainty perspectives, unauthorized user modification, interoperability, data leakage and data accuracy challenges, existing methodologies, and their limitations. From the review of existing methods proposed, the data sources used are mostly independent and dependency may exist when connected through social networks. The study presents the 10V's characteristics of big data and their correlations between the five challenges of data security. Thus, the paper serves as the outline for our future work to overcome incomplete or inaccurate data dependency problems which exist during data transformation. The proposed system is the recommended strategy for achieving full transparency and accountability in the corresponding provenance environment, where data loss can be avoided by keeping track of dependencies and substantial overheads detected in scientific workflow during data transformation.

## References

1. Dssouli, R., Serhani, M.A.: Big Data Pre-Processing: A Quality Framework. Conference Paper (2015)
2. Shakya, S.: A self monitoring and analyzing system for solar power station using IoT and data mining algorithms. *J. Soft Comput. Paradigm* **3**(2), 96–109 (2021)
3. Yin, X., Han, J., Philip, S.Y.: Truth discovery with multiple conflicting information providers on the web. *IEEE Trans. Knowl. Data Eng.* **20**(6), 796–808 (2008)
4. Missier, P., Paton, N.W., Belhajjame, K.: Fine-grained and efficient lineage querying of collection-based workflow provenance. In: Proceedings of the 13th International Conference on Extending Database Technology, pp. 299–310 (2010)
5. Huang, C., Wang, D., Chawla, N.: Scalable uncertainty-aware truth discovery in big data social sensing applications for cyber-physical systems. *IEEE Trans Big Data* (2017)
6. Hariri, R.H., Fredericks, E.M., Bowers, K.M.: Uncertainty in big data analytics: survey, opportunities, and challenges. *J. Big Data* **6**(1), 1–16 (2019)
7. Wang, X., He, Y.: Learning from uncertainty for big data: future analytical challenges and strategies. *IEEE Syst. Man Cybern. Mag.* **2**(2), 26–31 (2016)
8. Patil, P.J., Zalke, R.V., Tumasare, K.R., Shiwankar, B.A., Singh, S.R., Sakhare, S.: IoT protocol for accident spotting with medical facility. *J. Artif. Intell.* **3**(02), 140–150 (2021)

9. Bindhu, V., Ranganathan, G.: Hyperspectral image processing in internet of things model using clustering algorithm. *J. ISMAC* **3**(2), 163–175 (2021)
10. Buneman, P., Khanna, S., Tan, W.-C.: Why and where: a characterization of data provenance. In: Proceedings of International Conference on Database Theory (ICDT), pp. 316–330. London (2001)
11. Liao, C., Squicciarini, A.: Towards provenance-based anomaly detection in MapReduce. In: 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (2015)
12. Thudumu, S., Branch, P., Jin, J., Singh, J.J.: A comprehensive survey of anomaly detection techniques for high dimensional big data. *J. Big Data* **7**(1), 1–30 (2020)
13. Debnath, B., Solaimani, M., Gulzar, M.A., Arora, N., Lumezanu, C., Xu, J., Zong, B., Zhang, H., Jiang, G., Khan, L.: LogLens: a real-time log analysis system. In: 2018 IEEE 38th International Conference on Distributed Computing Systems (2018)
14. He, X., Chu, L., Qiu, R.C., Ai, Q., Ling, Z.: A novel data-driven situation awareness approach for future grids—using large random matrices for big data modeling. *IEEE Access* (2018). <https://doi.org/10.1109/ACCESS.2018.2805815>
15. Gao, Y., Chen, X., Du, X.: A big data provenance model for data security supervision based on PROV-DM model. *IEEE Access* **8**, 38742–38752 (2020)
16. Manoharan, J.S.: A novel user layer cloud security model based on Chaotic Arnold transformation using fingerprint biometric traits. *J. Innov. Image Process. (JIIP)* **3**(1), 36–51 (2021)
17. Haider, W., Hu, J., Xie, Y., Yu, X., Wu, Q.: Detecting anomalous behavior in cloud servers by nested-arc hidden semi-Markov model with state summarization. *IEEE Trans. Big Data* **5**(3), 305–316 (2017)
18. Sharma, R., Sungheetha, A.: An efficient dimension reduction based fusion of CNN and SVM model for detection of abnormal incident in video surveillance. *J. Soft Comput. Paradigm (JSCP)* **3**(2), 55–69 (2021)
19. Pang, G., Shen, C., Cao, L., Hengel, A.V.D.: Deep learning for anomaly detection: a review. *ACM Comput. Surv. (CSUR)* **54**(2), 1–38 (2021)
20. Missier, P., Belhajjame, K., Cheney, J.: The W3C PROV family of specifications for modelling provenance metadata. In: Proceedings of 16th International Conference on Extending Database Technology (EDBT), pp. 773–776. Genoa, Italy (2013)
21. Hammad, R., Barhoush, M., Abed-algumi, B.H.: A semantic-based approach for managing healthcare big data: a survey. *J. Healthc. Eng.* **12** (2020). Article ID 8865808
22. Cheung, K.H., Yip, K.Y., Smith, A., Deknikker, R., Masiar, A., Gerstein, M.: YeastHub: a semantic web use case for integrating data in the life sciences domain. *Bioinformatics* **21**(Suppl 1):i85–96 (2005)
23. Khan, F.Z., Soiland-Reyes, S., Sinnott, R.O., Lonie, A., Goble, C., Crusoe, M.R.: Sharing interoperable workflow provenance: a review of best practices and their practical application in CWLProv. *GigaScience* **8**(11), 95 (2019)
24. Zhang, O.Q., Ko, R.K.L., Kirchberg, M., Suen, C.H., Jagadpramana, P., Lee, B.S.: How to track your data: rule-based data provenance tracing algorithms. In: Proceedings of IEEE 11th International Conference on Trust, Security, Privacy Computer Communications, pp. 1429–1437. Liverpool, UK (2012)
25. Park, J., Nguyen, D., Sandhu, R.: A provenance-based access control model. In: Proceedings of 10th Annual International Conference on Privacy Security Trust, pp. 137–144 (2012)
26. Bates, A., Tian, D.J., Butler, K.R.B., Moyer, T.: Trustworthy whole system provenance for the Linux kernel. In: Proceedings of 24th USENIX Security Symposium (USENIX Security), pp. 339–334. Washington, DC, USA (2015)
27. Suen, C.H., Ko, R.K.L., Tan, Y.S., Jagadpramana, P., Lee, B.S.: S2Logger: end-to-end data tracking mechanism for cloud data provenance. In: Proceedings of 12th IEEE International Conference on Trust, Security, Privacy and Computer Communications, pp. 594–602 (2013)
28. Alabi, O., Beckman, J., Dark, M., Springer, J.: Towards a data spillage prevention process in Hadoop using data provenance. In: Proceedings of Workshop on Changing Landscapes HPC Securities (CLHS), pp. 9–13. Portland, OR, USA (2015)

29. Alneyadi, S., Sithirasenan, E., Muthukumarasamy, V.: A survey on data leakage prevention systems. *J. Netw. Comput. Appl.* **62**, 137–152 (2016)
30. Zhang, M., Huang, J., Shen, H., Xia, Z., Ding, Y.: Consecutive leakage-resilient and updatable lossy trapdoor functions and application in sensitive big-data environments. *IEEE Access* **6**, 43936–43945 (2018)
31. Shi, W., Zhu, Y., Philip, S.Y., Zhang, J., Huang, T., Wang, C., Chen, Y.: Effective prediction of missing data on Apache spark over multivariable time series. *IEEE Trans. Big Data* **4**(4), 473–486 (2017)
32. Sultana, S., Shehab, M., Bertino, E.: Secure provenance transmission for streaming data. *IEEE Trans. Knowl. Data Eng.* **25**(8), 1890–1903 (2012)
33. Agrawal, R., Imran, A., Seay, C., Walker, J.: A layer based architecture for provenance in big data. In: 2014 IEEE International Conference on Big Data (2015)
34. Korolev, V., Joshi, A.: PROB: a tool for tracking provenance and reproducibility of big data experiments. Reproduce'14. HPCA 2014, 2014—ebiquity.umbc.edu
35. Smys, S., Haoxiang, W.: Security enhancement in smart vehicle using blockchain-based architectural framework. *J. Artif. Intell.* **3**(2), 90–100 (2021)
36. Sun, L., Park, J., Nguyen, D., Sandhu, R.: A provenance-aware access control framework with typed provenance. *IEEE Trans. Dependable Secure Comput.* **13**(4), 411–423 (2015)
37. Hu, C., Li, W., Cheng, X., Yu, J., Wang, S., Bie, R.: A secure and verifiable access control scheme for big data storage in clouds. *IEEE Trans. Big Data* **4**(3), 341–355 (2017)

# An Edge-Based Disjoint Path Selection Scheme for FANETs



Orchu Aruna and Amit Sharma

**Abstract** FANET is a special type of ad hoc network. Now a days, FANET is used in different applications like civilian and military-based systems. Due to frequent changes of network topology, FANET faces unique challenges compared with previous networks such as mobile and ad hoc networks. The major problem of FANET is selection of an effective route without duplicate relay nodes in a network. In this paper, we introduce an EBDPS—an edge-based disjoint path selection scheme, which eliminates the redundant path selection and improves the energy efficiency and network lifetime. The proposed EBDPS scheme estimates the steadiness of the available link during communication and identifies the efficient relay nodes. The proposed algorithm computes the multiple robust link-disjoint paths during the process of route discovery. The selected disjoint paths effectively control the communication load and the energy efficiency during the data transmission phase.

**Keywords** FANET · Disjoint path selection · Energy efficiency · Minimum hops · Edge-based disjoint path selection scheme

## 1 Introduction

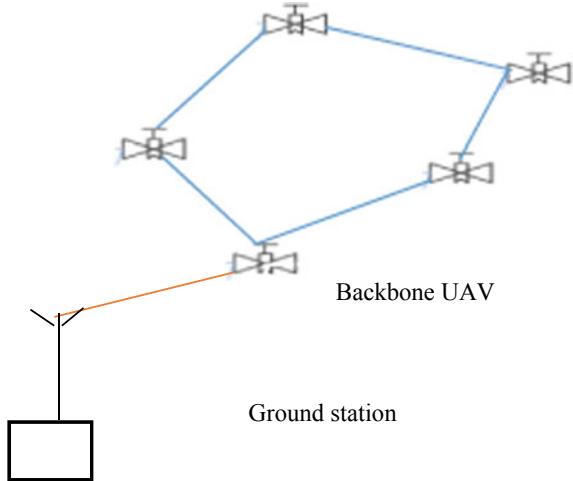
In both military and civilian domains, UAVs have been used extensively. In recent years [1], FANETs containing multiple UAVs have been investigated for enabling complicated applications that are difficult for using conventional mobile ad hoc networks or individual UAVs. The precise and prompt data delivery is required between UAVs in FANETs for applications such as wildfire monitoring and search and rescue operations. Several unique features are included in FANETs, such

---

O. Aruna (✉) · A. Sharma  
Lovely Professional University, Phagwara, Punjab, India  
e-mail: [arunasri52@gmail.com](mailto:arunasri52@gmail.com)

A. Sharma  
e-mail: [amit.25076@lpu.co.in](mailto:amit.25076@lpu.co.in)

**Fig. 1** Flying ad hoc network



as frequent topology changes, high mobility, and posing challenges on network connectivity [2].

Figure 1 shows the simplest flying ad hoc network. For maintaining the link between BS and other sub-UAVs, the part of a gateway node is included in the backbone UAV for FANETs. The wireless communication equipment will consider by gateway UAVs, and it can work under the constraints of long communication with ground stations, close communication with UAVs, low power consumption, and high power stations [3]. The same movement patterns, like direction and speed, need to be included in all connected UAVs for FANETs in order to sustain the reliable connection. For accomplishing the missions of autonomous aerial reconnaissance, the deployment of large number of same small UAVs can be done and used the communication architecture while performing those missions [4].

For data transmission, the shortest path is used by the reactive routing protocols like AODV in the route discovery process, and an alternative path is sought only if an active path is broken [5]. Link breakage is caused by the frequent route discoveries in FANETs, which characterizes based on a high degree of mobility and results in high overhead significantly [6]. Two significant problems are caused by this link failure. All packets that have transmitted on the broken route have dropped out, and the average packet delivery ratio (PDR) has reduced [7]. Until a new route is discovered, the data transmission is stopped, and the average end-to-end delay is increased.

The multiple paths' establishment is allowed in the disjoint path routing. A unique set of nodes is contained for each path between a source and destination. Two different types of paths are link-disjoint and node-disjoint paths. The common nodes are not included in the node-disjoint paths other than source and destination. In a similar manner, any common link doesn't include in the link-disjoint paths, but common nodes may exist. By comparing with the link-disjoint routes, less-effective links are

resulted due to the lower number of such disjoint routes although the links failure is guaranteed by the node-disjoint in case of main interest toward the fault tolerance during the path failure [8].

In this paper, we define a new metric called link steadiness, which defines the steadiness of the available link during data communication. The factors used to determine the link steadiness are minimum energy drain rate, node closeness, ETX, and link availability parameters, respectively. A new FANET routing protocol is proposed using this new factor, and it is known as EBDPS—an edge-based disjoint path selection scheme, which eliminates the redundant path selection and improves the energy efficiency and network lifetime. The proposed EBDPS scheme aims to reduce the routing complexity and improving energy efficiency on the links by avoiding redundant path selection using an edge-based disjoint path selection scheme. Two different components are included in the algorithm, such as route maintenance and route discovery process.

## 1.1 Contributions in This Paper

- In the proposed EBDPS scheme, a new metric called link steadiness, which defines the steadiness of the available link during data communication, is introduced in this proposed scheme.
- The newly introduced link steadiness parameter evaluates the link steadiness of the selected links based the multiple parameters like link steadiness is minimum energy drain rate, node closeness, ETX, and link availability factor.
- In the proposed scheme, the possible paths are selected and cached in the source node, and the steadiness of the selected primary path is constantly monitored to reduce the possibility of link breakage during data transmission. So, the packet delivery rate is not affected if the link breakage occurs.
- These selected disjoint paths control the number of hops required for forwarding the information to the destination via minimum hops. It achieves energy efficiency during data transmission.
- The proposed scheme of EBDPS provides improved results in delay, packet delivery ratio, and controls the overhead in the lower or higher mobility of network environments.

## 2 Literature Survey

If a UAV wants to engage in communication and the destination location is not known, the route discovery is adopted. Shirani et al. [9] have discussed the determination of the shortest path to the destination using route discovery. In the route reply (RREP) packet, the novelty is that the position of destination is included that shared with all intermediate nodes. UAVs exploit the greedy forwarding method if there is a

disconnection until the destination. As any connectivity factor is considered, the chosen path included in the links can be broken quickly as a drawback. Many route discoveries result that consuming more energy and resources.

Oubbati et al. [10] present the on-demand discovery path to address the issues of UAVs by considering the connectivity factor among UAVs. For establishing a robust routing path, the sequence of UAVs that are near each other is required. As this protocol can't able to determine various alternative solutions, it can't deal with sudden link breakages that occurred on multiple path links.

Among UAVs, unbalanced energy consumption is a serious issue. Shi and Luo [11] have demonstrated that the network is categorized into clusters, where CH is chosen using the relative velocity, energy level, and the connectivity degree with its members. The intra-cluster communications use by member nodes for direct communication. Since the residual energy is sufficient for communicating with other CHs that are located a little far, all communications will make via the CH. The residual energy is minimized as successive communications transit via the CH. It will run out of energy faster than other UAVs that results in strategy failure.

Aadil et al. [12] have focused on minimizing the overhead through the clustering formation based on a higher energy level. Based on the distance that separates the communications of UAVs, the dynamic transmission power is considered. For adequate CH selection, the formation of clusters is made using the k-means density (i.e., the neighborhood degree). This type of routing protocol is provided better performance for a path-planned mobility model, not in the case of FANET applications. To overcome the UAV's energy constraints, other types of schemes have been proposed for particular mobile nodes.

Oubbati et al. [13] have supported the exploitation of residual energy level and movement data of each UAV to ensure a high level of communication stability. The author has used the robust route discovery process to explore routing paths, which consider the discovered paths' connectivity degree, the prediction of link breakage, and the balanced energy consumption. The proposed scheme showed better results in reduced packet losses, minimized the number of path failures, and increased network lifetime.

Salam et al. [14] have proposed a bio-inspired mobility-aware clustering optimization for BIMAC-FASNET. Based on the algorithm of honey-bee optimization, the clusters are formed in FASNET. The simulation results prove that the proposed scheme shows better performance in the cluster formation time, CH lifetime, communication load, reaffiliation rate, link connection lifetime, and number of UAVs per cluster.

Smys et al. [15] have investigated the energy-effective protocols for wireless sensor networks. Based on the network hop selection, routing, and latency, the proposed protocols are categorized and examined each class for comparing the routing protocols' parameters. The NS-3 simulator is used to validate the performance outcomes. Based on the results, the routing task needs to be implemented for various technologies to prolong the network lifetime and ensure the better sensing coverage area.

Jacob and Darney [16] have considered an artificial bee colony algorithm to be implemented for routing enhancement in WSNs. The wireless communication is impacted positively by the evaluative features of an artificial bee colony optimization algorithm. It makes the routing decisions as decentralized and synchronous. By comparing with the previous state-of-the-art models, the proposed algorithm improves the performance in terms of reduced interference and increased throughput.

### 3 Proposed Framework

The motive of the proposed EBDPS—an edge-based disjoint path selection scheme is to eliminate the redundant path selection and improves the energy efficiency and network lifetime. The proposed EBDPS scheme computes stable and multiple link-disjoint paths based on the proposed metric of link steadiness, and an alternative reliable path is determined using high steadiness during link failure. It is designed primarily for low and high-mobility FANETs and where the link failures occur frequently. The new parameter of link steadiness is discussed for the route maintenance and route discovery of link-disjoint.

#### 3.1 Link Steadiness Metric

The factors used to determine the link steadiness are minimum energy drain rate, node closeness, ETX, and link availability factor, respectively (Table 1).

Let's assume that nodes  $i$  and  $j$  are in the communication range of each other.  $LST_{ij}$  refers to the link steadiness between  $i$  and  $j$ , and it can formulate as an integration

**Table 1** Factors considered in the proposed method

Factors considered	Description
LST	Link steadiness factor
MDR	Minimum-energy drain rate
NC	Node closeness
ETX	Expected transmission count
LA	Link available factor
$E_{res}$	Residual energy
DRI	Drain rate index
$D_{ij}$	Distance between node $i$ and node $j$
$R$	Communication radius
$d_f, d_r$	Forward delivery ratio and reverse delivery ratio
$T_x$	Transmission range

of the minimum energy drain rate  $\text{MDR}_{ij}$ , node closeness  $\text{NC}_{ij}$ , ETX as  $\text{ETX}_{ij}$  and link availability factor  $\text{LA}_{ij}$  as follows in Eq. (1):

$$\text{LST}_{ij} = \alpha_1 \text{MDR}_{ij} + \alpha_2 \text{NC}_{ij} + \alpha_3 \text{ETX}_{ij} + \alpha_4 \text{LA}_{ij} \quad (1)$$

where  $\alpha_1, \alpha_2, \alpha_3$  &  $\alpha_4$  indicate the weighting coefficients that constrained by the below Eq. (2):

$$\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1 \quad (2)$$

$$\text{MDR}_{ij} = \frac{E_{\text{res}}}{\text{DRI}_i(t)} \quad (3)$$

$$\text{NC}_{ij} = \frac{R - D_{ij}}{R} \quad (4)$$

where  $D_{ij}$  represents the two nodes' Euclidean distance and  $R$  is the node's communication radius.

$$\text{ETX} = \frac{1}{d_f x d_r} \quad (5)$$

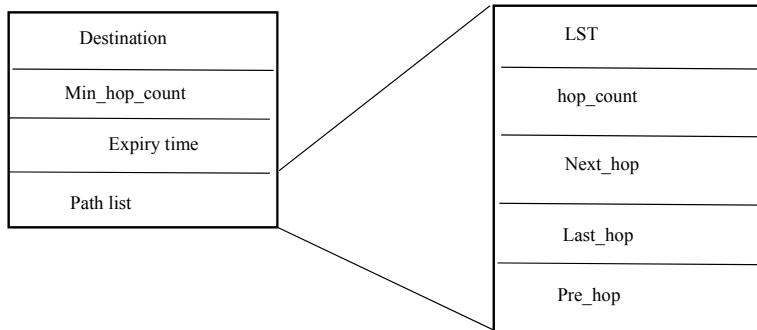
$d_f$  is defined as the measured probability of an arrived data packet at the recipient, and the reverse delivery ratio  $d_r$  is the probability of successfully received packets.

$$\text{LA}_n = \frac{1}{\left(1 - R_n/Tx_n + 1\right)} \quad (6)$$

Here,  $\text{LQ}_n$  denotes the link quality of the node  $n$ ;  $R_n$  denotes the radius of the node  $n$ ,  $Tx_n$ , and denotes the maximum transmission range of the node.

### 3.2 Route Discovery

For determining the multiple and stable link-disjoint paths between source and destination pairs, the control packet structures are modified and two additional fields called `ini_hop` and `LST` are added with them. Each routing table entry structure is shown in Fig. 2. The novelty in the route discovery mechanism is, each every link is estimated with link steadiness factor which comprises of minimum energy drain rate, node closeness, ETX, and link availability factor. The computed LST is added to the route discovery mechanism and considered as the primary parameter for route selection. Since all the estimated have the link steadiness metric, it is easier for the proposed method to share the LST value with the neighbor nodes during route

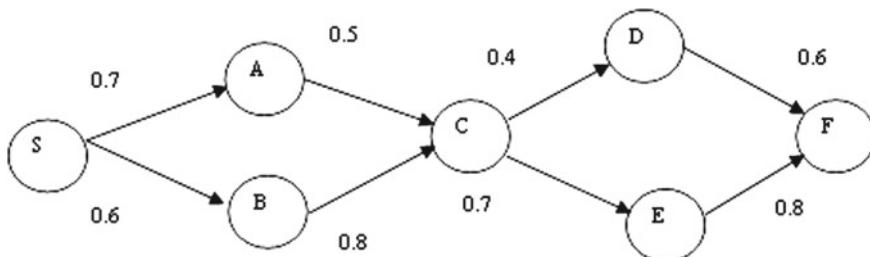
**Fig. 2** Routing table structure in nodes

discovery process. Since LST is the primary route selection parameter, it is easier for the source node to select and establish the route based on the estimated route stability.

Here, min\_hop\_count indicates all paths' minimum hop count to the same destination. The constraint of min\_hop\_count plus one is not shorter than the hop count should satisfy by each path in a route. Otherwise, it leads to routing loops. The pre\_hop field indicates the previous hop address from that the packet received and last\_hop is the last hop address to the destination.

Consider the below example for a better explanation of the route discovery process. The number near the link represents the link steadiness LST metric (Fig. 3).

Node C is paired with the reverse path, which has the largest metric if it receives the message from neighbor D. A complete route (S-A-C-D-F) is formed by transmitting a response message via node A. Some predefined threshold is smaller than the corresponding path metric difference if the previously arrived metric (0.4) is much smaller than the response copy from node E, which has link\_st (0.7). Good stability will not achieve by both the routes when a new response packet transmits via the

**Fig. 3** Route discovery

unused reverse route, ex: C-B-S. These two link-disjoint paths' combinations like S-B-C-E-F and S-A-C-E-F are sharing for maximizing the stability. The response message of 0.7 metric will send via node A, and the other message with 0.4 metric will send through node B.

### **Algorithm (Pseudocode)**

MDR—minimum energy drain rate, NC—node connectivity, ETX—expected transmission count, LA—link availability, LST—link steadiness metric;

NODE\_LST—LST value of the node; SELECT\_LST—threshold LST value to select the node;

DIFF\_THRESHOLD—difference of LST value between two response packets;

BREAK\_THRESHOLD—predefined threshold to consider the path is supposed to be broken.

```

For all nodes n
    Calculate MDR,NC,ETX,LA
    Calculate LST
End for
Node n store LST in routing table
Route discovery phase
    SOURCE broadcast control packets
    DESTINATION send response_packet
    If new_response_packet
        Intermediate node n checks for NODE_LST
        If (NODE_LST > SELECT_LST)
            Add node n into forwarder_list
    End if
    If duplicate_response_packet
        If (NODE_LST > DIFF_THRESHOLD)
            Add node n into forwarder_list
        Else
            Discard the packet
    End if
Route Maintenance Phase
    If (NODE_LST > BREAK_THRESHOLD)
        DESTINATION Intimate the link breakage
        SOURCE Check for BACKUP_PATH
            If BACKUP_PATH exists
                Retransmit the data through BACKUP_PATH
            Else
                Reinitiate the route discovery process
        End if
End if

```

## 4 Results and Discussion

The proposed protocol comparative analysis is done with the existing ECAD and BIMAC protocols. For the simulation requirement, totally of 25 nodes are considered, which are placed randomly in the network. Since our network is FANET, the random waypoint model movement was given to the nodes. MAC 802.15.4 utilizes in order to facilitate the FANET communication property. The initial energy capacity of 100 J is configured with every node in the network with each node in omni-antenna direction. The constant bit rate (CBR) traffic generator is used to generate consistent traffic during data transmission. The data communication is carried out by UDP as no acknowledgment needed from the receiver node. The implementation of the protocol is done in NS-2, and performance is compared with ECAD and BIMAC protocols. Table 2 represents the simulation table of network process.

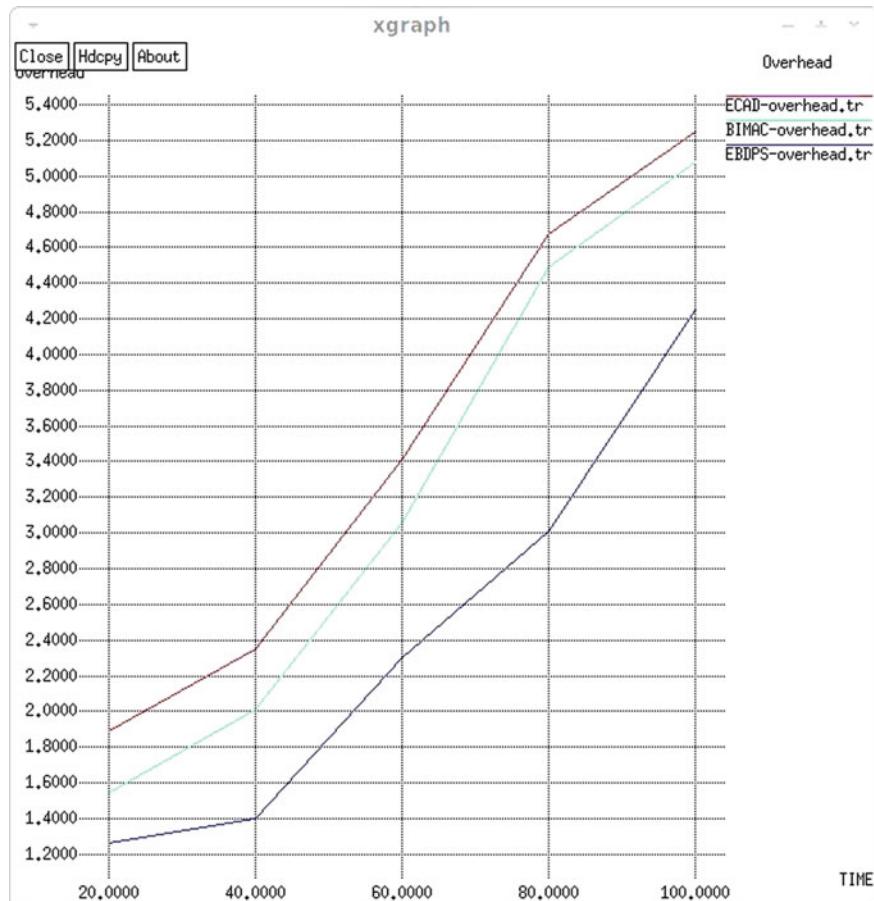
Figure 4 shows the simulation results of routing overhead for the proposed algorithm and previous existing techniques. Routing overhead refers to the transferred or transmitted total number of packets from one node to another. The routing process overhead, packet preparation, and routing table in a node are included. In our proposal, the routing overhead is minimized by avoiding path redundancy and calculating the path steadiness metric. It avoids the selection of unstable routes thus fewer path failures. The simulation results listed in the Table 3 prove that the proposed method reduces the overhead up to 0.12 compared with the existing methods.

The ratio of a delivered packet to the total sent packet from source to destination is the packet delivery ratio (PDR). Figure 5 shows the results of the packet delivery ratio for the proposed algorithm EBDPS and other previous methods like ECAD and BIMAC. The maximum number of data packets is reached to the destination. The efficient route selection improves the PDR by selecting the efficient relay nodes. The proposed mechanism measures the link stability using the steadiness metric; hence, the appropriate nodes are only selected for routing. Based on the analysis of simulation results, the efficiency of PDR is achieved by the proposed algorithm than the previous techniques listed in Table 4.

Figure 6 displays the throughput results of proposed algorithms and existing methods. Throughput is defined as the total units of data in a system is measured that can process for a given time. The high throughput rate ensures high data deliverability to the intended destination. The quick path change affects the network throughput

**Table 2** Simulation table

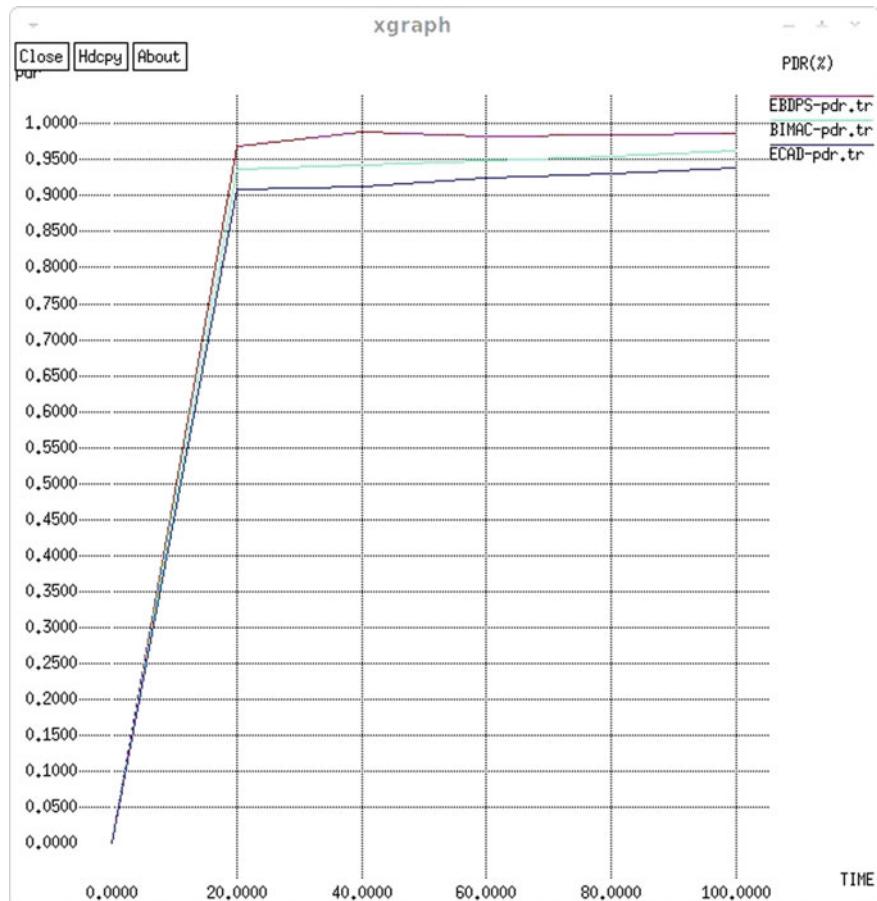
Parameter	Value
Number of nodes	25
Network area	$1000 \times 1000 \text{ m}^2$
Initial energy	100 J
MAC type	802_15_4
Routing protocol	AODV
Simulation time	100 s



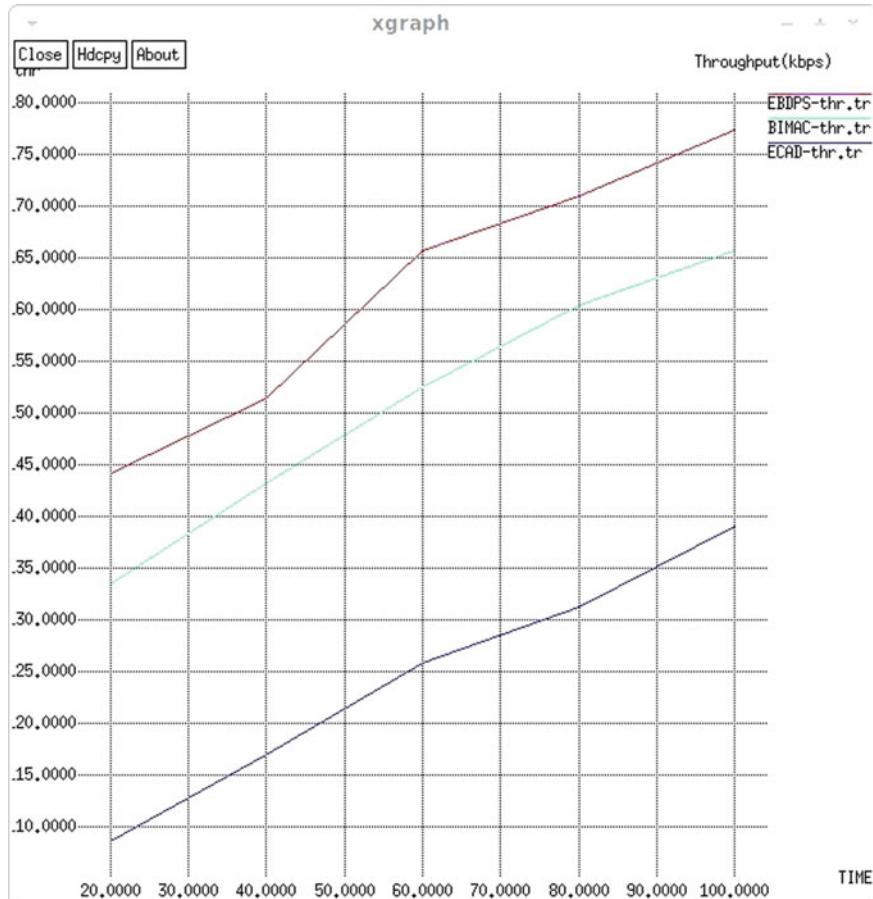
**Fig. 4** Routing overhead

**Table 3** Time versus routing overhead

Time	ECAD	BIMAC	EBDPS
20	1.89	1.54	1.26
40	2.35	2.01	1.40
60	3.41	3.06	2.30
80	4.68	4.49	3.01
100	5.25	5.08	4.25

**Fig. 5** Packet delivery ratio**Table 4** Time versus packet delivery ratio

Time	ECAD	BIMAC	EBDPS
20	0.9085	0.9369	0.9695
40	0.9124	0.9430	0.9891
60	0.9238	0.9497	0.9826
80	0.9317	0.9548	0.9839
100	0.9397	0.9635	0.9861



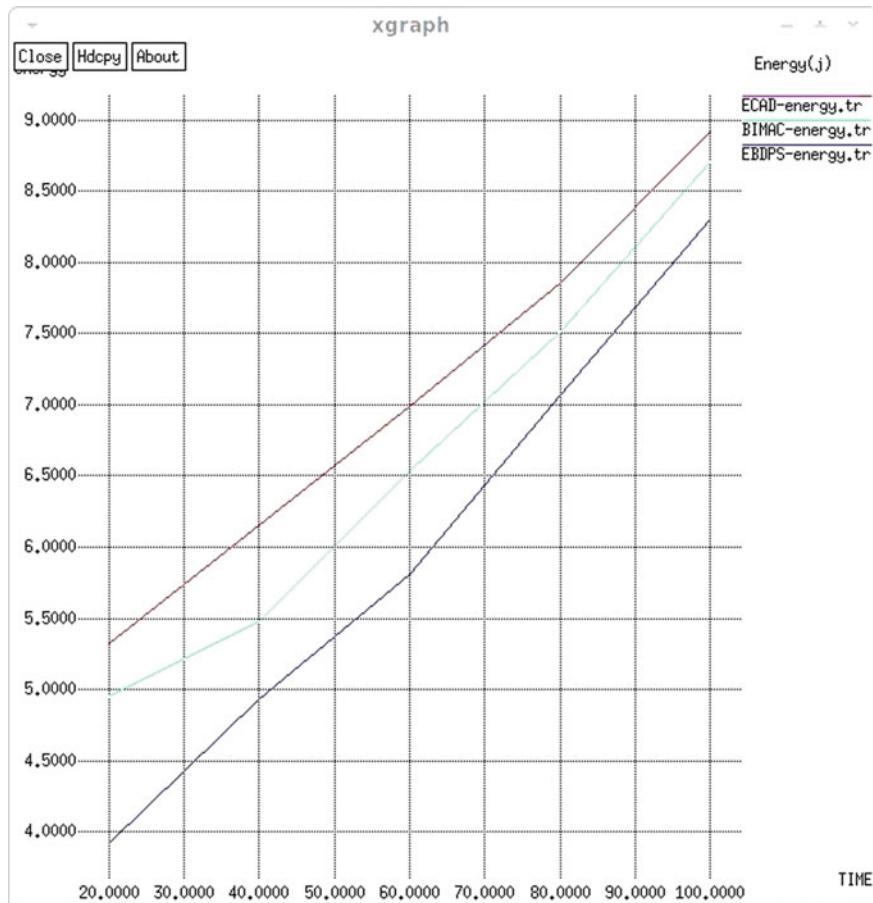
**Fig. 6** Throughput

majorly. In our proposal, the path change issue was tackled by selecting the reliable relay nodes by estimating the steadiness of every link. And the relay nodes are selected based on their steadiness and available energy. Due to this, the selected path will be stable for a long time and ensures a high data delivery rate. Thus, the proposed method shows better throughput results listed in Table 5 over the existing ones like ECAD and BIMAC.

Figure 7 shows the simulation results of energy consumption for the proposed algorithm and previous existing techniques. Maintaining sufficient energy helps the UAVs to fly high and longer duration. The UAV's flying ability is directly connected with the available energy. The high energy consumption leads to quicker energy drain, and they lost their flying capability. The proposal method ensures optimized energy utilization by selecting the appropriate relay nodes and avoiding path redundancy. By comparing with the existing methods of BIMAC and ECAD, the proposed method

**Table 5** Time versus throughput

Time	ECAD	BIMAC	EBDPS
20	108.56	133.42	144.13
40	116.87	143.17	151.42
60	125.78	152.46	165.75
80	131.20	160.42	171.01
100	139.05	165.64	177.39

**Fig. 7** Energy consumption

**Table 6** Time versus energy consumption

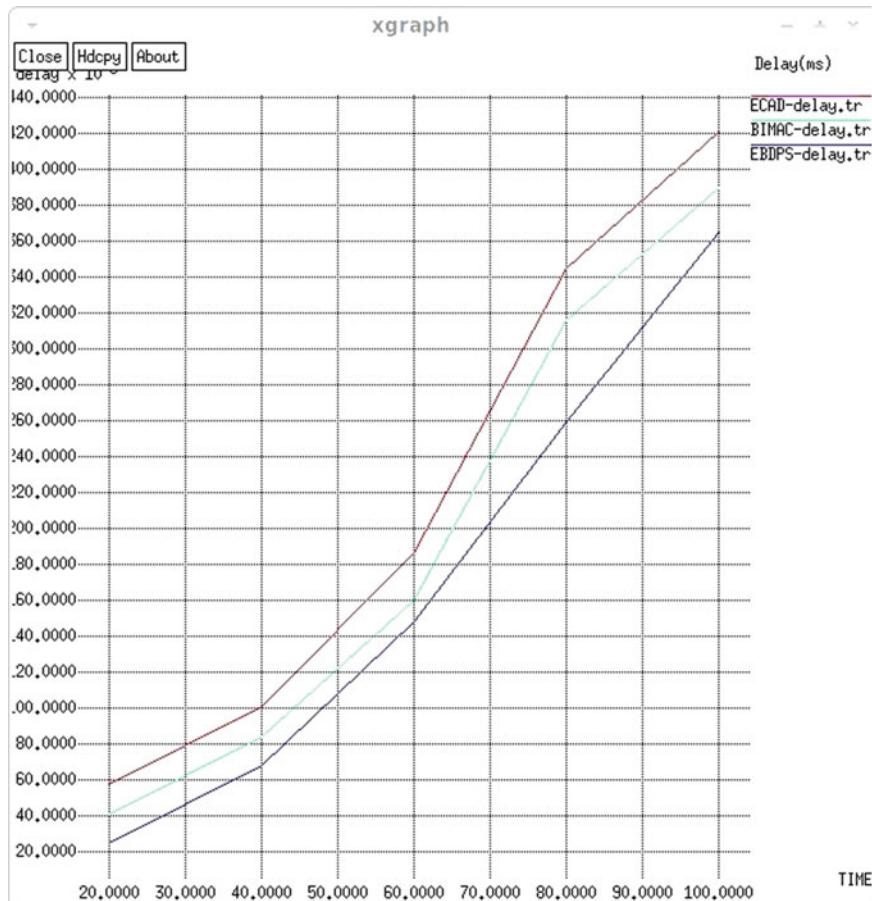
Time	ECAD	BIMAC	EBDPS
20	5.328	4.954	3.920
40	4.987	5.480	4.933
60	6.987	6.540	5.811
80	7.854	7.510	7.067
100	8.918	8.70	8.296

EBDPS saves a considerable amount of energy to improve the network lifetime it has listed in Table 6.

Figure 8 illustrates the end-to-end delay results of the proposed method EBDPS. Delay is an important QoS parameter for forwarding data in a time constraint environment. Minimizing the communication delay improves the network lifetime to a great level. The selection of relay nodes for communication ensures a high data delivery rate within the estimated time. Also, the consideration of the link steadiness metric ensures the participation of stable relay nodes for data communication. The simulation results demonstrate that the proposed algorithm shows effective results listed in Table 7 in end-to-end delay comparative to other methods of ECAD and BIMAC.

## 5 Conclusion

The high mobility and frequent topology changes are the major concern in achieving efficiency in FANETs. It includes various distinctive challenges owing to the dynamic topological structure, selecting the effective reliable relay node without redundancy. Redundant link and relay node selection played a vital role in controlling the communication load and energy efficiency. In this work, we introduce an EBDPS—an edge-based disjoint path selection scheme, which eliminates the redundant path selection and improves the energy efficiency and network lifetime. The proposed scheme uses minimum energy drain rate, node closeness, ETX, and link availability factor to estimate the link steadiness metric. The selected disjoint paths effectively control the communication load and the energy efficiency during the data transmission phase. The proposed algorithm is outperformed by the existing schemes in terms of control overhead, packet delivery ratio, and end-to-end delay in either lower or higher mobility.



**Fig. 8** End-to-end delay

**Table 7** Time versus end-to-end delay

Time	ECAD	BIMAC	EBDPS
20	0.058	0.041	0.025
40	0.101	0.084	0.068
60	0.187	0.160	0.148
80	0.345	0.316	0.259
100	0.421	0.390	0.365

## References

- Shahhatreh, H., Sawalmeh, A.H., Al-Fuqaha, A., Dou, Z., Almaita, E., Khalil, I., Othman, N.S., Khreishah, A., Guizani, M.: Unmanned aerial vehicles (UAVs): a survey on civil applications and key research challenges. *IEEE Access* **7**, 48572–48634 (2019)
- Mahmud, I., Cho, Y.-Z.: Adaptive hello interval in FANET routing protocols for green UAVs. *IEEE Access* **7**, 63004–63015 (2019)
- Islam, N., Rashid, M.M., Pasandideh, F., Ray, B., Moore, S., Kadel, R.: A review of applications and communication technologies for internet of things (IoT) and unmanned aerial vehicle (UAV) based sustainable smart farming. *Sustainability* **13**(4), 1821 (2021)
- Noor, F., Khan, M.A., Al-Zahrani, A., Ullah, I., Al-Dhlan, K.A.: A review on communications perspective of flying ad-hoc networks: key enabling wireless technologies, applications, challenges and open research topics. *Drones* **4**(4), 65 (2020)
- Nayyar, A.: Flying ad-hoc network (FANETs): simulation-based performance comparison of routing protocols: AODV, DSDV, DSR, OLSR, AOMDV, and HWMP. In: 2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), pp. 1–9. IEEE (2018)
- Srivastava, A., Prakash, J.: Future FANET with application and enabling techniques: anatomization and sustainability issues. *Comput. Sci. Rev.* **39**, 100359 (2021)
- Duan, Y., Lam, K.-Y., Lee, V.C.S., Nie, W., Li, H., Ng, J.K.-Y.: Packet delivery ratio finger-printing: toward device-invariant passive indoor localization. *IEEE Internet Things J.* **7**(4), 2877–2889 (2020)
- Park, P., Ghadikolaei, H.S., Fischione, C.: Proactive fault-tolerant wireless mesh networks for mission-critical control systems. *J. Netw. Comput. Appl.* **186**, 103082 (2021)
- Shirani, R., St-Hilaire, M., Kunz, T., Zhou, Y., Li, J., Lamont, L.: On the delay of reactive-greedy-reactive routing in unmanned aeronautical ad-hoc networks. *Proc. Comput. Sci.* **10**, 535–542 (2012)
- Oubbati, O.S., Lakas, A., Zhou, F., Güneş, M., Lagraa, N., Yagoubi M.B.: Intelligent UAV-assisted routing protocol for urban VANETs. *Comput. Commun.* **107**, 93–111 (2017)
- Shi, N., Luo, X.: A novel cluster-based location-aided routing protocol for UAV fleet networks. *Int J Digital Content Technol. Appl.* **6**(18), 376 (2012)
- Aadil, F., Raza, A., Khan, M.F., Maqsood, M., Mahmood, I., Rho, S.: Energy-aware cluster-based routing in flying Ad-Hoc networks. *Sensors (Basel, Switzerland)* **18**(5), 1413–1428 (2018)
- Oubbati, O.S., Mozaffari, M., Chaib, N., Lorenz, P., Atiquzzaman, M., Jamalipour, A.: ECaD: Energy-efficient routing in flying ad hoc networks. *Int. J. Commun. Syst.* **32**(18), e4156 (2019)
- Salam, A., Javaid, Q., Ahmad, M.: Bioinspired mobility-aware clustering optimization in flying ad hoc sensor network for internet of things: BIMAC-FASNET. *Complexity* **2020** (2020)
- Smys, S., Bashar, A., Haoxiang, W.: Taxonomy classification and comparison of routing protocol based on energy efficient rate. *J. ISMAC* **3**(2), 96–110 (2021)
- Jacob, I.J., Darney, P.E.: Artificial bee colony optimization algorithm for enhancing routing in wireless networks. *J. Artif. Intell.* **3**(1), 62–71 (2021)

# Sign Language Interpreter



Ramya Srikanteswara , C. B. Niveditha, A. Sindhu Sai, Reddigari Keerthi Reddy, and S. A. Akshayanjali

**Abstract** Sign language is the main source of communication for the deaf-mute people. These people go through many kinds of problems whilst communicating in person or through any other devices. To overcome this communication barrier, they need an interpreter which converts the sign language into text. In some situations, these impaired people may be unknown with sign language. Thus, necessity of sign interpreter is unpreventable. Developing this kind of interpreter needs a wide range of knowledge in fields such as deep learning, image processing, and convolution networking. The crucial point of this analysis is to know whether recognizing the gesture can succeed in assisting the self-learners in learning the sign language. This ideology can avoid their quarantine from the rest of the society notably. Results from this literature review could help in development of an efficient sign interpreter which helps for the communication between non-signer and a signer.

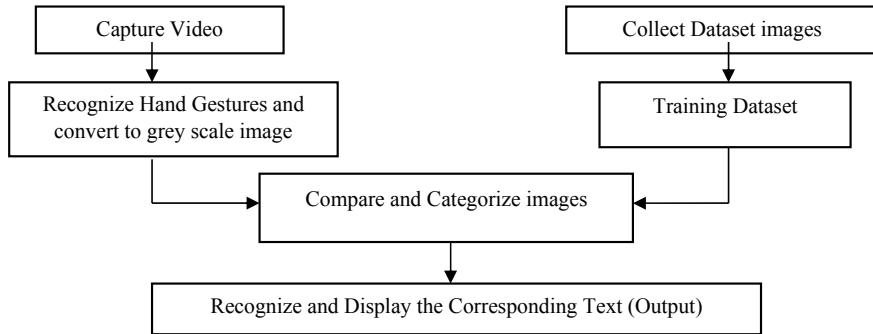
**Keywords** Sign language · Interpreter · Speech · Text · Signer

## 1 Introduction

In this modern world, humans often use the technological advancements such as mobile devices [1–3]. Vocally, impaired people are often off the beaten track of accessing social interactions, education. Assistive technology is the hope for these people which can solve their problems and encourage to lead a normal life. To put numbers in context, census 2020, World Health Organization (WHO) says that there are 466 million people all over the world (5% of world population) suffering with hearing problems of whom 34 million are children. When it comes to employment, these people will be facing a lot of issues and difficulties. Some of deaf-mute people feel they are discriminated because of their disability whilst some avoid socializing as they feel them also maybe unskilled in sign language. 68% of people with hearing loss feel off the beaten track at work as they cannot communicate properly. According to

---

R. Srikanteswara (✉) · C. B. Niveditha · A. Sindhu Sai · R. K. Reddy · S. A. Akshayanjali  
Nitte Meenakshi Institute of Technology, Bengaluru, India  
e-mail: [ramya.srikanteswara@nmit.ac.in](mailto:ramya.srikanteswara@nmit.ac.in)



**Fig. 1** Block diagram

the World Federation of the Deaf, 300 sign languages are in use around the world and 70 million impaired individuals are using them. Sign languages, like other languages, are naturally germinated. They are considered as highly structured systems governed by a set of linguistic rules [4] (Fig. 1).

## 2 Objective

The main objective of this survey paper is to analyze the different techniques or methodologies used to interpret the sign languages so far. This will give clear idea about the pros and cons of techniques used in the previous papers, which in turn helps to build an effective interpreter.

## 3 Background

Communication is the most important part of everyone's day to day life. Sign languages and gestures are helping them a lot to communicate with rest of the world. Limited research has been done about the Indian sign language due to the complex pattern of gestures [5]. The sign can be a word or a spelling which can be represented using fingers [6]. The challenge is to recognize the hand postures using two-dimensional representation provided by image or video. In most of the intelligent system, architecture of classifier for recognizing sign language uses convolution neural network architecture. The depth sensors help in capturing extra information to improve accuracy [7]. The sign language techniques are majorly classified into 3 groups: wearable sensor-based models, computer-based models, and hybrid systems. Sensor-based sign language recognition makes use of stain sensors, pressure sensors, or inertial sensors. Unlike camera-based systems, sensor-based systems are less likely

to get affected by environmental conditions. The major disadvantage of this technique is that the user may feel uncomfortable due to restricted movement as according to sensor configuration. To overcome this, sensors can be fixed into wearable hand gloves, wrist bands, or watches.

Computer-based systems make use of camera and image processing techniques to classify the gestures. This model is quite user-friendly as it the gestures are performed using plain hand and it costs less. But this approach is limited by external factors such as lightening, background, position of the camera, and shadows. More than one camera can be used to take three dimensional images which give more accurate results. Hybrid system is combination of both sensor-based and computer vision-based approaches. This typically makes use of camera and wearable sensors. This approach gives high recognition rate. But this technique restricts the mobility of the user.

## 4 Methodology

Majorly four steps are used to translate the sign language.

### A. Data Collection

Data collection is a method of collecting the images of signing hand which represents various gestures. Images can be captured using several different varieties of cameras. Database will be created for the training as well as testing purpose. The resolution of the images varies from device to devices and even the background of the image matters. To reduce the computational errors and for better comparison, image should be pre-processed. So that all the images will be of equal scale.

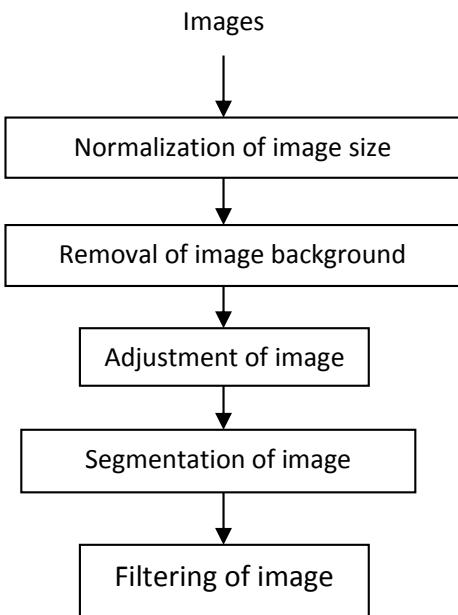
### B. Hand Partitioning

Hand partitioning is the method used to take out the hand sign from the collected images [6]. Hand region needs to be extracted by removing the background. The colour-based segmentation is one of the simplest ways of extracting structure of hand. The resultant image of segmentation process will be a binary image, in which white colour act for skin pixels and dark black colour act for background of the image. This may contain segmentation errors which can be reduced by filtering and morphological operations (Fig. 2).

### C. Feature Extraction

It is important to convert the data collected into some representation for many applications before training a network. To classify the images obtained from the previous step, it is necessary to extract some key features of the image. Hand shape is most prominent feature amongst all. Many techniques are available to represent the shape.

**Fig. 2** Hand partitioning/data pre-processing

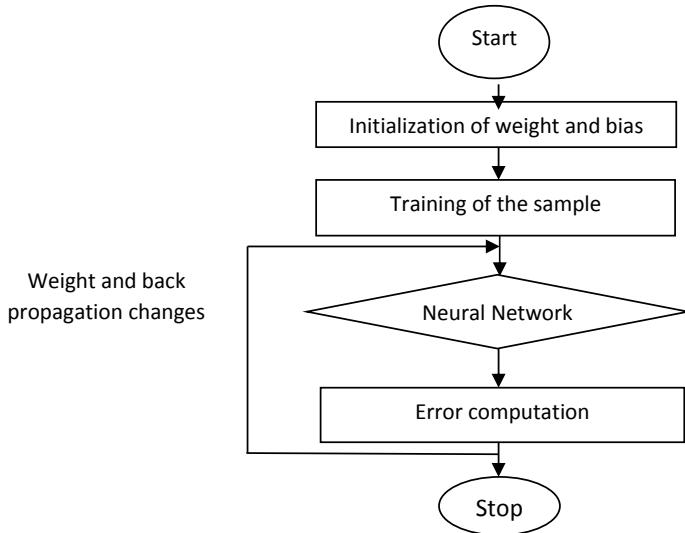


#### D. Categorization

The features extracted in the previous method are utilized here as classifier input. This step will recognize the sign. The ANN is one of the most used tools. Artificial neural network has many applications in pattern recognition field. This step includes training and testing phase. Training is done to configure the neural networks in the way that the set of inputs should produce the set of desired outputs. Testing is done to check the efficiency and accuracy of the system (Fig. 3).

### 5 Techniques

1. **Computer Vision Technique [1]:** To enable the camera for capturing instantaneous movements, we use this computer vision technique. These images should be converted into grey scale using canny-edge algorithm. The speech recognition library is used to translate the voice recorded into plain text.
2. **Transfer Learning [8]:** Transfer learning is one of the ML techniques. In this technique, the models are trained on huge datasets and restructured to fit more distinct data. This can be done by reprocessing the part of weights from the existing pre-trained model and changing or reinitializing the weights at superficial layers. For instance, a fully trained network can be able to classify few more additional sets of data by reinitializing its weights at the final classification

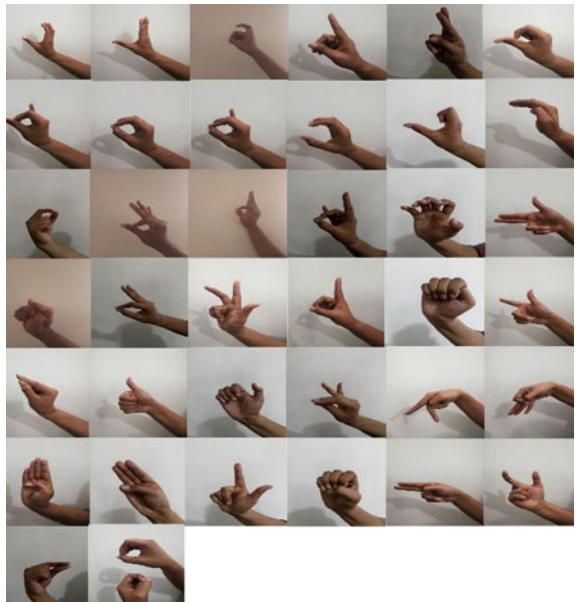


**Fig. 3** Categorization/classification

layer. The major advantage of this is that it offers less data requirements and time.

3. **GF with PCA** [9]: Gabor filter is used for extracting features from the images. Combination of an elliptical Gaussian with a tangled exponential which represents sinusoidal plane wave produces a 2-dimensional Gabor filter. PCA is one of the statistical techniques used in the field of pattern recognition of data of high dimension.
4. **LSTM Encoding and Decoding:** This NN model was put forward by Hochreiter et al. [10]. This network implements succession model which helps to control input frames. This model operates the sequence data. This helps in learning information with longer learning cycle, and it will also avoid the problem of gradient disappearance.
5. **Convolutional Neural Network:** CNN is multi-layered neural network by which visual patterns can be recognized [11]. It is better than traditional machine learning techniques. Difficult to implement on small dataset. Often over fitting takes places whilst implementing on small datasets [12]. To avoid over fitting, most of the researchers refer to use existing model that are implemented on large dataset like ImageNet. They use fine-tuning for training small dataset using the pre-defined models that are implemented on large dataset. They used VGG19 network, and then modify it to detect the BDSL alphabets (Fig. 4).
6. **Digital Image Processing and ANN:** An algorithm is used for image processing; this method is done using digital computer. ANN has 3 interconnected layers input hidden and output layers [11]. In this system, it takes sign input through MATLAB image processing and converts into text. This system

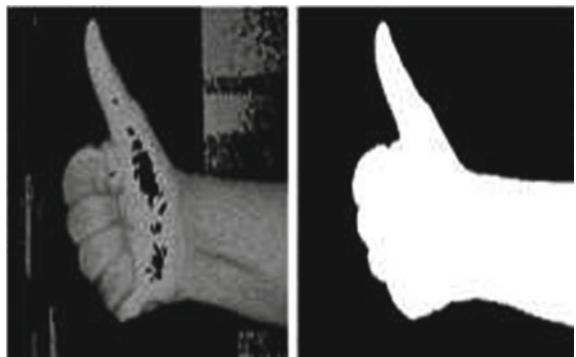
**Fig. 4** Overview of the network [12]



provides two-way communication and helps in easy interaction between normal people and visually impaired people. There are three levels of processing image low-level, middle-level, and high-level process [12]. The histogram of oriented gradient (HOG) is a feature description used in image processing. In this system, HOG is used for object detection (Fig. 5).

7. **CNN and Image Pre-processing:** An algorithm is used for image processing; this method is done using digital computer. ANN has 3 inter-connected layers input hidden and output layers [13]. In this system, it takes sign input through MATLAB image processing and converts into text. This system provides two-way communication and helps in easy interaction between normal people and

**Fig. 5** Using MATLAB to recognize sign language [11]



visually impaired people. There are three levels of processing image low-level, middle-level, and high-level process [14]. The histogram of oriented gradient (HOG) is a feature description used in image processing. In this system, HOG is used for object detection.

8. **Minimum Eigen Value Algorithm** [15]: There are diverse types of methods exists: edge, corners, and blobs. This algorithm focuses on corners because they are resistant to opening problem here Shi-Tomasi method is used it is also called minimum Eigen value algorithm. The corner tips in the image are scalar object corner can be detected by looking at intensity value of image.
9. **CNN and Deep Learning:** This model can translate the Indian sign language (ISL) to spoken English [15]. The following steps are used to convert the signs to English:
  1. Create the database containing the hand gestures of sign language.
  2. Hand gestures can be recognized by input application of neural networks and algorithm.
  3. Processing is done by using categorizing and deep learning methods to increase the working of model.
  4. Converts sign to spoken English.

Convolutional neural network (CNN) has basic features:

1. Insert information will be processed for algorithm use.
2. Sign and gestures are classified and train the system with number of iterations on training dataset.
3. Provides the output for the gesture.

Using this, sign is converted to text messages. But converting text-to-speech is done by number of libraries and APIs exit functions are used [16]. Algorithm is developed on CNN using Kera's deep learning library using this identification and classification [17]. This model can provide accurate output (Table 1).

## 6 Conclusion

Technology is an obsession for this modern world. Mobile phone has a major impact on this technology. In this application, we can change sign language into text so that communication will be easy for deaf and mute people. We have surveyed several techniques and methodologies employed in sign language interpretation technique. In most of the cases, the translation is one way, either from sign language to text/speech or text/speech to sign language. A complete two-way translator is exceedingly rare for ISL. We believe that this research is a great start for implementing this for deaf community.

**Table 1** Techniques and features used in this application

Authors	Algorithm/notation used	Advantages	Limitations	Accuracy	Year
Shinde et al. [1]	Computer vision, canny-edge detection algorithm	A GUI based, user-friendly mobile application	Unable to create an API and the application can not utilize the cloud	Phase 1–85% Phase 2–95%	2020
Sruthi et al. [2]	CNN using DL technique,	More accurate due to large dataset	Videos taken directly from mobile camera cannot be used	98.64%	2019
Soewito et al. [3]	Hand gestures extraction algorithm	Evaluating accuracy of image recognition, AI system for the conversion of images of gestures into Indonesian text	Less accuracy for blurry images	87%	2020
Adithya et al. [6]	Distance transformation, Fourier descriptors, Kurtosis, ANN	No need to wear any device by user, low computational complexity	Images captured in poor lighting with colourful background may show low accuracy	91.11%	2013
Beena et al. [7]	PDNN implementation of the CNN, GPU-enabled system with Theano	The application translates and speaks the translated sentences w.r.t to the gesture	Applicable only for static gestures	94.6774%	2017
Admasu et al. [9]	EMA using GF with PCA, ANN	Fault tolerant, able to identify unknown input sign fast	Gave low performance when the orientation level is increased	–	2010
Siming et al. [10]	Linde-Bunzō-grey (LBG) algorithm	No gradient disappearance	Dataset is restricted to range	91.5%	2019
Satheesh et al. [15]	Minimum Eigen value algorithm	Takes less space and less computational time	If image is not stored in database, system cannot give output	–	2015

(continued)

**Table 1** (continued)

Authors	Algorithm/notation used	Advantages	Limitations	Accuracy	Year
Kishore et al	Elliptical Fourier descriptors, ANN	Less computational time	Works only in dark background	95.1%	2015
Tripathi et al. [17]	DWT, HMM	Not restricted only to dark background	Person should use gloves which are overly expensive	80.4%	2015
Pandey et al. [13]	K-means clustering, MATLAB	Video of person's signing hand is taken instead of images	Performance was satisfactory in case of segmenting the region of hand from image	More than 90% for most of the alphabet	2015
Athira et al. [18]	SVM for classification, multi-class C-SVC to train dataset	Good accuracy, no wrist band required, economical	Can give better results only in uniform background and under proper lighting conditions	89% for single handed dynamic gesture 91% for figure spelling gesture	2019
Daphne et al. [19]	OpenCV, TensorFlow, Keras	High accuracy in identifying the ASL static gestures	Will give accurate results only in well-lit room	Original-74% Skin mask- 72% Sobel filtered 71%	2019
Harini et al. [20]	Python OpenCV library, CNN	Memory requirement is less, and accuracy is high because of csv dataset	Accuracy is less with poor lighting	99.91%	2020

## References

- Shinde, A., Daddona, R.: Two-Way Sign Language Converter for Speech-Impaired (February 2020)
- Sruthi, C.J., Lijiya, A.: Signet: A Deep Learning based Indian Sign Language Recognition System (April 2019)
- Soewito, B., Khrisna, A., Satyadhana, R.: Communication on Mobile Phone for The Deaf Using Image Recognition (August 2020)
- Bragg, D., Koller, O., Bellard, M., Berke, L.: Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective (August 2019)
- Jayadeep, G., Venugopal, V., Vishnupriya, N.V., Vishnu, S., Geetha, M: Mudra: Convolutional Neural Network based Indian Sign Language Translator for Banks. Department of Computer Science and Engineering Amrita Vishwa Vidyapeetham, Amrita Puri, India

6. Adithya, V., Vinod, P.R., Gopalakrishnan, U.: Artificial Neural Network Based Method for Indian Sign Language Recognition. *ICT* (2013)
7. Beena, M.V., Agni Sarman Namboodiri, M.N.: Automatic Sign Language Finger Spelling Using Convolution Neural Network: Analysis, vol. 117, no. 20, pp. 9–15 (2017)
8. Garcia, B., Alarcon Viesca, S.: Real-time American Sign Language Recognition with Convolutional Neural Networks. *Stanford University Stanford, CA*
9. Fantahun Admasu, Y., Raimond, K.: Ethiopian Sign Language Recognition Using Artificial Neural Network. Department of Electrical and Computer Engineering, Addis Ababa University, Addis Ababa, Ethiopia (2010)
10. He, S.: Research of a Sign Language Translation System Based on Deep Learning. *Ridley College, St. Catharine's, Canada* (2019)
11. Sign language and gesture detection for deaf and dumb people. *Int. J. Dev. Res.* **4**(3), 749–752 (2014)
12. CVPR2021W: Cha Learn Papers Gruber Mutual Support of Data Modalities in the Task of Sign CVPRW 2021 Paper
13. Pandey, P et al.: An efficient algorithm for sign language recognition. (*IJCSIT*) *Int. J. Comput. Sci. Inf. Technol.* **6**(6) (2015)
14. Ananth Rao, G., Kishore, P.V.V.: Selfie Video Based Continuous Indian Sign Language Recognition System. Department of Electronics and communication Engineering (February 2017)
15. Raju, S.K., Anil Kumar, G.S., Arokia Swamy, S.: Double Handed Indian Sign Language to Speech and Text. Department of Electrical and Electronics Engineering (2015)
16. Kishore, P.V.V., Prasad, M.V.D., Raghava Prasad, Ch., Rahul, R.: 4-Camera Model for Sign Language Recognition Using Elliptical Fourier Descriptors and ANN (2015)
17. Tripathi, N., Nandi, G.C.: Continuous Dynamic Indian Sign Language Gesture Recognition with Invariant Backgrounds (2015)
18. Athira, P.K., Sruthi, C.J., Lijiya, A.: A Signer Independent Sign Language Recognition with Co-articulation Elimination from Live Videos: An Indian Scenario. Accepted 5 May 2019
19. Tan, D., Meehan, K.: Implementing Gesture Recognition in a SL Learning Application. Department of Computing Letterkenny Institute of Technology Letterkenny, Ireland (2019)
20. Harini, R., Janani, R., Keerthana, S., Madhubhala, S., Venkatasubramanian, S.: Sign Language Translation (2020)

# Architectural Insight of Neural Information Extraction, Retrieval, and Processing for Multimodal Neural Search



Jainal S. Gosaliya , Adarsh K. Gupta , Akshay Ashok , and Swapnil M. Parikh

**Abstract** In the growing world of digitization, digital media is engendered in abundance. With the ascension of the utilization of the Internet, there has been a prodigious increase in the engendering of digital content which includes images, audio, video, and documents such as pdf and text data. Information is free and more accessible than in any other era of humanity. Due to such a cognizance explosion, there is a vigorous need to make it more accessible. This can be achieved with semantic search. The quandary of processing, indexing, and storing such content has grown exponentially. At the same time, the infrastructure to handle such length has to be efficient and scalable. The current scenario of erudition explosion resulted in sizably voluminous data having a high performant scalable and resilient architecture which can parallelly process this multimodal binary file, can be gamely transmuting, and is becoming a requirement of the future. Different from the subsisting approaches that design handcrafted and task-concrete architectures for neural search to address only a single task, our architecture is tuned to handle multimodality which fundamentally denotes those data types (modalities) that can be audio, video, documents, images. This paper discusses the solution available to make digital content more accessible which is engendered as a result of the cognizance explosion. The proposed architecture will explore the domains of information extraction from this digital media securely and efficiently with various deep learning approaches for some categorical use cases.

**Keywords** Neural search · Information retrieval · Semantic search network

---

J. S. Gosaliya · A. K. Gupta · A. Ashok  
Babaria Institute of Technology, Vadodara, Gujarat, India

S. M. Parikh ()  
Parul Institute of Engineering and Technology, Parul University, Vadodara, Gujarat, India  
e-mail: [swapnil.parikh@gmail.com](mailto:swapnil.parikh@gmail.com)

## 1 Introduction

Twenty-first century has been called the cognizance age due to the ascension of engendering of digital content which has been made possible due to facile access to the Internet. There has been rapid digitalization across the industries. Information is free and more accessible than in any other era of humanity. Due to such an erudition explosion, there is a vigorous need to make it more accessible. This can be achieved with semantic search. In general, semantic search betokens as search with sense [4].

The general process related to semantic search can be divided as mentioned below:

1. **Media Cleaning and Transforming:** Today, a common trend is that most of the large-scale media is stored on cloud storage. Hence using a message broker and multiprocessing environment, one can download all the files in parallel and perform a series of format checks. Then, it can be categorized into media types and loading can be done on distributed file system (DFS). The DFS file id is then used for information retrieval (IR).
2. **Media Processing and Information Retrieval:** The pipelines are decentralized so that there is no single point of failure. The pipelines are basically stream processing queues that have a neural information retrieval model attached which acts as a consumer of the stream. Based on the file types and information retrieval algorithm, it performs the task of downloading the file from the DFS using file id, performs inference on the media file using the IR model, and then stores the extracted information in a document database.
3. **Indexing and Searching:** The traditional document database is not suitable for high performance and fast search; hence, one always indexes the stored document to elastic search which is a distributed, multitenant-capable full-text search engine. It stores the indexes into the main memory (RAM) and hence offers fast search to the input queries.

So nowadays to cope with such challenges, this research paper presents an infrastructure to take the semantics of the data for making a neural information retrieval system and later accommodating a semantic predicated probe which avails organizing the digital media assets efficiently and accurately. This paper endeavors to address a genuine-world scenario for all types of media files like documents, images, audio, and videos.

This research paper also utilizes the concept of multimodal for pdfs, jpg, mp3, etc., and making them searchable is the main challenge in this era. Not only this but it is withal an arduous task to integrate all the processing and probing systems under a mundane genuine-time backend infrastructure.

## 2 Literature Survey and Related Work

The systematic literature review always serves the purpose of exploring state-of-the-art approaches for neural information extraction, retrieval, and processing for multimodal neural search.

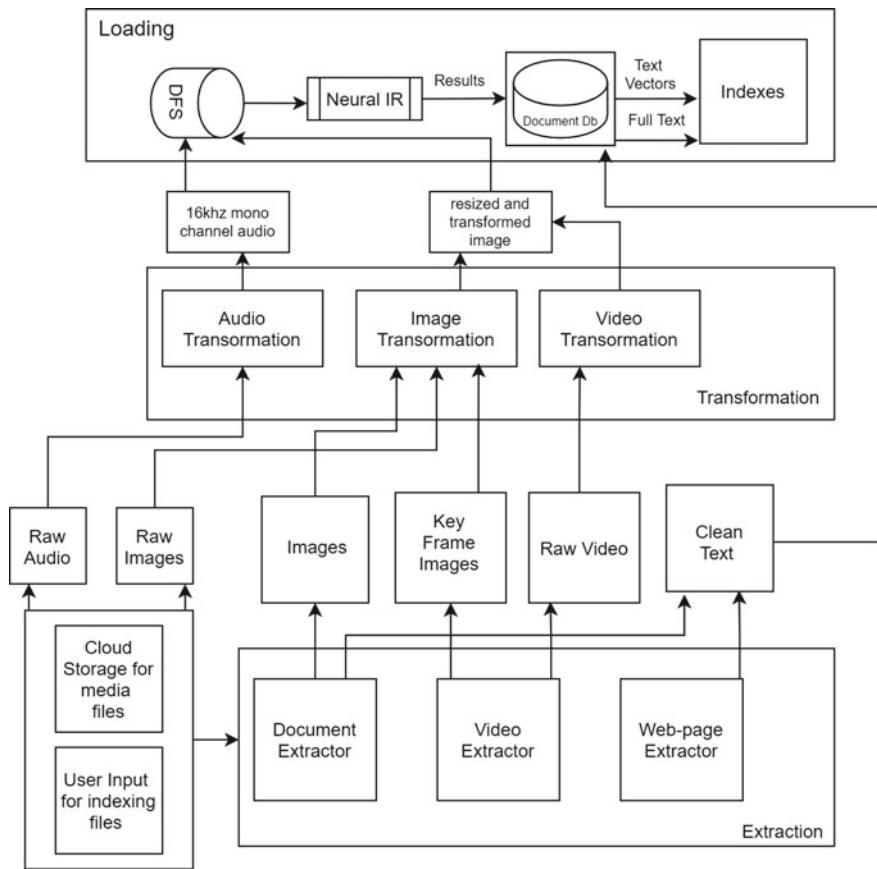
Authors [21] have reviewed various feature extraction techniques in content-based image retrieval. Authors focused on extracting color and texture features. Authors showed that Gabor wavelet transform is used during efficient discrimination of texture feature. Authors [2] have reviewed various text mining approaches and techniques like classification, clustering, and extraction. Authors have also discussed text mining approaches in healthcare and biomedical domain. Authors [1] showed that highly scalable and computationally efficient and consolidated information extraction approaches are in need for dynamic and unstructured big data. Authors told that quality, usability, sparsity, dimensionality, heterogeneity, scarcity, etc., are prominent challenges in information extraction field. Authors [7] proposed framework for semantic search engine through new ranking approach. The proposed framework can be further extended to detect more accurate semantic information from social network. Authors [16] presented a scalable real-time visual search system on JD.com's e-commerce platform. Authors have designed, implemented, and evaluated proposed approach through various optimization techniques. Authors [28] developed LOD backend infrastructure for research information of the social sciences. Authors have used GESIS use case for the same. Authors [25] had presented semEHR, a unified information extraction and semantic search system for obtaining clinical insight from unstructured clinical notes. This approach is ontology-based approach. semEHR is open source. Authors [12] reviewed various approaches or techniques related to Web data extraction and prepared comparative analysis for the same. Here, authors tried to identify efficiency of extraction process. Authors [27] invented a generalized deep multimodal neural architecture search (MMnas) framework for various multimodal learning tasks. To achieve this, authors had constructed a unified encoder-decoder backbone. Authors [17] introduced the multimodal multi-domain conversational dataset (MMConv). This dataset is fully annotated accumulation of human-to-human role-playing dialogues spanning over multiple domains and tasks. Authors also provided realistic user settings, structured venue database, crowd-sourced knowledge database as well as annotated image repository. Authors [24] proposed NaLa-Search, a novel semantic Web search and navigation architecture. It allows users to explore data stored in the LOD cloud through a multimodal, interaction-based (voice and touch) mobile application.

So as mentioned above, subsisting work majorly fixates on single media data extraction from the given text content but does not extract that from an image in pdf or ppt. So, our proposed solution tackles this circumscription by scaling up this to different file formats. The proposed architecture utilizes different state-of-the-art models on respective file types maximizing the quantity of index metadata for semantic search.

### 3 Proposed Architecture and Approaches

Figure 1 shows an ETL architecture in a nutshell which contains many miniature parts which are explained in further subsections. This diagram has three serving layers with specific purposes.

1. Extraction: This is the layer that majorly deals with the extraction of data from various sources and supplies it to the further pipelines for processing in our architecture. This is the first or gateway layer for our architecture, as everything commences from here. The extraction phase itself contains three main extracting processes.



**Fig. 1** Extract–transform–load

- (a) Document Extraction: Here, the data will be extracted from all the different document types like .doc, .ppt, .xls, .pdf, etc. The data which are extracted are generally of 2 types, extracting images and extracting text.
  - (b) Video Extraction : Here, the audio/video frame is extracted and stored separately.
  - (c) Web-page Extraction: Here, the HTML tree has been parsed to text. Then, this phase will extract text and image data from the Web sites.
2. Transformation: Transforming all the raw data and files from the extraction process and converting them into a normalized format and storing them utilizing an asynchronous task distributed system is the responsibility of the layer. This layer consists of 3 blocks: audio transformation, image transformation, and video transformation.
3. Loading: It is the heart of the implementation as all the important and heavy processes fall under this layer. All the information retrieval processing is performed with the output indexing at this layer. It also gives an abstraction layer for the searching of all the indexed data.

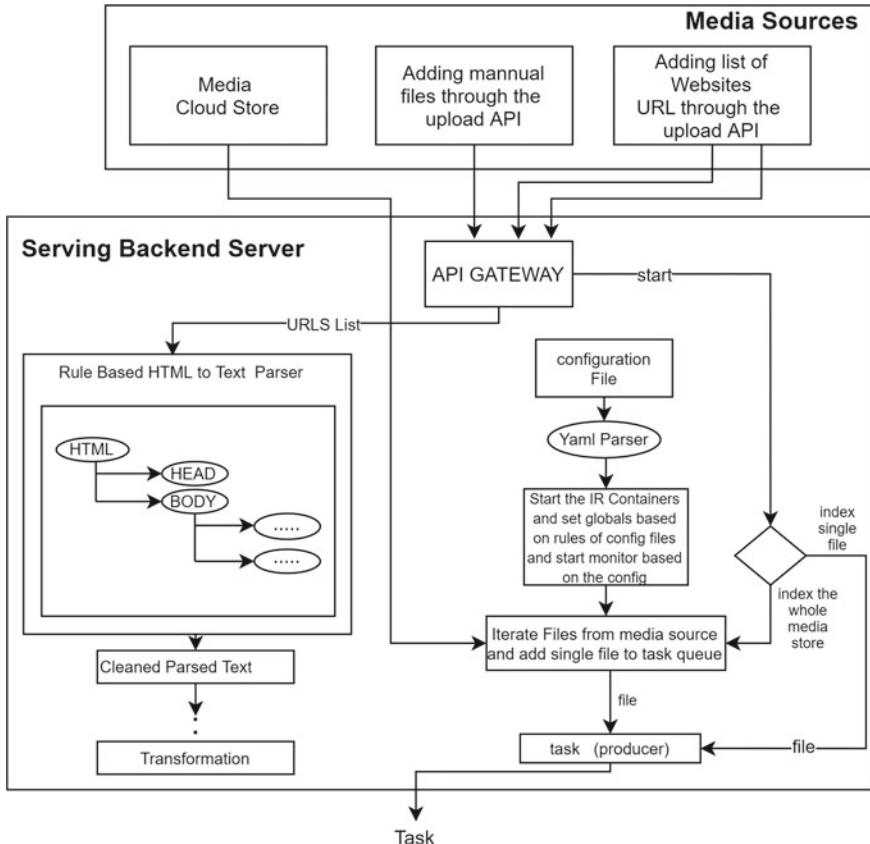
### ***3.1 Extraction Phase***

Figure 2 shows a bird's-eye view of the extraction phase along with the components of the gateway serving backend and its components.

The whole indexing process starts with defining a config.yaml file. This config file contains configuration for various file types and also contains configuration for choosing various deep learning-based information retrieval techniques for further processing. Additionally, information retrieval (IR) methods (e.g., for image–image captioning, OCR, scene recognition), etc., need to be performed. And inside it, of the mentioned IR methods which containers need to be started automatically and used as part of the pipeline. For further references, see <https://github.com/semantic-search/indexing-main/blob/master/config.yaml>

1. Media Sources: The entry point for whole ETL process is the API gateway. The source of the blobs can be a MEDIA Cloud store which is a storage bucket containing blobs within any organization. The source can also be list of Web site URL, which needs to be indexed. The source can also be a manual file upload via API call.
2. Serving Backend Server: As per the API endpoints, the URL, files from media store, and single files are passed to a necessary block.
3. Extract Web sites: Text is extracted from URLs by a rule-based HTML to text parser. Rules like Ignore URLs, ESCAPE SNOB, Ignore Images and sanitizes the text by replacing any unwanted spaces, new line characters, tab spaces etc.

In the case of media files, the backend either starts media extraction process by iterating the media store or receives a single file to extract.



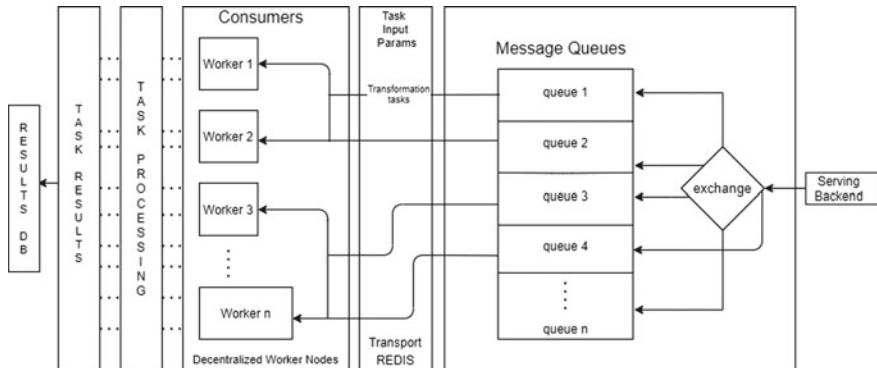
**Fig. 2** Extraction phase

In the document extraction, text and images are extracted from documents based on their respective parsers.

These images along with other images are parsed through the EXIF parsers, extracting the GPS coordinates, and getting their location info by calling OpenStreetMap Database and indexing the address info which will be used as an attribute during searching.

### 3.2 Transformation Phase

Figure 3 shows the first phase of transformation in our implementation. It is the main agent of the process with powerful asynchronous job queue used for running tasks in the background. It contains components and data structures like queues that connect



**Fig. 3** Transformation—distributed processing

the rear and front part of the architecture. Tasks which are the query or in a particular process that has to be executed, which are provided.

The following implementation consists of different internal blocks in its workflow.

1. Consumer Block: Contains worker nodes which processes tasks from queue.
2. Producer Block: Enqueues new task based on events in backend.
3. Task Queues: Task queues are used as a mechanism to distribute work across consumers.

In the architecture, the setup has been made in which the number of files denotes the number of tasks and each worker consisted of doing a finite dedicated process. So different processes don't make a worker rather a particular process becomes a worker and have their independent queue as well which had benefits like where one can get to load a particular process on our pipeline so when the resource management is done so it can be done optimally and wastage or scarcity of resources will not occur. Each worker will be hosting in different pods during deployment and have independent resources.

Number of files that are queued can be seen as

$$\text{Number of workers} = \text{Number of processes} \quad (1)$$

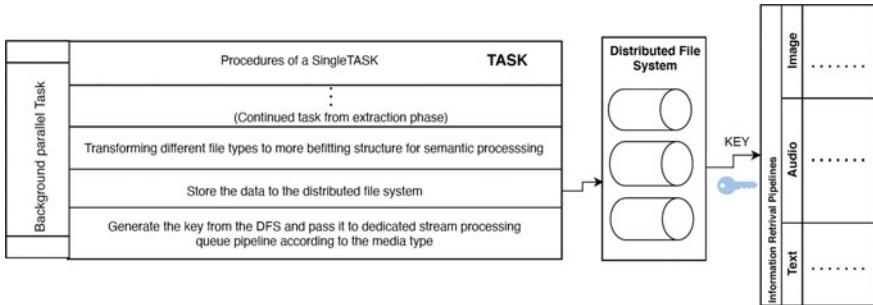
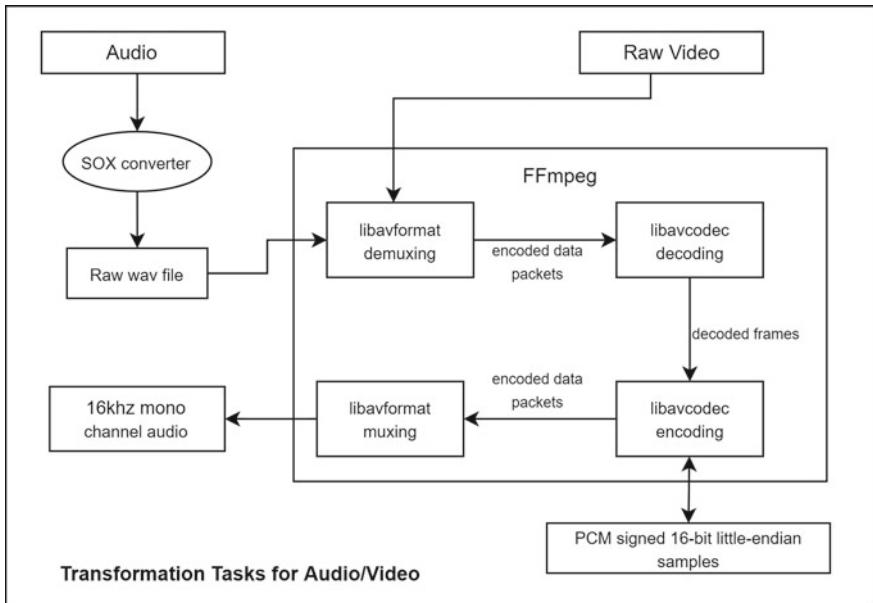
$$\text{Concurrency} = \text{System Capacity Dependent} \quad (2)$$

$$\text{Queued files} = \text{Number of files}/(\text{Number of workers} * \text{Concurrency}) \quad (3)$$

So, the remaining time can be calculated by this through multiplying queued files with constant time variable.

After a task is processed by the worker node, the output is stored into our database for future processing and access.

Figure 4 shows the summarized block view of our task which is processed by workers in the asynchronous job queue. The task originates in our extraction phase,

**Fig. 4** Transformation—task flow**Fig. 5** Transformation—task flow—audio/video

and in the transformation phase, its primary work is to send the correct files and documents to the distributed file system and from there to the dedicated pipelines which will be the genesis of the loading phase.

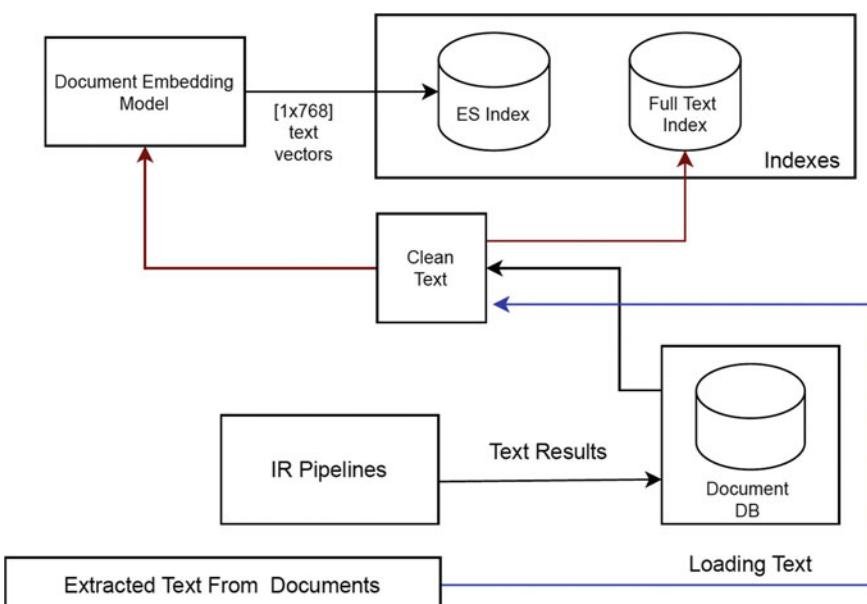
Figure 5 shows media transformation task flow for audio and video. The WAV audio format is mainly used through the IR pipelines by the deep learning models. So, our goal is to convert the incoming audio to the desired 16Khz mono-channel WAV audio format which is the requirement of the ORM models. Lib SoX is used to convert incoming audio data to WAV format. The WAV file generated by SoX needs to be converted to 16KHz mono-channel format, for that FFmpeg is used. FFmpeg is also used to extract audio data from incoming video data to WAV audio format.

### 3.3 Loading Phase

Figure 6 named loading architecture aims at indexing all the documents in elastic search index as well as in full-text search index. This is the final stage of our implementation. It is the continuation of our last loading process. Here, the data is coming to the loading process from information retrieval pipelines like image captioning results, image recognition results, audio transcribe results, and many more. So, these neural semantic extraction results are converted into text form. The text is then supplied to document embedding model and ultimately to index the text in elastic search and secondly to full-text index.

This process integrates text similarity search by adding vector fields into elastic search. The vectors are generated by passing each text statement from the cleaning process through a sentence embedding model which is Bert Large Cased model with 24-layer, 1024-hidden, 16-heads, 340 M parameters. This results in  $1 \times 768$  dimension dense vectors with dims matching with the BERT model of which are indexed into the elastic search.

So, survey had been done that BERT architecture was best suited for our implementation as this model can understand the meaning of each word based on context both to the right and to the left of the word; this represents a clear advantage in the field of context learning.



**Fig. 6** Loading phase—text

When the user enters a query, the query goes through the same embedding model, and a vector is generated. Then, the calculation of the cosine similarity between the ranked vectors of elastic search and the user query provides accurate semantic similar results.

This whole bulk-indexing process is done in parallel using Redis Queue for real-time queue-based indexing based on Redis DB with real-time monitoring and queue-based controls for the process.

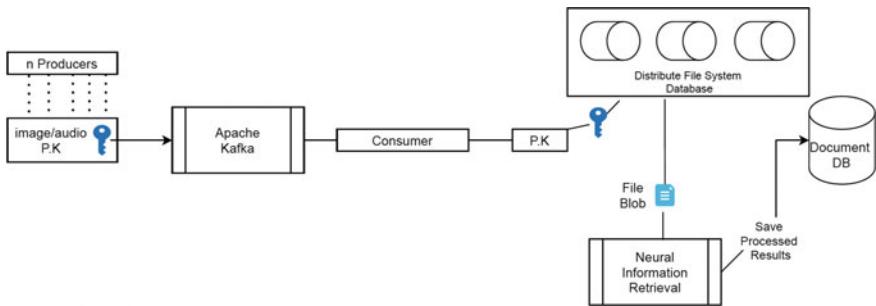
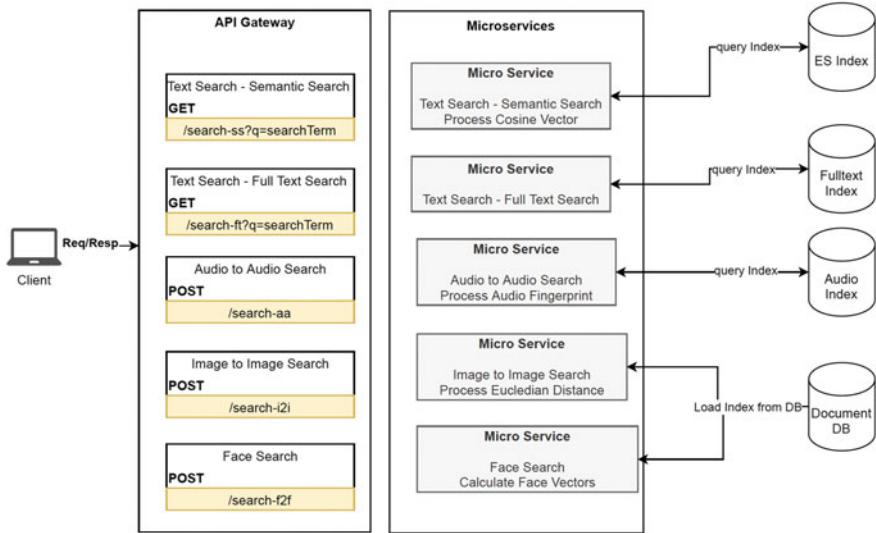
Secondly, for the subordinate full-text search index, the process is as follows. Incoming text and its metadata are stored in the full-text search index in the primary memory for fast retrieval of data. When the user enters a search query, first, a Levenshtein distance is calculated between the query and the text in the full-text search index, and the results with high scores are sent as a response. Due to this, these typo-tolerant search results are obtained.

```
{
  "settings": {
    "number_of_shards": 2,
    "number_of_replicas": 1
  },
  "mappings": {
    "dynamic": "true",
    "_source": {
      "enabled": "true"
    },
    "properties": {
      "doc_id": {
        "type": "text"
      },
      "text": {
        "type": "text"
      },
      "text_vector": {
        "type": "dense_vector",
        "dims": 768
      }
    }
}
```

Below is the explanation of the above configuration:

The text vector configuration is of dense vector type with dimensions matching with the BERT model ( $1 * 768$ ). Also, other elastic search cluster parameters are no. of shards and no. of replicas. Lastly, the dynamic mapping of ongoing data indexing is set to be true.

Figure 7 represents loading phase for media. In the working of the whole pipeline process, once the extraction and transformation stages are completed, it invokes the neural information retrieval pipeline by ingesting the DFS key by producing in the

**Fig. 7** Loading phase—media**Fig. 8** Search

appropriate APACHE KAFKA topic. The end connector which is the consumer of the IR pipeline and which has the deep learning code performs the information retrieval processes and stores the output into the database.

This whole process is asynchronous for all the IR methods belonging to a file type. The different topics of kafka are same as the method name and are taken config.yaml which helps the producers to send the key to the appropriate consumer based on the config.yaml.

### 3.4 Search Process

Figure 8 named search procedure starts from the client—a browser or any application that calls the REST API. The request from the client is passed to the API gateway where it is decided which micro-service is to be called based on the HTTP method and route of the REST API.

There are basically four kinds of search such as text search, audio search, image to image search, and face search. Indexing of data for all of the above search methods is a part of the neural information IR pipelines.

## 4 Category-Wise Summary of Implemented Models/Approaches

Table 1 provides the different deep learning models that are utilized in the paper for a particular extraction category. These algorithms and their corresponding models are openly available and widely used models. For the experimentation, all the information retrieval models have been integrated into a pipeline to the proposed architectures. Additionally, as a component of the open source, all these models are packaged into their separate containers with all the indispensable deep learning models and source files inside it and open-sourced the source code along with the containers in a public domain.

## 5 Experimental Results

Figures 9 and 10 illustrate experimentation results with various ways of search. All the information retrieval models have been implemented and integrated into a pipeline to the proposed architectures.

Figure 9 consists of three different implementations; the topmost implementation is of full-text search, shows its typo-tolerant search, and displays accurate results using them. Next in the middle is the elastic search implementation which brings relevant results of the query using the vector similarity search. The one, in the end, is the image captioning search which takes a text query as an input and searches the result of a query over the text consisting of words and sentences which were extracted using the image captioning model during the loading process. This gives us the relevant images according to query context.

Figure 10 also consists of three different implementations; the topmost implementation is of face search which returns us with pertinent images using the face recognition model, according to the input query image given. Next in the middle is the reverse image search which gives us similar images as results according to the image submitted. At the end, there is OCR search that matches the query text to the

**Table 1** Category-wise summary of implemented models/approaches

Extraction category	Implemented models/approaches
Image recognition	ResNet50 [10]
Image recognition	ResNet152 [10]
Image recognition	ResNet101v2 [10]
Image recognition	ResNet152v2 [10]
Image recognition	vgg16 [23]
Image recognition	vgg19 [23]
Image recognition	Inceptionv3
Image recognition	NASNetLarge [29]
Image recognition	MobileNet [11]
Image recognition	MobileNet large [11]
Scene recognition	Places365
Scene recognition	Places365
Scene recognition	places365kerashybrid [10]
Scene recognition	places365kerasbase [10]
Scene recognition	IBM Max Scene Classifier
OCR	EASY OCR [3]
OCR	Keras OCR [3]
Object detection	mask-rcnn-senet [9]
Object detection	YOLOv4 [5]
Object detection	yolov4-voc [5]
Object detection	ppyolo [20]
Object detection	openimages [14]
Object detection	ms-coco [19]
Object detection	retina-net [18]
Image search	XCEPTION MODEL [6]
Image captioning	Image captioning [26]
Image captioning	self-critical.pytorch [22]
Image captioning	IBM MAX image Caption (im2txt)
Speech to text	QUARTZNET [13]
Speech to text	Jasper speech to text [15]
Speech to text	Deep speech [8]
Sound classification	IBM MAX audio classifier
Audio fingerprint	Audio search

### Typesense Search

Art Public Speaking

The art of speaking (1763).pdf

art of speaking - Containing, an essay; in which are given rules for expressing properly the principal passions and humours, which occur in reading, or public speaking. And lessons taken from

art of speaking - Containing, an essay; in which are given rules for expressing properly the principal passions and humours, which occur in reading, or public speaking. And lessons taken from

art of eloquence.pdf

art of eloquence - a guide to effective speech with selected addresses of the authors; a governor and a scientist look at public speaking.

The art of speaking.pdf

The art of speaking - containing, an essay, in which are given rules for expressing properly the principal passions and humours, which occur in reading, or public speaking, and lessons, taken from

The art of speaking.pdf

The art of speaking. - Containing, I. An essay; in which are given rules for expressing properly the principal passions and humours, which occur in reading and public speaking; and II. Lessons taken

### ElasticSearch Search

How to impact people, how to earn companion and effect people

Human Hacking - Win Friends, Influence People, and Leave Them Better off for Having Met You.pdf

Summary - How to Win Friends and Influence People by Dale Carnegie.pdf

How to Analyze People : Use the Laws of Power - Analyze and Win Friends Using Subliminal Manipulation, Persuasion, Dark Psychology, Hypnosis, NLP Secrets, Body Language, and Mind Control Techniques.pdf

How Outlaws Win Friends and Influence People.pdf

How to Win Friends and Influence People : (Vermillion Classics).pdf

### Image Caption Search

blue and green bird



Name: peacock pic 2  
Score: 93%



Name: peacock pic 3  
Score: 91%



Name: peacock pic 4  
Score: 82%



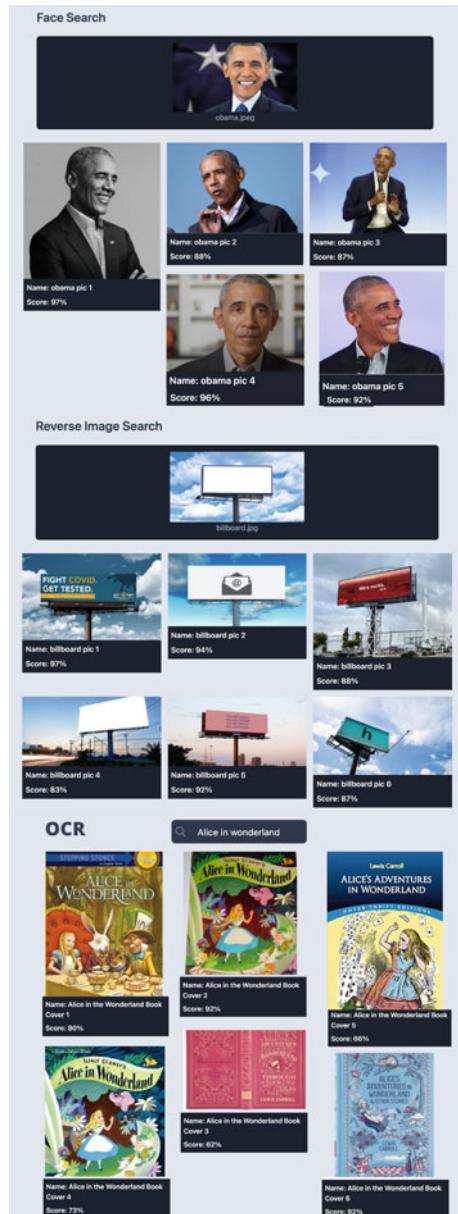
Name: peacock pic 5  
Score: 86%



Name: peacock pic 4  
Score: 82%

**Fig. 9** Typesense, elastic, and image caption search

**Fig. 10** Face, reverse image, and OCR search



OCR text (extracted from all the images during the loading phase) in the full-text search index and gets results according to the text present in the images along with typo-tolerance.

## 6 Conclusion and Future Work

With the ascension utilization of the Internet, plenty of digital content/media has been engendered. Due to such a cognizance explosion, there is a vigorous need to make it more accessible. This can be achieved with neural semantic search. This paper gives a novel architecture cognate to the whole extract transform load (ETL) process and information retrieval. The proposed architecture is robust enough to let the utilizer decide the data source which needs to be indexed; its cloud provider agnostic thus provides more flexibility. The architecture can be custom-tailored to the Desiderata by designating which file types or models need to be utilized for processing the data source. Different from the subsisting approaches that design handcrafted and task-concrete architectures to address only a single task, our architecture can be generalized to automatically engender optimal architectures of different tasks.

This paper further provides a summary table of all the approaches utilized in this ETL process. The experimental search result shows the cessation-to-end working of the novel architecture along with the precision of the results and all the respective model. The implementation of the proposed architecture can be found at <https://github.com/semantic-search>; it contains the model architecture with all the neural information retrieval pipelines with the coupling code and withal the faculty to integrate your own neural information retrieval pipeline model along with the facility of customization of culled IR models utilizing the config mechanism as discussed above in the paper. Adscititiously, it withal contains a mechanism to explicitly mention the source of blob store on any cloud, for example, Google Cloud or AWS. These cloud storages are majorly used as the primary blob stores in most backend infrastructures. Hence, the architecture in this paper provides flexibility to the user to integrate search as a service in the real-time application on top of their existing blob store, which allows anyone to explore relevant content at scale. Providing a deeper understanding of the intent of the search query significantly improves search results as compared to the traditional search methods using the approaches mentioned in this research paper. Hope that this research work may serve as a solid baseline to inspire future research on multimodal neural search.

## References

1. Adnan, K., Akbar, R.: An analytical study of information extraction from unstructured and multidimensional big data. *J Big Data* **6**(1), 1–38 (2019)
2. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B., Kochut, K.: A brief survey of text mining: Classification, clustering and extraction techniques. [arXiv:1707.02919](https://arxiv.org/abs/1707.02919) (2017)
3. Baek, Y., Lee, B., Han, D., Yun, S., Lee, H.: Character region awareness for text detection (2019)
4. Bast, H., Björn, B., Haussmann, E.: Semantic search on text and knowledge bases. *Foundations Trends Inf Retrieval* **10**(2–3), 119–271 (2016)
5. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection (2020)
6. Chollet, F.: Xception: Deep learning with depthwise separable convolutions (2016)
7. El-gayar, M., Mekky, N., Atwan, A.: Efficient proposed framework for semantic search engine using new semantic ranking algorithm. *Int J Adv Comput Sci Appl* **6**(8) (2015)
8. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., Ng, A.Y.: Deep speech: Scaling up end-to-end speech recognition (2014)
9. He, K., Gkioxari, G., Dollr, P., Girshick, R.: Mask r-CNN (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)
11. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications (2017)
12. Kamanwar, N., Kale, S.: Web data extraction techniques: A review. In: 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave). pp. 1–5. IEEE (2016)
13. Kriman, S., Beliaev, S., Ginsburg, B., Huang, J., Kuchaiev, O., Lavrukhin, V., Leary, R., Li, J., Zhang, Y.: Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions (2019)
14. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krashin, I., Pont-Tuset, J., Kamali, S., Popov, S., Mallochi, M., Kolesnikov, A., Duerig, T., Ferrari, V.: The open images dataset v4: unified image classification, object detection, and visual relationship detection at scale (2018). <https://doi.org/10.1007/s11263-020-01316-z>
15. Li, J., Lavrukhin, V., Ginsburg, B., Leary, R., Kuchaiev, O., Cohen, J.M., Nguyen, H., Gadde, R.T.: Jasper: An end-to-end convolutional neural acoustic model (2019)
16. Li, J., Liu, H., Gui, C., Chen, J., Ni, Z., Wang, N., Chen, Y.: The design and implementation of a real time visual search system on jd e-commerce platform. In: Proceedings of the 19th International Middleware Conference Industry. pp. 9–16 (2018)
17. Liao, L., Long, L.H., Zhang, Z., Huang, M., Chua, T.S.: Mmconv: An environment for multimodal conversational search across multiple domains. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 675684. SIGIR '21, Association for Computing Machinery, New York, NY, USA (2021). 10.1145/3404835.3462970, <https://doi.org/10.1145/3404835.3462970>
18. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollr, P.: Focal loss for dense object detection (2017)
19. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollr, P.: Microsoft coco: Common objects in context (2014)
20. Long, X., Deng, K., Wang, G., Zhang, Y., Dang, Q., Gao, Y., Shen, H., Ren, J., Han, S., Ding, E., Wen, S.: Pp-yolo: An effective and efficient implementation of object detector (2020)
21. Patel, J.M., Gamit, N.C.: A review on feature extraction techniques in content based image retrieval. In: 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET). pp. 2259–2263. IEEE (2016)
22. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning (2016)

23. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014)
24. Snchez-Cervantes, J.L., Alor-Hernndez, G., Paredes-Valverde, M.A., Rodrguez-Mazahua, L., Valencia-Garca, R.: Nala-search: A multimodal, interaction-based architecture for faceted search on linked open data. *Journal of Information Science* **0**(0), 0165551520930918 (0). <https://doi.org/10.1177/0165551520930918>, <https://doi.org/10.1177/0165551520930918>
25. Wu, H., Toti, G., Morley, K.I., Ibrahim, Z.M., Folarin, A., Jackson, R., Kartoglu, I., Agrawal, A., Stringer, C., Gale, D., et al.: Semehr: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *Journal of the American Medical Informatics Association* **25**(5), 530–537 (2018)
26. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention (2015)
27. Yu, Z., Cui, Y., Yu, J., Wang, M., Tao, D., Tian, Q.: Deep multimodal neural architecture search. In: Proceedings of the 28th ACM International Conference on Multimedia. p. 37433752. MM '20, Association for Computing Machinery, New York, NY, USA (2020). <https://doi.org/10.1145/3394171.3413977>, <https://doi.org/10.1145/3394171.3413977>
28. Zapilko, B., Boland, K., Kern, D.: A lod backend infrastructure for scientific search portals. In: European Semantic Web Conference. pp. 729–744. Springer (2018)
29. Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition (2017)

# A Comparative Study Analysis on Air Monitoring and Purification Systems



Nida Praveen, Lipika Goel, and Sonam Gupta

**Abstract** Industrial activities and vehicles are the foremost cause of degradation of air quality (AQ) level. The main cause is the emission of harmful gases or air pollutants, solid particles, smoke, and dust from these industries. Air pollutants can be smog, nitrogen oxides, carbon monoxide, carbon dioxide, ammonia, sulphur dioxide, and other pollutants. Outdoor as well as indoor AQ level is affected by these types of pollutants. Other case of indoor AQ level degradation may be the release of air-borne chemicals from the furniture and during cooking. This degraded air that is polluted air is the main (AAFSs). The aim of this paper is to perform a comparative study analysis on the research that have cause of disease in humans. Hence, there is a need of indoor automatic air filtration systems in order to generate the lucrative AAPSs in last few years, thereby controlling the AQ level.

**Keywords** Air quality · Machine learning · Cloud computing · IoT

## 1 Introduction

Basically, air is the combination of various gases that are present in a fix ratio to compose the pure air. However, any imbalance in the ratio of gases may cause to excessive harm to the life of living beings on the Earth. This imbalance may occur due to either the mixture of new harmful gas in air or by the increment of any particular gas in air, such types of gases are called contaminants in air. Various contaminants may be smog, nitrogen oxides, carbon monoxide, carbon dioxide, ammonia, sulphur dioxide, and other pollutants [1]. The exposure of these harmful particles in the atmosphere causes various chronic health diseases like respiratory infections and lung cancer. This problem has been overcome by various researchers in recent years that promote the real-time monitoring of unfavourable situations occurring in the air; hence, this is a concerning research topic as various methods are proposed to monitor AQ in the real-time scenario. Due to these adversarial effects, air purification is also the

---

N. Praveen · L. Goel · S. Gupta (✉)  
Ajay Kumar Garg Engineering College, Ghaziabad, India  
e-mail: [guptasonam6@gmail.com](mailto:guptasonam6@gmail.com)

seeking topic in research field. Also, government is looking for the challenging ways to threat the air pollution. Various simple air filtration systems have been developed by simply putting the filter behind a fan. However, until the AQ is measured, respective actions cannot be taken on the basis of bad or good AQ levels. The other method adapted to purify the indoor air is the ventilation facilities in buildings. This type of AAFSs works by simply exchanging the outdoor and indoor air. Conjunction to these ventilations, there can also be openings in walls which permit inclusion of fresh air from outside into indoors, thus can be behaved as infiltration walls. It can be said that proper ventilation in conjunction with AFSs can be a good solution to tackle the problem of degrading AQ. Poor AQ can also cause poor mortality rate as polluted air causes various diseases in any normal person's body that can damage the functioning on life giving organs of human body such as lungs [2]. Moreover, modern lifestyle is also causing poor indoor AQ [3] as people are using air conditioners and heating systems which leads to sick building syndrome (SBS). In order to increase morbidity and mortality rate, various researchers are proposing different air purification systems as this is the serious issue to be considered. Additionally, since the outbreak of life taking COVID-19, AQ monitoring is a serious concern. Therefore, it is a worldwide topic to be researched. Remaining of the paper is organized as follows. Section 2 presents the survey performed by us on various technologies and countermeasures proposed by various researchers in last 7–8 years. Section 3 gives the comparative analysis of the studies along with the gaps in the studies proposed. Section 4 highlights the number of studies, focusing on functioning of the proposed system such as AQ monitoring, AQ monitoring, and AAFS. Lastly, Sect. 5 concludes the paper followed by some proposed future researchers directions in Sect. 6.

This paper is aiming to perform a comparative study analysis on the research that have been performed in order to generate the lucrative AAPSs in last few years.

## 2 Literature Survey

Degrading AQ is the serious concern in last few years; hence, AAFSs is trending topic among most of the researchers. Some of the research that have been proposed to develop air purifiers/AAFSs/Air purification systems are discussed in this section.

In this paper, the authors propounded air purification system (APS) having AQ monitoring systems. The system was designed for indoors [4]. In this paper, the researcher propounded an AQ monitoring system. They collected the real-time AQ data using microcontroller (Arduino) and stored the data in the database located at remote server. This data was updating as per the time and also this was accessible by any user via an active Internet connection from the server [5]. The author propounded AQ monitoring model based on wireless gas sensors. The proposed system was able to detect the harmful compounds like  $\text{CH}_2\text{O}$ ,  $\text{CO}_2$ , and CO. They implemented the smartphone application to show the real-time sensed data. The system was designed in the way that whenever the sensed data reaches the threshold value, alert message was displayed on the user's screen via Bluetooth and Wi-Fi [6].

The author propounded IoT-based indoor AAFS. They utilized the indoor plants, Chlorophytum comosum and Sansevieria trifasciata, to develop the proposed AAFS. The proposed system was less costly, and also, less maintenance was required to grow the plants. The plants were put in houses, offices, and small buildings, and testing was performed on the basis of before and after AQ checking in the surrounding atmosphere. The proposed study concluded that the two plants were a cheap solution to purify the air at a great extent [7].

The researcher designed an intelligent and controlled AAFS and sweeper robot. They presented the software and hardware design of the proposed control system. The proposed system was designed for a control system that was controlling the door (to be open or close) on the basis of the dust absorption in air utilizing the stepper motor [8]. In this paper, the author utilized activated carbon fibre to model an AAFS. They combined metallic collection rod with fibre brush, and fibre sheet, in order to generate the proposed model. Fibre sheet applied electrostatic force that collects the charged particles on collection rods, and those charged particles absorbed the harmful gaseous particles from the air, thereby cleaning the air. The model's performance was compared to the HEPA purifier. The comparative results of the proposed model and the HEPA filter showed that the proposed method outperformed by 35% accuracy [9].

In this paper, the author propounded a cheap sensing system to monitor the AQ. They studied the actual-time concentration of different harmful gaseous particles. They deployed the proposed system in several cities and collected the sensors data. They saved the collected data in uSense database that was accessible via different users through Internet. This system was mainly deployed to make people aware about the surrounding's AQ, so that preventive actions can be taken if necessary. However, the proposed system was only to check the AQ and could not give any purification system [10].

The researcher propounded an AQ monitoring-enabled AAFS. AQ monitoring task was performed utilizing gas sensors [11]. A threshold was set to trigger the alarm that was displayed on the LCD display, if the AQ pollution level exceeds the threshold value. Thus, the air filtering process was enabled to purify the air [12]. The author utilized AHP technology to select the best AAFS. They proposed three alternatives: Bap706, Bap600, and Bap1700 to choose the best AAFS for urban areas. The best purifier was selected to be Bap1700 that was chosen by the panel of four members [13].

In this paper, the researcher propounded an indoor AAFS having mobile pedestrian tracking capability. This system can track the pedestrian and monitor various pollution causing particles in air around the pedestrian and thereby can purify the air accordingly. The performance of the system was assessed using various algorithms. Pedestrian tracking was performed by utilizing Cam Shift algorithm, and Harr + AdaBoost algorithm was used to check the air quality, thus implementing the mobile AAFS [14].

Researcher propounded a negative ion-based AAFS. They developed a negative ion production circuit having 8 and 4 kV outputs; this generated voltage was used for air ionization process. These generated ions reacted with dissipate gas, thereby

cleaning the air. The performance was verified by preparing a prototype and circuit analysis [15].

The author [16] performed a study on use of mechanical air filter utilized for indoor polluted air. They showed that high efficiency particulate air (HEPA) AP is most widely used while purifying air due to its efficient pollution removing capability that is 99.97%, when the size of polluting particles is not more than  $0.3\text{ }\mu\text{m}$  (diameter). The author [17] propounded an AQ sensing and controlling system that was specially designed for health centres facing the issues caused by polluted air. They applied the data analytics on the sensed data about the concentrations of  $\text{CO}_2$ , temperature, and the  $R_H$  value to analyse the trends in AQ. Based on the sensed data, the control mechanism was able to control the speed of AAFS. They utilized ZigBee technology as the communication channel. Based on three different actions, the complete software was divided into respective sections as: (a) data monitoring system; (b) AQ analysing system; (c) application and purification system.

The author [18] propounded a sensor-based AQ monitoring system. The proposed system was implemented by distributing the sensors in a particular area and information was exchanged using serial bus communication. The researcher [19] proposed real-time AQ monitoring, and purification system specifically designed for varied particulate matters. They performed air purification using five-stage filtering using sterilization cotton filters, HEPA, activated carbon, cold catalyst, and fine dust filters. However, the system was particularly designed for smaller areas only, and also, no method for purification time has been estimated. Researcher [3] proposed indoor AQ monitoring, and analytics system especially designed from the perspective of COVID-19 outbreak. They used IoT sensors for sensing eight types of pollutants in clinical lab and utilized machine learning models to classify indoor AQ. However, no purification method was proposed, only analytics has been done on AQ recorded data. (SBC) network with the rate of 20KBPS to 1MBPS. The hardware thus accurately monitored the AQ in less time and more efficiently.

### 3 Comparative Analysis and Gap Study

Table 1 shows the comparative analysis of the studies that have been discussed in this paper.

### 4 Analysis

Above studies have been analysed, and the results are represented in Fig. 1. From the graph, it can be seen that only 40% (total 6) studies have only performed AQ monitoring process, 33% (total five) studies have performed the air purification along with the monitoring process; also, only 7% (total 1) study has performed the data analytics on the trend of AQ, however, could not provide the good air purification

**Table 1** Comparative analysis and gap study

Year	Summary	Technology used	GAP study	Authors(s)	References
2019	Development of an effective indoor air purification system. Works effectively for particulate matter metrics	Based on micro-controllers and sensors	Manual starting of air purifier. Not effective on coarse particles	Marinov et al.	[4]
2018	An app was designed as an interface to alert the user when the pollution (gas) concentrations were above a certain level	Mobile app (Wi-Fi-based), pollution and data monitoring using gas sensors	Only data monitoring, pollution monitoring, and alarm generation but not any purification method	Huang et al.	[6]
2018	Utilized the cloud technology to analyse the quality of air in real-time. Usage of PPM metric. Wi-Fi needed to access Internet to access the data from the environment	Data recording on Excel sheet, remote server monitoring using microcontroller (Arduino)	Only data monitoring and pollution monitoring, but not any purification method	Okokpujie et al.	[5]
2018	Development of a drone to monitor the quality and thereby purifying the air when needed	Sensors and filters embedded drones	Effective for indoors only; however, swarm bots were proposed for model development for larger space	Parvatekar et al.	[7]
2018	Development of a natural air purification system using indoor plant, thereby cheap	IOT based system utilizing indoor plants-Chlorophytum comosum and Sansevieria trifasciata	Not feasible for corporate office buildings, needed proper ventilation	Shitole et al.	[8]

(continued)

**Table 1** (continued)

Year	Summary	Technology used	GAP study	Authors(s)	References
2017	An automatic system based upon electrostatic forces, development of an air purifier and sweeping machine	Dust sensor, stepper motor, and control system	Biased in dust concentration measurement than air purification	Qijun et al.	[9]
2017	Based on novel electrostatic forces gave better and faster results	An automatic system based upon electrostatic forces (ACF, metallic collection rod, carbon fibre brush)	High power consumption	Kim et al.	[12]
2017	Design and development of a semi-automated air purification system gave buzzer when pollution levels exceed a threshold point	Arduino microcontroller-based gas sensors	Non-automated air purifier was used, manual starting of air purifier	Sharma et al.	[13]
2017	Comparison among the three air purification systems and found out the best of three	Analytical hierarchy process (AHP)	Lack of suggestions in circuitry design	Delgado et al.	[14]
2016	An improvement for outdoor air quality (AQ), hence a good approach to track AQ for pedestrian	Kalman filter, CamShift algorithms, AdaBoost	Not an efficient approach for outdoor air purification system	Fan et al.	[10]
2015	To initiate the program to make people aware the air quality, which was also a benefit for local government	<i>uSense</i> database and gas sensor	Continuous looking into <i>uSense</i> domain, but not any alert system was implemented	Brienza et al.	[15]

(continued)

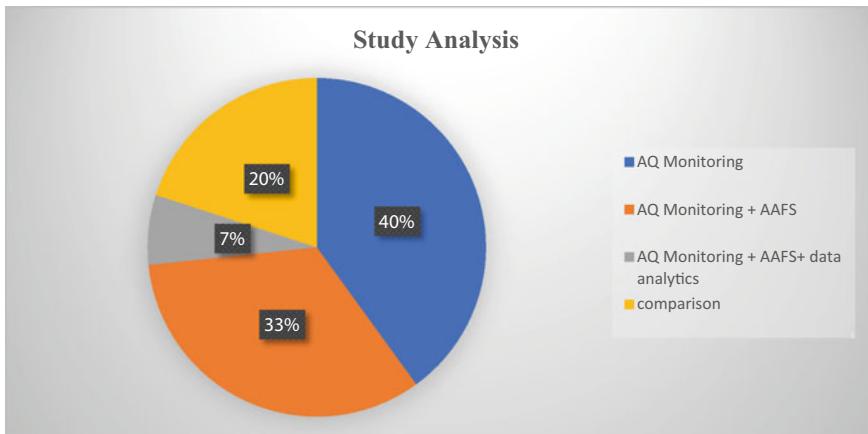
**Table 1** (continued)

Year	Summary	Technology used	GAP study	Authors(s)	References
2013	Works on the basis of reaction of active and exhaust air particles	-ve ion generating circuitry design	Efficiency levels not included in results	Lai et al.	[16]
2015	AQ sensing and controlling system that was specially designed for health centres facing the issues caused by polluted air	Data analytics using Auto regressive integrated moving average (ARIMA)	Not an efficient method of Air purification	Yu and Lin	[17]
2020	Real-time AQM and purification system specifically designed for varied particulate matters	MQ135 for data collection and air purification using five-stage filtering. Sterilization cotton filters, HEPA, activated carbon, cold catalyst, and fine dust filters have been utilized for the same	System is particularly designed for smaller areas only and also no method for purification time has been estimated	Niranjan and Rakesh	[18]
2021	Indoor AQM and analytics system especially designed from the perspective of COVID-19 outbreak	Use of IoT sensors for sensing eight types of pollutants in clinical lab. Use of machine learning models to classify indoor AQ	No purification method was proposed, only analytics has been done on AQ recorded data	Mumtaz et al.	[19]

results, and rest 20% (total 3) of the studies have compared the different air purification methods. Thus, it can be inferred that there is not such system that as-a-whole can be counted as a good system having all the facilities like AQ monitoring, data collection, alerting facility, Web-based access, air purification, and data analytics on trends on air quality.

## 5 Conclusion

This paper aimed to provide a detailed study on AAFSs. From the papers presented in this study, it is inferred that development of AAFSs mainly focus on the indoor pollutants like: nitrogen oxides, carbon monoxide, carbon dioxide, and more powered



**Fig. 1** Study analysis

is required to create the electrostatic force of attraction to attract the harmful gaseous particles to clean the air. Also, not all the research have the monitoring systems, and some were also lacking in good air purification system, but provide only monitoring. Hence, the system that can monitor the AQ can store the data in databases to provide access to user to monitor data to perform the air purification task and also to predict the future possibilities of air pollution in any particular area would be a tough compatible solution to the degraded AQ problem.

## 6 Future Research Directions

Based on the survey performed in this work, we can suggest some future research directions that can aid the respective researchers.

- Developing more robust and efficient system that can generate instant alert via mail or SMS in the situation when AQ reaches the threshold level.
- Calculation of estimated time that can be taken by the purifier to clean the environment in which the system is deployed based upon the values collected of AQ parameters.
- Development of AQM systems for real-time environment in spite of just for lab environments.
- More development of mobile applications that can be used by the common, non-technical people as well.
- Automatic switch off and on system based on the quality parameters' value of the deployed environment.

## References

1. Roegiers, J., Denys, S.: Development of a novel type activated carbon fiber filter for indoor air purification. *Chem. Eng. J.* **417**, 128109 (2021)
2. Cheek, E., et al.: Portable air purification: review of impacts on indoor air quality and health. *Sci. Total Environ.* **766**, 142585 (2021)
3. Shakya, S.: A self monitoring and analyzing system for solar power station using IoT and data mining algorithms. *J. Soft Comput. Paradigm* **3**(2), 96–109
4. Marinov, M.B., Iliev, D.I., Djamiykov, T.S., Rachev, I.V. Asparuhova, K.K.: Portable air purifier with air quality monitoring sensor. In: 2019 IEEE XXVIII International Scientific Conference Electronics (ET), Sozopol, Bulgaria, 2019, pp. 1–4. <https://doi.org/10.1109/ET.2019.8878570>
5. Okopujie, K., Noma-Osaghae, E.: A smart air pollution monitoring system. *Int. J. Civ. Eng. Technol.* (2018)
6. Huang, C. et al.: A multi-gas sensing system for air quality monitoring. In: 2018 IEEE International Conference on Applied System Invention (ICASI), Chiba, 2018, pp. 834–837. <https://doi.org/10.1109/ICASI.2018.8394393>
7. Parvatekar, K.V., Zacharia, S.M., Sheelvant, S.V., Nanaiah T., Ambika, K.: EnviDron—a drone that purifies air. In: 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT), Bangalore, India, 2018, pp. 614–618. <https://doi.org/10.1109/ICGCIoT.2018.8753058>
8. Shitole, S., Nair, D., Pandey, N., Suhagiya, H.: Internet of Things based indoor air quality improving system. In: 2018 3rd International Conference for Convergence in Technology (I2CT), Pune, 2018, pp 1–4. <https://doi.org/10.1109/I2CT.2018.8529813>
9. Qijun, X., Chaoying, L., Yifang, L., Wei, H., Zhonghui, L.: Design of control system for an intelligent air purifier and sweeper combined robot. In: 2017 12th IEEE Conference on Industrial Electronics and Applications (ICIEA), Siem Reap, 2017, pp. 227–230. <https://doi.org/10.1109/ICIEA.2017.8282847>
10. Fan, L., Hongdou, C.: Indoor mobile biological air purifier with pedestrian tracking system. In: Chinese Control and Decision Conference (CCDC), 2016
11. Madhura, S.: IoT based monitoring and control system using sensors. *J. IoT Soc., Mob., Anal., Cloud* **3**(2), 111–120 (2021)
12. Kim, H., Han, B., Woo, C.G., Kim, Y., Lim, G., Shin, W.G.: Air cleaning performance of a novel electrostatic air purifier using an activated carbon fiber filter for passenger cars. *IEEE Trans. Ind. Appl.* **53**(6), 5867–5874. <https://doi.org/10.1109/TIA.2017.2745499>
13. Sharma, M., Kumar, A., Bachhar, A.: I2P air purifier with air quality monitoring device. In: 2017 2nd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, 2017, pp. 478–481. <https://doi.org/10.1109/CESYS.2017.8321326>
14. Delgado, A., Flor, H.: Selection of the best air purifier system to urban houses using AHP. In: 2017 CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), Pucon, 2017, pp. 1–4. <https://doi.org/10.1109/CHILECON.2017.8229622>
15. Brienza, S., Galli, A.: Low-cost sensing system for cooperative air quality monitoring. *Sensors*, pp. 12242–12259 (2015)
16. Lai, C.M., Pan, M.H.: Ionization based air purifier. In: Consumer Electronics (ISCE), IEEE 17th International Symposium, 2013
17. Vijayan, V.K., Paramesh, H., Salvi, S.S., Dalal, A.A.K.: Enhancing indoor air quality, the air filter advantage. *Lung India* **32**(5), 473–479 (2015)
18. Yu, T.C., Lin, C.C.: An intelligent wireless sensing and control system to improve indoor air quality: monitoring, prediction, and preaction. *Int. J. Distrib. Sens. Netw.* **2015**, 140978 (2015)
19. Pillai, M.A., Veerasingam, S., Yashwanth, S.D.: Implementation of sensor network for indoor air quality monitoring using CAN interface. In: 2010 International Conference on Advances in Computer Engineering, Bangalore, 20–21 2010

# Malicious Node Detection and Prevention for Secured Communication in WSN



Vaibhav Dabhade and A. S. Alvi

**Abstract** Wireless sensor networks (WSN) are composed primarily of resource-constrained sensor nodes, different monitoring characteristics and terminal nodes. Such types of infrastructure have been used in areas such as health care, defense, agricultural sectors, and emergency management, as communications systems and intrusion detection. Due to various growing use of wireless sensor networks, essential information is shared in an insecure channel among network entities including sensors, communication gateways, participants, etc., and the existence of critical and confidential information in the network emphasizes the difficulty of security risks. Through this work, we present the Hybrid Pairwise Key Establishment Scheme (HPKE) for connected devices with large scale sensors. This work often deals with the shortest path of computing among two nodes using broadcast tree construction (BTC) that optimizes the usage of internal resources and the consumption of energy. The intrusion detection scheme (IDS) recognizes certain possible factors of detecting malicious, potentially unreliable nodes during communication between nodes. Eventually, we will implement some dynamic application in network simulation environment of proposed system.

**Keywords** Wireless sensor network · BTC · Hybrid Pairwise Key Establishment Scheme

## 1 Introduction

A wireless sensor network (WSN) typically comprises several wireless small, low-power, and low-cost wireless network nodes. Every sensor node is powered by a

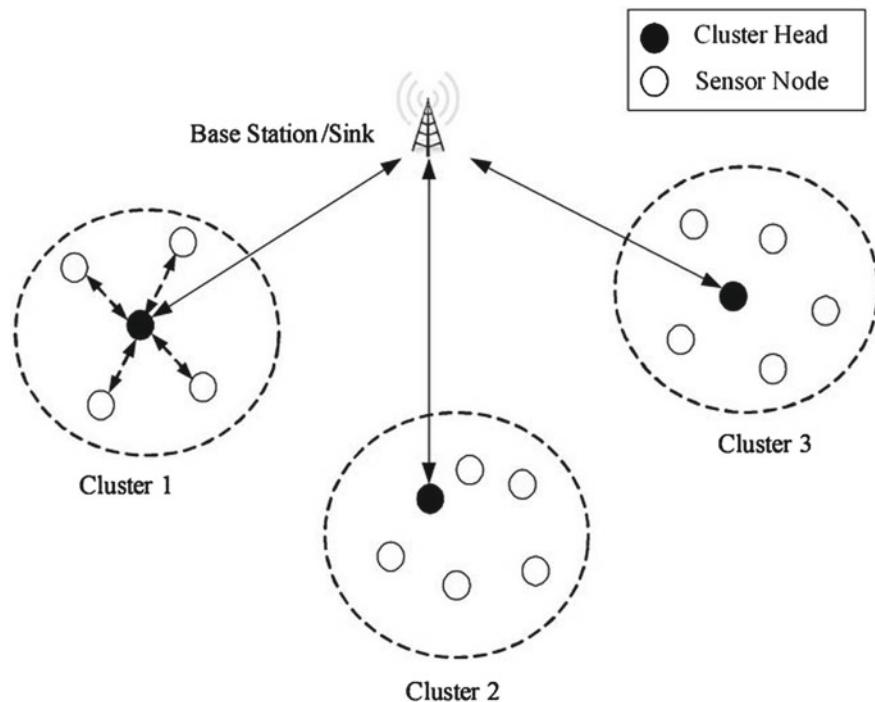
---

V. Dabhade (✉)  
Computer Engineering, MET's BKC, IOE, Nashik, India  
e-mail: [vaibhavd\\_ioe@bkc.met.edu](mailto:vaibhavd_ioe@bkc.met.edu)

A. S. Alvi  
Information Technology, Prof. Ram Meghe Institute of Technology & Research, Badnera, Amravati, India

battery, with local storage of data process technology. Sensor devices can communicate at a small distance, wireless communication. Wireless sensor networks are helpful in different scenarios, such as gathering information to a remote topography or tracking a hostile situation, due to low overhead computational and communication costs. Two negotiating parties need to share a hidden key in the symmetric-key practice before communicating with each other. Because of the volatile network topology, safely and efficiently transmit the encrypted data in the sensor nodes is a significant challenge when designing a wireless sensor network. Several quick redistribution schemes have been examined to solve wireless communication's critical quality problem in recent studies. The basic concept here is to pre-load a collection of symmetric keys onto the sensor network before deploying them. After implementation, if two nodes share one standard key, they can be communicated with each other. Most of the strategies suggested are either focused on the theories of probability and random graphs or based upon polynomial bivariate. As we discussed in the literature, that scheme has its limitation and cannot attain both the protection and efficiency specifications of wireless sensor networks simultaneously (Fig. 1).

This work suggests an enhanced pairwise robust authentication scheme for wireless sensor networks to tackle current critical pre-distribution frameworks' weaknesses. Our system is effective against selective forwarding attacks compared to



**Fig. 1** Cluster network in WSN for secure data transmission

existing approaches. Our scheme's other advantages include low overhead storage, full network access, large network scale, poor communication, and computational overheads. The framework is also capable of detecting malicious activities using the IDS technique and the minimal use of resources using the nearest neighbor approach.

The major contribution of paper is illustrated as below,

- To develop a poetization-based cluster head (CH) selection in WSN.
- To design and develop a cluster-based WSN for data aggregation by CH to reduce the data redundancy.
- To develop an algorithm for detection malicious and network attacks such wormhole, sinkhole, sybil attack, etc.

Moreover, the further sections of the paper are divided as follows: Sect. 2 describes various existing methods done by previous researchers for malicious nodes detection in WSN. Sect. 3 describes material and methods with implementation of the proposed system, while Sect. 4 depicts the algorithm specification for proposed implementation. Sect. 5 describes the experimental setup for evaluating the proposed work and results achieved with our methodology and comparative analysis with various state-of-the-art methods. Section 6 discusses the conclusion and its future scope.

## 2 Review of Literature

According to [1], numerous segments and sub-size sensor nodes are often utilized as multi-hop communication points via the radio channel in wireless sensor networks. In today's world, wireless sensor networks are used for a variety of purposes, including health care, the environment, agriculture, flood control, and military applications [2]. Essential process generation, key distribution, crucial update, and post node setup are all part of the key management system. IoT devices are minimum objects that rely on themselves in a wireless network. By encrypting pairs of nodes in between, the access control method allows for four different types of keys to guarantee message synchronization. They are also crucial for information, community, and networking. The security key is used to encrypt previously transmitted data as well as to authenticate new nodes joining the network. A cluster's nodes are connected by a standard group key. Cryptographic keys in pairs are used to secure node-to-node communications and the whole network.

The WSN has a single or multiple placed hop route that are used to detect the condition and transmit the information to the access point. The technology that displays the AODV routing mechanism has already been modified to monitor and avoid wormholes in the real world. As a result, this improved AODV has been included in the wormhole threat detection and mitigation approach [3].

WSN nodes are vulnerable to a variety of assaults, including capture attempts, data movement jamming, and denial of service, due to their wireless networking, arrangement, and intimate contacts with their physical surroundings. As a result, all

network contacts must be secured. The distribution of cryptographic keys during the initial setup is a must for safeguarding communication between nodes [4].

For WSNs, a number of key management methods have been suggested. It is possible to divide them into two groups. The first is crucial pre-distribution methods, in which secret keys are selected from a pool and disseminated before to the nodes being deployed to the sensor nodes. Essential estimation is the second category. The primary calculation is based on a certain mathematical concept, such as matrices and polynomials [5].

Specific keys are kept in a node store in pairwise systems that have been suggested ( $n - 1$ ). This method offers sufficient security since the penetration of a single node does not reveal the secrets of other nodes. However, when the number of nodes grows, the device fails to scale because more memory is required to hold the keys [6].

Because they offer real-time monitoring and are possibly low-cost solutions, wireless sensors are becoming more popular as a solution to a variety of real-world issues. When sensors are used in harsh environments, they are susceptible to physical assault. Before a WSN may exchange data, encryption keys must be established between sensor nodes. Computational power and energy transfer will be saved using a hierarchical primary management system [7].

The wireless network outperforms traditional sensors because it is built with tiny sensor nodes that have limited processing and computing capabilities. These sensor nodes may detect, measure, and gather information from the environment and send it to the user or sink node depending on some local decision process or network application installed in the region [8].

When sensor networks are placed in an adversarial environment, they are susceptible to a variety of malicious assaults. An adversary listens to traffic rapidly and impersonates one of the network nodes. The communication should be encrypted and verified to ensure secrecy. A comprehensive examination of critical agreement problems in generic network settings has been conducted. Trusted-server, public-key, and key pre-distribution schemes are the three kinds of key arrangements accessible for agreements [9].

IoT relied heavily on wireless sensor networks (WSNs) to maintain reversal contact among nodes at the higher technology layer of the system. Key implementations can be categorized into decentralized protocols which are delegated. A centralized authority opts-in is solving problems and making to delete those keys or nodes inside the system. The decentralized model is smoother and requires fewer messages to submit but more complicated integration [10].

Intermediate network nodes can incorporate power-conservation findings from individual sensors. It gathers results from multiple sensors and calculates a minor post. This group activity involves safe communications from individuals. Team key establishment for safe group contact in WSN is the bottleneck [11].

Chan et al. [12] address three new key setting frameworks using a pre-distribution system to collect keys from each node randomly. Second, comparatively Its difficult to identify small scale attacks than large one. Second, in the multipath-reinforcement framework, the system demonstrates whether authentication between any pair of nodes can be improved by leveraging the security of other connections. The design

provides a random-pair credentials scheme that completely protects the confidentiality of the entire network when every node is detected and allows node-to-node authorization and qualified majority-based termination.

Cheng and Agrawal [13] suggested several key pre-distribution schemes create pair keys among sensor networks in the studies; they are too cumbersome or unreliable for some common attacks. To fix those vulnerabilities, they present the enhanced pairwise key setup scheme (IPKE) for wireless sensor networks of enormous size. Compared to the prior main proportion of total systems, the device scheme will provide complete network access, maximum assisted network size, lowest overhead transmission and energy consumption. The study results also show that the proposed framework scheme is resistant to various types of attacks.

Chen Zong-Gan et al. [14] state that the adaptive threshold NEDA addresses the LM-RAS problem. Three novel strategies, i.e., LPFE, NSS, HRS, are integrated into NEDA to achieve more effective CSS for extending the network's lifetime. LIFE measures individual fitness, which indicates the participation a person makes to the lifespan of the network. The NSS uses neighborhoods to create probabilistic models to identify new things, which helps to produce varied CSS. The Hours uses the residual power of the detectors as laboratory prepared for the improvement or elimination activities to fine-tune the exposure strategies.

Chen Ling et al. [15] studied WSN SCC algorithms primarily. The research indicates that attention should be paid to the sensor nodes and the channel's energy balance during the algorithm's operation. This also reflects a significant benefit of the SCC algorithm relative to the LEACH algorithm. The application of the SCC algorithm will effectively lengthen the network's lifetime and decrease network power loss through the simulation test line. The following two aspects indicate its core implementation advantages. Cluster head nodes are distributed uniformly in the network. They are also in a sparse location, effectively avoiding problems such as weak access to the network caused by inadequate energy supply.

Iqbal Siddiq et al. [16] state that the primary purpose is to estimate and improve definite issues such as node attack, node communication stability, energy consumption as well as throughput ratios. The principle of node communication by keys and creating a connection safely helps protect a network. The structure used to protect the sensor network in the paper is Single Space Key Distribution by Advanced Blom.

Farhadi Mostafa Moghadam et al. [17] defined the proposed protocol as a security attack based on security problems. Additionally, to resolve wireless nodes security concerns, the protocol scheme proposed by Alotaibi has implemented a shared authentication and essential agreement procedure based on Diffie–Hellman elliptic curve cryptography algorithm.

The energy-efficient hierarchical key controlling protocol inhomogeneous hierarchical WSN is defined by T. Kavitha and Rajadurai Kaliyaperumal [18]. This protocol is evaluated using network simulation, and its EEHKMP output consequences are matched with the E2ESCP in terms of energy, quality of service, and throughput. The ordinary delay transpired in EEHKMP evaluated with the E2ESCP approach

for efficient completion of contact between the nodes stands reduced to 0.23%. For EEHKMP and E2ESCP, the average packet transmission ratios are 0.77 and 0.69, respectively.

Aggarwal and Gupta [19] proposed an energy-efficient essential pre-distribution approach designed to create trust between nodes. This method considers the energy of the nodes when determining the keys and their relative distances from one another. In addition, the current critical pre-distribution techniques give very minimal or zero security from node compromise attacks. The scheme proposed could detect node compromise attacks without the need to share a key ring.

Ugur Yildiz Huseyin et al. [20] proposed a linear programming system to resolve the issue of incomplete stable communication in terms of its effect on network existence, path length, queue size, and energy dissipation. The results of numerical analysis of the device show that having a low (below 0.1) likelihood of primary protection association has a significant effect on network lifetime; however, as this likelihood increases, the decrease in network lifetime is rapidly curving down.

Jasim et Ahmed et al. [21] proposed a technique that generates a random value and a secret key unexpected time sign to improve authentication. The base station node will validate the false aggregated data when the packets are received using the previously generated key. Secure node authentication, data fragmentation techniques, fully homomorphic encryption, and model access control are also used to identify and prevent threats. The stable node authentication method prevents attackers from gaining access to the network. To minimize network latency, the base station node uses distance information between the participating nodes. To verify the dependency of the suggested procedure, the device replicates two well-known assaults: Sybil and sinkhole attacks.

### GAP Analysis

- Due to absence of data aggregation, most of system generate high network overhead during data transmission [5, 6].
- High power consumption for direct communication to cluster member to base station [11, 12].
- Low detection accuracy for malicious nodes detection [8–10].
- No prevention for internal attack detection like buffer overflow, DDOS, etc. [4].

### 3 Proposed System Design

This research basically emphasizes on secure data communication in WSN with pairwise semantic key encryption scheme and detection of internal malicious nodes

when it misbehaviors. System also explores intrusion detection system (IDS) technique with shortest path generation approach using BTC. The different phases of proposed work have been described below.

### **Problem statement**

This proposed work to design and develop a system for dynamic network attack detection and prevention from WSN; the major objective of proposed system it to detect malicious behaviors of nodes and prevent it.

### **Designing of WSN**

In the first phase, we create network simulation with a few nodes in the NS2 environment. The NAM simulator generates and sets each node location as well as public parameters. The particular node has specific energy in Joules, location details, radius, number of neighbor's nodes, etc.

### **Pairwise Symmetric-key generation and Encryption**

In that phase, we create pairwise key generation using a 64-bit AES Encryption Algorithm, and we collaborate some secure hash function methods, some RSA techniques, and numerous MD5-based one-way functions. We maintain one large [64 \* 64] matrix for pair matching for each receiver nodes. That maintains higher security for encrypted data.

### **Broadcast Tree Construction**

This approach is used for minimum resource utilization during data transmission. Whenever two parties communicate, it is mandatory to use section channel and the trustworthy shortest path. Using BTC, we construct the most straightforward path based on each neighbor node and its trust values. This approach gives us assurance about perfect path selection where never generates any cut during communication.

### **IDS approach**

Intruder nodes are dynamically monitoring their adjacent neighbors using operating messages that identify the black hole node. If a node is forwarding packets, a watchdog node can examine its next node's validity and forward the packet. So any node inside the coverage area is reviewed by the watchdog objective function. If any node does not deliver the incoming packet, the watchdog function counts delays and compares it to a predefined threshold for transmitting the message. Unless the wait is much less than the threshold, watchdog assumes maybe the next node is collaborative; else, the node would be marked as malicious. Even the watchdog node checks the packet so as not to change it. Whether any node only acknowledges but does not forward through the watchdog node, this watchdog recognizes it as a sinkhole node and eliminates the node from both the transmission of the packet path. The proportion of transmitting and receiving network packets to one sinkhole node is infinite.

## Intruder Revocation

Whenever any internal node exhibits misbehavior and is detected by the watchdog, it takes complete control and coordinates for the next transmission. Moreover, BTC immediately constructs another path to collaborating with remaining neighbor nodes and forwards data to the destination node. This process repeatedly executes until the destination is reached. The essential advantage of this method is, it never pauses the transmission process and attacker does not get complete information about hacked data.

## 4 Algorithm Design

This algorithm describes the execution of malicious node detection during the communication. It has validated each internal nodes-based IP table. Once data transmission has started from source node, it validates the entire process and collects the list of malicious nodes.

### 4.1 Algorithm for Malicious Node Detection

**Input:** All Network Nodes N, Source Nodes SRCnd, Destination node DESTnd, PREQ internal nodes

**Output:** detection of malicious node Mnode

**Step 1:** Initialized network with N nodes

**Step 2 :** select SRCnd and DESTnd for transmit the data

**Step 3 :** Find the route between SRCnd and DESTnd using path discovery and send PREQ packet to all nodes using below formula

$$Fx = \sum_{k=0}^n SRCnd \rightarrow PREQ \rightarrow DESTnd$$

**Step 4 :** Generate individual ip table at DESTnd for each source

**Step 5 :** Store all received packet data and information in respective ip table.

**Step 6 :** if(All ip tables are similar)

No malicious node found in network

**Step 7 : else**

$Ip[i] \leftarrow$  irrelevant ip table received by DESTnd

$Ip[j] \leftarrow$  any relevant ip table received by DESTnd

For each ( $Adj(N)$  into  $ip[i]$ )

$If(Adj(N)) != Ip[j])$

$MaliciousNode.add(Adj(N))$

*End if*

*End for*

**Step 8 :**  $MaliciousNode.List$

## 5 Results and Discussions

We provide the assessment of both the planned and current systems in this part. We quantitatively assess the study after explaining our experimental setup using the various parameters utilized, such as throughput, packet delivery ratio, cost, and time. Our tests are carried out using the NS2 simulator version 2.35, proven to provide realistic results. The NS simulator runs TCL code, while the header input uses both TCL and C++ code. For communication in our simulations, we utilize an infrastructure-based network architecture. The network selection is used to provide access to the wireless network at any time. In NS2, WMN simulates. The TCL file depicts the simulation of the whole architecture that was suggested. To run in the NS2 simulator, TCL uses the EvalVid Framework, which also helps to record running connection information messages using the connection pattern file us1. The NS2 trace file. Tr may be used to examine the findings. Filtering, processing, and presenting vector and scalar data are all supported. The project folder's results directory includes us. tr file, which provides the simulation's performance results. Using the graph tool, we plot the result parameters against the x and y-axis parameters based on us.tr file. The graph files have. awk extensions and may be plotted using the graph program (Table 1).

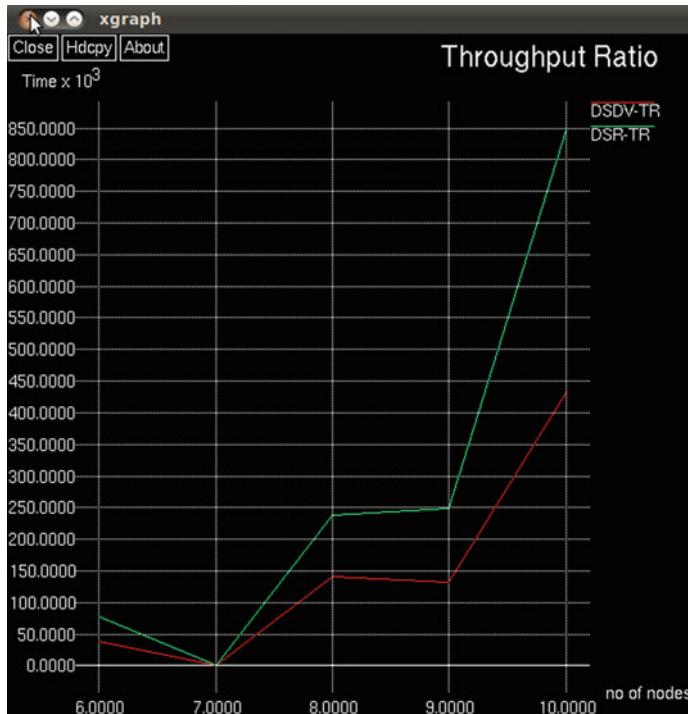
### 5.1 Performance Metrics and Parameters

#### 5.1.1 Throughput

The ratio of the total amount of transfer by the destinations to the entire simulation time is what it is called. The throughput is a result of the mobility pattern; for example, if two nodes are constantly nearby and move together, the TCP connection connecting them will have the same throughput as a single hop. Throughput, on the other hand, is 0 if the two nodes are constantly in separate network partitions. TCL

**Table 1** Parameters and values

Parameter	Values
Simulator	NS-allinone 2.35
Simulation time	25 s
Channel type	Wireless channel
Propagation model	2 Ray Ground
Standard	MAC/802.11
Simulation size	1000 * 1500
Max packet length	1000
Ad hoc routing	AODV, SAODV, DSR, DSDV
Traffic	CBR, PBR



**Fig. 2** Simulation results for number of nodes versus throughput

script is used to improve congestion prevention for TCP vegas. When the TCL script is run, it creates a trace file that contains all of the simulation events. The throughput is estimated with the help of an awk script that analyzes the trace file and saves the results to a file. Figure 2 shows the throughput displayed using the data received from running the awk script.

The diagram below shows throughput in an IDS system with four distinct protocols. Using DSR in simulation, the throughput numbers are somewhat improved.

### 5.1.2 End to End Delay

It is the time it takes for a packet to travel from source to destination across a network. The greater the performance, the shorter the delay. High channel delays in a mobile network that cause the TCP timer to expire would require TCP to retransmit the delayed packet needlessly, wasting time and energy, and degrading network performance.

Figure 3 illustrates the fluctuation of delay defined on the y-axis with regard to the number of nodes defined on the x-axis, and it can be shown that delay is not constant



**Fig. 3** Simulation results for number of nodes versus delay

with respect to nodes due to incomplete environmental awareness. In addition, it has been discovered that IDS with DSR has a shorter latency than DSDV.

### 5.1.3 Control Overhead

Overhead is defined as any combination of extra or indirect computing time, memory, bandwidth, or other resources needed to achieve a certain objective. Routing overhead increases the length of time it takes to transmit and receive routing packets, and the routes selected influence which nodes lose energy fastest. Proactive protocols may cause a significant increase in bandwidth and energy consumption for networks with highly dynamic topologies. Because the route to the destination is created when it is required based on an initial discovery between the source and the destination, reactive protocols trade off this overhead with increased latency. The simulation graph in Fig. 4 shows the change of control overhead (y-axis) with relation to the number of nodes (x-axis), and it can be shown that DSR has lower control overhead than DSDV. The number of nodes grows, so does the control overhead of IDS DSR.



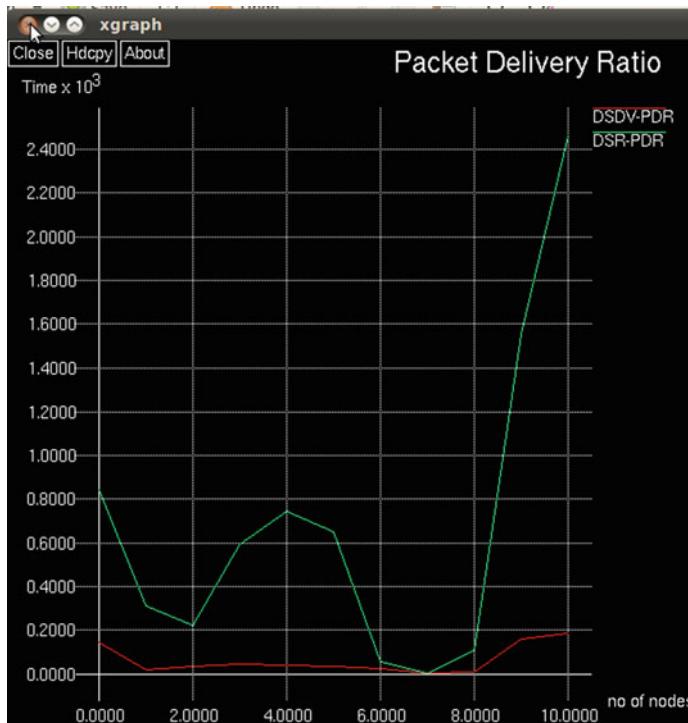
**Fig. 4** Simulation results for number of nodes versus control overhead

#### 5.1.4 Packet Delivery Ratio

The ratio of the total number of packets received by the destination node to the number of packets sent by the source node is what it is called. The simulation graph shows the change of packet delivery ratio (y-axis) with regard to the number of nodes (x-axis), and it can be shown that PDR of DSR is higher than DSDV IDS protocol (see Fig. 5).

#### 5.1.5 Packet Drop Ratio

The ratio of the total number of packets dropped by the destination node to the total number of packets sent by the source node is what it is called. The simulation graph depicts the change of packet loss ratio (y-axis) with relation to the number of nodes (x-axis), and it can be shown that DSR has a lower packet drop rate than DSDV IDS protocol (see Fig. 6).



**Fig. 5** Simulation results for number of nodes versus packet delivery ratio

In Fig. 7, we demonstrate the comparative analysis between proposed systems versus various state-of-art systems. The evaluation has done based on detection of nodes.

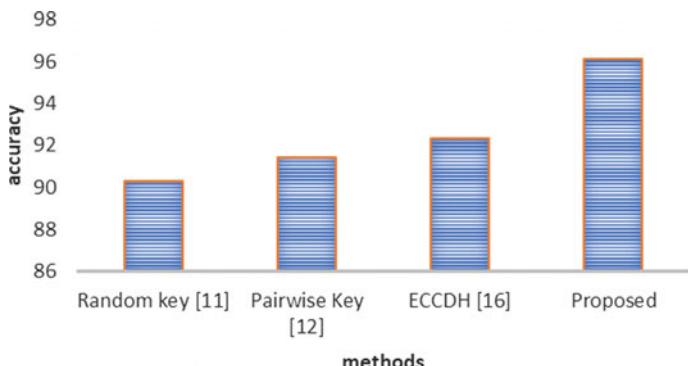
For this experiment, we used KDDCUP99 dataset for generate the malicious connection, and according to detection of those connection, accuracy has defined. The accuracy achieve by proposed system is around 96% which is better almost 5% than [12, 13, 17].

## 6 Conclusion and Future Work

This system focuses on the detection and prevention of network attacks during data transmission. The suggested methods improve ad hoc network security while also preventing various types of network assaults. It helps to increase the packet delivery ratio (PDR) and reduce network overhead by fostering the implementation of the appropriate routing protocol [22]. In the future, it will be necessary to enhance the table entries at the receiving node in order to identify wormhole nodes more quickly. In addition, wireless ad hoc networks' security will be improved. With the use of a



**Fig. 6** Simulation results for number of nodes versus packet drop ratio



**Fig. 7** Comparative analysis of proposed system

novel described method, it is possible to avoid different kin fog network assaults. Wormhole attacks have been recognized as attacks that may be strong and inflict significant network damage, even when communications need authentication and encryption, for future study.

## References

1. Zhang, X., Wang, J.: An efficient key management scheme in hierarchical wireless sensor networks. In: ICCCS, IEEE, 2015
2. Kaur, R., Kaur Sandhu, J.: A study on security attacks in wireless sensor network. In: 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2021, pp. 850–855. <https://doi.org/10.1109/ICACITE51222.2021.9404619>
3. Raghu Vamsi, P., Kant, K.: A taxonomy of key management schemes of wireless sensor networks. In: 5th IEEE, ICACCT, 2015, pp. 290–296
4. Nisha, Dave, M.: Storage as a parameter for classifying dynamic key management schemes proposed for WSNs. In: ICCTICT, IEEE, 2016
5. Rehman, S., Gang, C., Purevsuren, D.: A hybrid key management scheme for adhoc wireless sensor networks to improve secure link establishment. In: 6thICSCSS, IEEE, 2015
6. Chakavarika, T.T., Chaurasia, B.K., Gupta, S.K.: Improved energy efficient key management scheme in wireless sensor networks. In: ICCSP, pp. 2135–2140. IEEE, 2016
7. Tian, B., Xin, Y., Lu, S., Yang, X., Li, D., Gong, Z., Yang, Y.: A novel key management method for wireless sensor networks. In: 3rd International Conference on Broadband Network and Multimedia Technology IEEE, pp. 1106–1110, 2010
8. Barad, J., Kadhiwala, B.: Improvement of deterministic key management scheme for securing cluster-based sensor networks. In: 1st IEEE International Conference on Networks and Soft Computing (ICNSC), 2014, pp. 55–59
9. Du, W., Deng, J., Han, Y.S., Varshney, P.K., Katz, J., Khalili, A.: A pairwise key pre-distribution scheme for wireless sensor networks. ACM Trans. Inf. Syst. Secur. **8**(2), 228–258 (2005)
10. Mansur, I., Chalhoub, G., Lafourcade, P.: Key management in wireless sensor networks. J. Sens. Actuator Netw., pp. 251–273 (2015)
11. Wang, E.K., Hui, L.C.K., Yiu, S.M.: A new key establishment scheme for wireless sensor networks. IJNSA **1**(2), 17–27 (2009)
12. Chan, H., Perrig, A., Song, D.: Random key predistribution schemes for sensor networks. In: 2003 Symposium on Security and Privacy, 2003 May 11, pp. 197–213. IEEE
13. Cheng, Y., Agrawal, D.P.: Improved pairwise key establishment for wireless sensor networks. In: 2006 IEEE International Conference on Wireless and Mobile Computing, Networking and Communications 2006 Jun 19, pp. 442–448. IEEE
14. Chen, Z.G., Lin, Y., Gong, Y.J., Zhan, Z.H., Zhang, J.: Maximizing lifetime of range-adjustable wireless sensor networks: a neighborhood-based estimation of distribution algorithm. IEEE Trans. Cybern. 2020 Apr 1
15. Chen, L., Liu, W., Gong, D., Chen, Y.: Cluster-based routing algorithm for WSN based on subtractive clustering. In: 2020 International Wireless Communications and Mobile Computing (IWCMC) 2020 Jun 15, pp. 403–406. IEEE
16. Iqbal, S., Prerana, S., Sukrutha, H., PurushottamShanbhag, G.: Attack resistant secure key management in wireless sensor networks. In: 2019 1st International Conference on Advances in Information Technology (ICAIT) 2019 Jul 25, pp. 475–479. IEEE
17. Moghadam, M.F., Nikooghadam, M., Al Jabban, M.A., Alishahi, M., Mortazavi, L., Mohajerzadeh, A.: An efficient authentication and key agreement scheme based on ECDH for wireless sensor network. IEEE Access. **13**(8), 73182–73192 (2020)
18. Kavitha, T., Kaliyaperumal, R.: Energy efficient hierarchical key management protocol. In: 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS) 2019 Mar 15, pp. 53–60. IEEE
19. Aggarwal, C., Gupta, B.B.: Energy efficient key pre distribution scheme in WSN. In: 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) 2018 May 18, pp. 2480–2484. IEEE
20. Ciftler, B.S., Yildiz, H.U., Tavl, B., Bıçakçı, K., Incebak, D.: The impact of incomplete secure connectivity on the lifetime of wireless sensor networks (2018)

21. Jasim, A.A., Idris, M.Y., Azzuhri, S.R., Issa, N.R., Noor, N.B., Kakarla, J., Amiri, I.S.: Secure and energy-efficient data aggregation method based on an access control model. *IEEE Access* **11**(7), 164327–164343 (2019)
22. Jacob, I.J., Ebby Darney, P.: Artificial Bee Colony optimization algorithm for enhancing routing in wireless networks. *J. Artif. Intell.* **3**(01), 62–71 (2021)

# Performance Analysis and Assessment of Various Energy Efficient Clustering-Based Protocols in WSN



Trupti Shripad Tagare and Rajashree Narendra

**Abstract** Currently, the world is witnessing a rapid development in the area of wireless sensor networks. It holds a potential to transform many features of our economy and life, starting from bio diversity mapping to industry automation, transportation, healthcare monitoring. These networks primarily aim to develop protocols to utilize the sensor node energy efficiently and thereby maximize the lifetime of the network. In this research paper, we compare the energy efficient clustering-based protocols used in wireless sensor networks. The work involves implementation of homogeneous protocols, namely low-energy adaptive clustering hierarchy (LEACH), energy aware multi-hop multi-path hierarchical (EAMMH). Next, the heterogeneous protocols such as enhanced distributed energy efficient clustering (EDEEC) and stable election protocol (SEP) are executed. The analysis is carried out in terms of number of nodes and probability of election for cluster head (CH). The observations and results obtained for these protocols are presented.

**Keywords** Wireless sensor networks · Low-energy adaptive clustering hierarchy · Energy aware multi-hop multi-path hierarchical · Enhanced distributed energy efficient clustering · Stable election protocol · Average energy · Dead nodes · Energy efficiency · Cluster head

## 1 Introduction

Wireless sensor networks (WSNs) consist of large number of sensor nodes. These nodes are spatially distributed in the area to be monitored. These nodes are tiny and have limited power. These sensor nodes are usually placed in inhospitable environments to sense various physical parameters like temperature, pressure, vibrations,

---

T. S. Tagare (✉)  
Dayananda Sagar College of Engineering, Bengaluru, India  
e-mail: [truptitagare-ece@dayanandasagar.edu](mailto:truptitagare-ece@dayanandasagar.edu)

R. Narendra  
Dayananda Sagar University, Bengaluru, India  
e-mail: [rajashree-ece@dsu.edu.in](mailto:rajashree-ece@dsu.edu.in)

etc. The node consists of sensing device, trans-receiver, processing unit, and a tiny battery. The basic requirement of these networks is to consume minimum power and enhance the network lifetime [1, 2]. Energy is the most important resource for a WSN. This demands for efficient use of node energy as the exchange or recharging of node batteries in these conditions is a herculean task. Today, WSNs are employed in large number of applications like military applications, healthcare monitoring, telematics, smart buildings, etc. Hence, a number of routing protocols are proposed to reduce the consumption of energy in the nodes.

The energy efficient protocols in WSN are classified into the following categories:

- Flat routing-based protocols
- Hierarchical-based protocols
- Location-based protocols.

**Flat routing-based protocols:** Each and every node in the WSN has an important role. The sink requests for some data and then waits for each sensor node to sense and provide the data centric characteristics [3]. Few techniques used in the implementation are direct transmission and minimum transmission energy (MTE)-based protocols.

**Hierarchical-based protocols** are implemented in two steps. First, the cluster head is chosen, and then, the routing is carried out. The data are aggregated to make the network scalable and energy efficient [3]. These are further classified into homogeneous protocols which can be implemented using LEACH, LEACH—centralized (LEACH-C), EAMMH protocols, and heterogeneous protocols which can be implemented using DEEC, EDEEC, SEP techniques. **Chain-based protocol** is also an approach from hierarchical routing protocols which reduce the consumption of energy in the network. It can be implemented using power efficient gathering in sensor information systems (PEGASIS), hierarchical PEGASIS, mobile sink improved energy efficient PEGASIS-based (MIEEP) routing protocol.

**Location-based protocols** track the node location. The two techniques used to estimate the nearest neighboring node distance, i.e., finding the coordinate of the neighboring node, and the other is the use of global positioning system (GPS) [3].

In this work, we analyze the hierarchical-based clustering routing protocols of WSN and evaluate their performance. Our proposed concept is to compare the homogeneous and heterogeneous protocols using simulations with MATLAB and has presented details comparative evaluation and inferences.

## 2 Literature Survey

- Paper [1] elaborated on the need of energy efficiency in WSN and implemented direct communication, LEACH, energy, and distance-based algorithms for few nodes and concluded that LEACH outperformed the other techniques.
- In [4], the detailed survey of flat routing protocols was carried out, and several parameters were taken into account to establish a comparison between them.

- Paper [5] provided a detailed review of all the efficient energy routing protocols and techniques to increase the network lifetime
- In [3], the author gave a detailed insight of the hierarchical-based protocols and carried out a comparative study between MODLEACH and MIEEPB routing protocols in WSN with results showing better performance of MIEEPB over MODLEACH.
- While paper [6] carried out implementation of clustering techniques, namely EAMMH, LEACH, SEP, and TEEN routing protocols for WSN. The results proved that the change in these parameters affects the performance of cluster protocols in WSN.
- [7] Demonstrated the implementation of LEACH and SEP protocols. It employed MATLAB to carry out the analysis.
- In [8], the comparative study of LEACH, N-LEACH, and SEP routing protocols of WSNs was carried out. The results showed that SEP protocol provided a longer network lifetime due to its heterogenous nature. Also, the SEP's advance nodes helped in increasing the network lifetime.
- In [9], the author discusses the classification and comparison of different energy routing protocols. Here, NS simulator was used for simulation and carried out the performance analysis.
- Paper [10] proposed a hierarchical clustering protocol with a load balanced approach. The parameters considered were energy efficiency, scalability, and reliability. An ant lion optimizer was used in the selection of CH. The results showed an improvement in all the parameters.
- [11] proposed the dragonfly algorithm which used the LEACH-C protocol. Here, the flocking attitude of the flies was utilized in CH selection. The simulated results showed an improvement in the number of alive nodes, increase in life time of the network.

### 3 Objective

The efficient energy consumption in WSN is of primary importance to enhance the network lifetime. It is clear from the literature survey that many researchers have proposed different clustering protocols, namely LEACH, LEACH-C, DEEC, EDEEC SEP, EAMMH, and EEHC. However, this study, we compare and analyze the performance of only the following variants, namely

- Homogeneous Protocols: LEACH, EAMMH
- Heterogeneous Protocols: SEP, EEDEC.

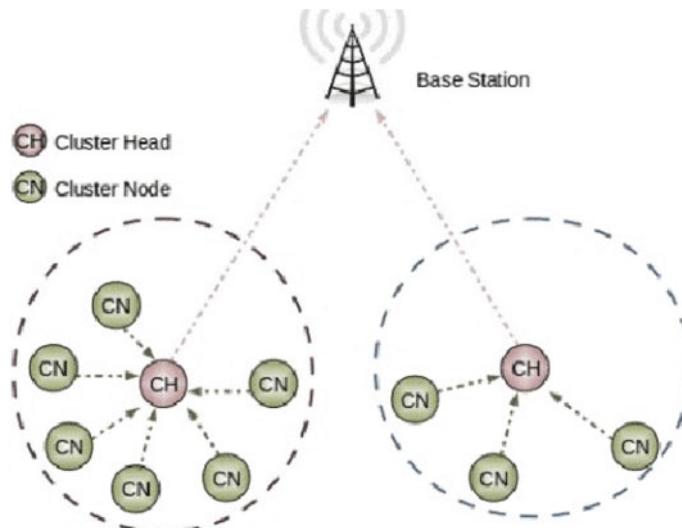
## 4 A Detailed Study

### 4.1 Homogeneous Protocols

Homogeneous protocols are the type of protocols that operate on a homogeneous WSN. A homogeneous WSN is one where all the sensor nodes have the same capabilities in power, storage, processing, and communication [12].

The homogeneous protocols are discussed below:

**Low-energy Adaptive Clustering Hierarchy (LEACH):** It is established on hierarchical routing. In this approach, all the nodes in a network divide themselves into several local clusters and a cluster head (CH) which is selected on the basis of probability from randomly placed sensor nodes. Figure 1 depicts the structure of the CH and cluster nodes. Here, the data aggregation is performed in the CH, and the data are transmitted to base station (BS) using single hop. Basically, a stochastic algorithm is applied to each node at each round. When not sensing, these nodes go into sleep mode and conserve their energy. Each round helps the node in determining, whether it will become a CH. A node elected as CH cannot become CH for next P rounds , where P is the desired percentage of CHs which will be set by the designer.



**Fig. 1** The architecture of cluster head and cluster nodes [6]

Thus, in each round, a node has a probability of  $1/P$  in being elected as CH. The normal nodes select their nearest CH to transmit its sensed data using time division multiple access (TDMA) approach [13, 14].

**Energy Aware Multi-hop Multi-path Hierarchical (EAMMH) Protocol:** It combines the energy aware routing and multi-hop intra-clustering approaches. It divides the network into clusters, but it includes multi-hop to transmit the aggregated data to the base station. Similar to LEACH, EAMMH also has the set-up phase wherein the cluster organization takes place. Next, in the steady-state phase, the aggregated data are sent to the BS.

## 4.2 Heterogeneous Protocols

Heterogeneous protocols are the type of protocols that operate on a heterogeneous WSN. A heterogeneous WSN is one where all the sensor nodes have different capabilities in terms of computing power or range of communication. The concept is based on irregular sensor model which is used to calculate fairly accurately the performance of sensor nodes.

The heterogeneous protocols are discussed below:

**Stable Election Protocol (SEP):** This protocol aims at understanding the heterogeneity present in the nodes using hierarchical protocols. Here, the nodes that are elected as CH, aggregate the data, and transmit it to BS. Here, we consider a heterogeneity in nodes by classifying them into two types. First are the normal nodes enriched with few Joules of energy, and second type are advanced nodes which are facilitated with more energy when compared with normal nodes (in Joules). SEP applies at each node the weighted election probabilities to elect a CH node. The protocol delays the death of first node, and that period is referred to as the stability period. This approach is powerful for closed loop applications with feedback are in consideration.

**Enhanced Distributed Energy Efficient Clustering (EDEEC):** This protocol works in line with SEP and EDEEC algorithm. However, the only enhancement it makes is to consider the energy present in the entire network as also a criterion for the selection of CH at each round. It considers heterogeneity in the nodes by considering the normal advanced nodes and super nodes. Here, the probabilities are set for normal nodes, advanced nodes, super nodes remaining energy level and total energy of the network to select a CH in every round of election.

## 5 Simulations and Analytical Observations

In this section, we carry out the analysis and performance evaluation of all the hierarchical clustering protocols discussed using MATLAB R2020a. The complete codes are written, and simulation results obtained for different cases are provided in this section. We provide the comparison between the homogeneous and heterogeneous protocols. The homogeneous protocols have all nodes with same energy (in Joules) while the heterogeneous protocols have variety of nodes namely the normal, advanced, and super nodes which differ in their initial energy. The protocols for homogeneous, namely LEACH and EAMMH and for heterogeneous, namely SEP and EDEEC, are compared here.

### Homogeneous Protocols

#### Simulation Parameters

The simulation parameters considered are shown in Table 1.

#### Simulation Details

First, the homogeneous protocols LEACH and EAMMH are considered. All the simulation parameters are kept same during the simulation of LEACH and EAMMH protocols.

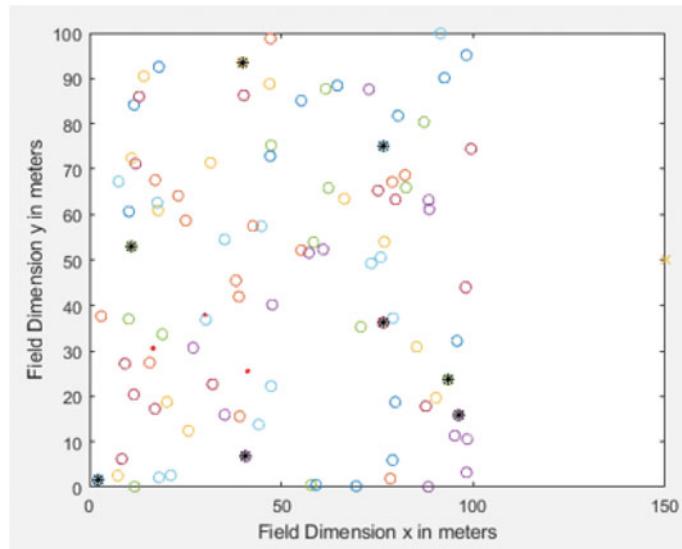
#### Assumptions

- Nodes are static in nature.
- The energy present in each node is same and hence homogeneous.

Figure 2 represents the random distribution of heterogeneous cluster nodes in a  $100 \text{ m} \times 100 \text{ m}$  network field depicting the normal nodes (alive), cluster heads, dead nodes.

**Table 1** Simulation parameters for homogeneous protocols: LEACH and EAMMH

Simulation parameters	Values
Network area	$100 \text{ m} \times 100 \text{ m}$
Number of nodes (varied)	CASE 1&3:100 CASE 2&4:250
Initial energy: $E_0$	0.1 J
$E_{elec} = E_{tx} = E_{rx}$	50 nJ
Probability of a node to become cluster head during election (varied)	CASE 1:0.1 CASE 2:0.2
Energy spent in transmit amplifier types: $E_{fs}$ $E_{mp}$	10 pJ 0.0013 pJ
Energy required for data aggregation: $E_{da}$	5 nJ
Maximum number of rounds: $r_{max}$	100
Heterogeneity percentage	0.0



**Fig. 2** The random distribution of homogeneous cluster nodes in a  $100 \text{ m} \times 100 \text{ m}$  network field

The simulations are executed with respect to the following parameters:

- Average energy in each node per round (with variation of probability and number of nodes)

The simulation results are shown in Table 2 for four cases.

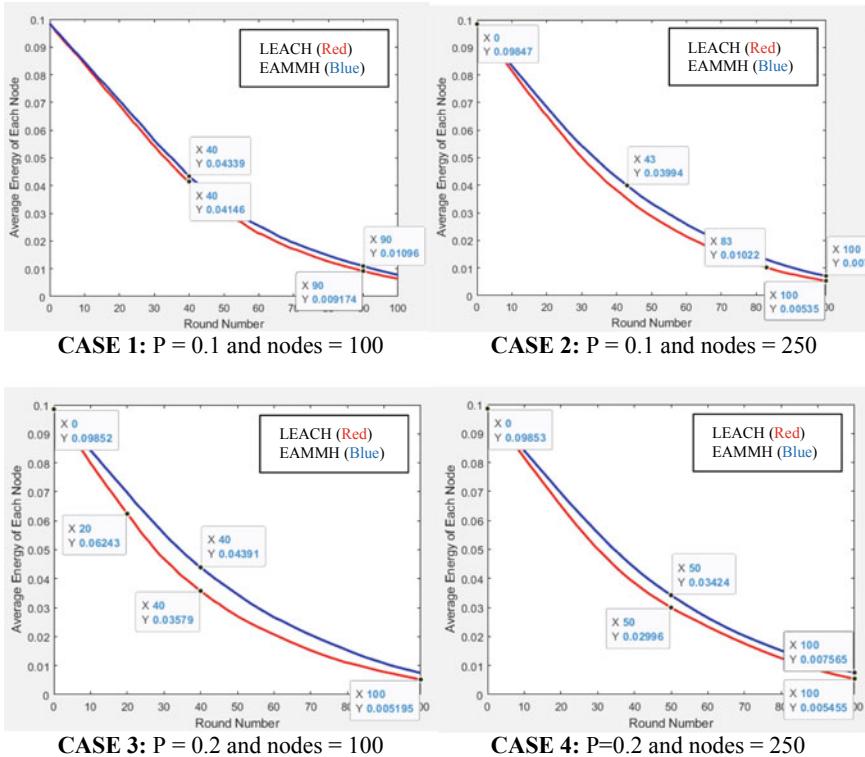
### Observation

As seen in Fig. 3 and from Table 2,

**CASE 1:** With  $P = 0.1$  and number of nodes = 100

**Table 2** Average energy of each sensor node per round

Round No rmax	CASE 1: $P = 0.1$ No. of nodes = 100		CASE 2: $P = 0.1$ No. of nodes = 250		CASE 3: $P = 0.2$ No. of nodes = 100		CASE 4: $P = 0.2$ No. of nodes = 250	
	Average energy in sensor node (in J)		Average energy in sensor node (in J)		Average energy in sensor node (in J)		Average energy in sensor node (in J)	
	LEACH	EAMMH	LEACH	EAMMH	LEACH	EAMMH	LEACH	EAMMH
0	0.09482	0.09482	0.09847	0.09847	0.09852	0.09852	0.09853	0.09853
20	0.06903	0.07069	0.06527	0.06835	0.06943	0.06973	0.06507	0.06945
40	0.04146	0.04339	0.03827	0.04293	0.03579	0.04391	0.03851	0.04378
60	0.02276	0.02549	0.02148	0.02587	0.02078	0.02669	0.02339	0.02645
80	0.01262	0.01477	0.01147	0.01439	0.01097	0.01531	0.01254	0.01517
100	0.00642	0.00776	0.00535	0.00716	0.00519	0.00743	0.00545	0.00756



**Fig. 3** LEACH and EAMMH protocol: average energy in each node per round (in Joules) of each node for four cases

Results show that in EAMMH protocol, there is more average energy in each node compared to LEACH protocol with each round. At 100th round, EAMMH shows a 0.1343% increase in average energy in the sensor node when compared with that using LEACH.

#### **CASE 2:** With $P = 0.1$ and number of nodes = 250

EAMMH has more average energy in each sensor node per round when number of nodes is increased. At 100th round, EAMMH shows a 0.182% increase in average energy in the sensor node when compared with that using LEACH.

#### **CASE 3:** With $P = 0.2$ and number of nodes = 100

Results show that EAMMH still performs better than LEACH. At 100th round, EAMMH shows a 0.224% increase in average energy in the sensor node when compared with that using LEACH.

#### **CASE 4:** With $P = 0.2$ and number of nodes = 250

Results again show that EAMMH still performs better than LEACH. At 100th round, EAMMH shows a 0.211% increase in average energy in the sensor node when compared with that using LEACH.

**Table 3** Number of dead nodes per round

Round No rmax	CASE 1: P = 0.1 No. of nodes = 100		CASE 2: P = 0.1 No. of nodes = 250		CASE 3: P = 0.2 No. of nodes = 100		CASE 4: P = 0.2 No. of nodes = 250	
	No. of dead nodes		No. of dead nodes		No. of dead nodes		No. of dead nodes	
	LEACH	EAMMH	LEACH	EAMMH	LEACH	EAMMH	LEACH	EAMMH
30	3	1	18	11	9	3	16	10
40	7	5	47	36	23	11	48	27
70	45	42	122	101	47	38	109	99
80	54	51	141	117	57	45	124	119
90	64	56	160	136	67	53	146	134
100	69	63	186	166	74	61	174	161

- Number of dead nodes per round (with variation of probability and number of nodes)

### Tabulation

Table 3 shows comparison of LEACH and EAMMH protocol with respect to number of dead nodes per round with respect to four cases were the probability for CH election and number of nodes is varied.

### Observations and Result

As seen in Fig. 4 and from Table 3,

CASE 1: With P = 0.1 and number of nodes = 100.

In LEACH protocol, the first dead node is seen at round number 25, while for EAMMH, it occurs at round number 28.

CASE 2: With P = 0.1 and number of nodes = 250.

The simulation results for LEACH and EAMMH protocol show that in LEACH protocol, the first dead node is seen at round 15, while for EAMMH, it occurs at round 16.

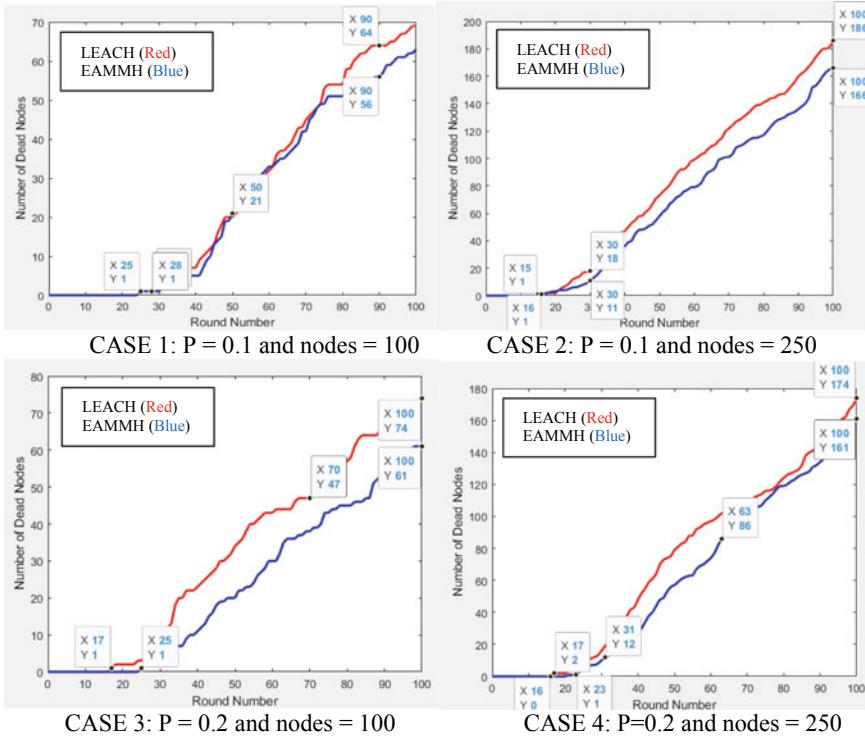
CASE 3: With P = 0.2 and number of nodes = 100.

In LEACH protocol, the first dead node is seen at round number 17, while for EAMMH, it occurs at round number 25.

CASE 4: With P = 0.2 and number of nodes = 250.

In LEACH protocol, the first dead node is seen at round number 17, while for EAMMH, it occurs at round number 23. Compared to CASE 1, with more probability and number of nodes, the time of death of first node in the network is delayed by double number of rounds.

Conclusion for homogenous protocols between LEACH and EAMMH: From the tabulations for all the cases, it is clear that EAMMH provides longer network lifetime, and average energy per node is greater compared to LEACH protocol.



**Fig. 4** LEACH and EAMMH protocol: number of dead nodes per round with  $P = 0.1$  and nodes = 100

## Heterogeneous Protocols

### Simulation Parameters

The simulation parameters are shown in Table 4.

### Simulation Details

The heterogeneous protocols SEP and EDEEC are simulated using MATLAB [15].

### Assumptions

- The nodes are static in nature.
- The energy in normal nodes is less than that in advanced nodes.

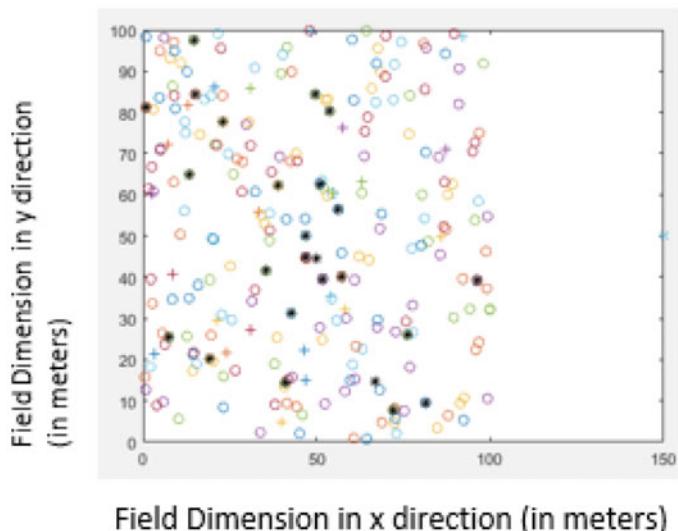
Figure 5 represents the random distribution of heterogeneous cluster nodes in a  $100 \text{ m} \times 100 \text{ m}$  network field depicting the normal nodes (alive), advanced nodes (alive), cluster heads, dead nodes.

The simulations are executed with respect to the following parameters:

- Average energy in each node per round (in Joules)

**Table 4** Simulation parameters for heterogeneous protocol: SEP

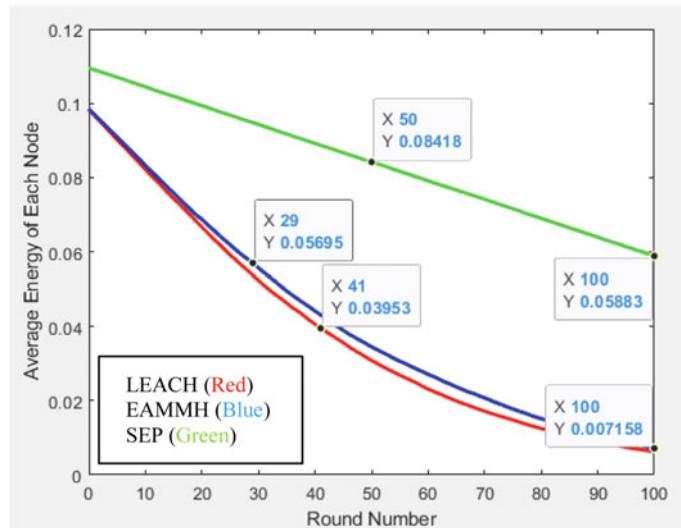
Simulation parameters	Values
Network area	100 m × 100 m
Number of nodes (varied)	250
Initial energy: E0	0.1 J
Eelec = Etx = Erx	50 nJ
Probability of a node to become cluster head during election (varied)	0.1
Energy spent in transmit amplifier Types: Efs Emp	10 pJ 0.0013 pJ
Energy required for data aggregation: EDA	5 nJ
Maximum number of rounds: rmax	100
Heterogeneity percentage	0.1

**Fig. 5** The random distribution of heterogeneous cluster nodes in a 100 m × 100 m network field

Comparison between homogeneous protocols LEACH, EAMMH, and heterogeneous SEP is carried out for 250 nodes as shown in Fig. 6.

### Observations and Result

As seen in Fig. 6 and Table 5, the simulation is carried out for a network field with 250 nodes. Also, probability  $P = 0.1$  is fixed. Thus, for same settings, the simulation results show that heterogeneous SEP protocol is more energy efficient protocol when compared with homogeneous LEACH and EAMMH techniques. However, EAMMH is better than LEACH as observed in our previous results. But these results hold good as long as the number of nodes is less, but when number of nodes increase



**Fig. 6** Comparison between LEACH, EAMMH, and SEP protocols for average energy (in Joules) in each node per round

(to 250 as shown in results), the average energy curve of both LEACH and EAMMH shown same results.

Thus, LEACH protocol is suitable when the number of nodes is less because of its simplicity.

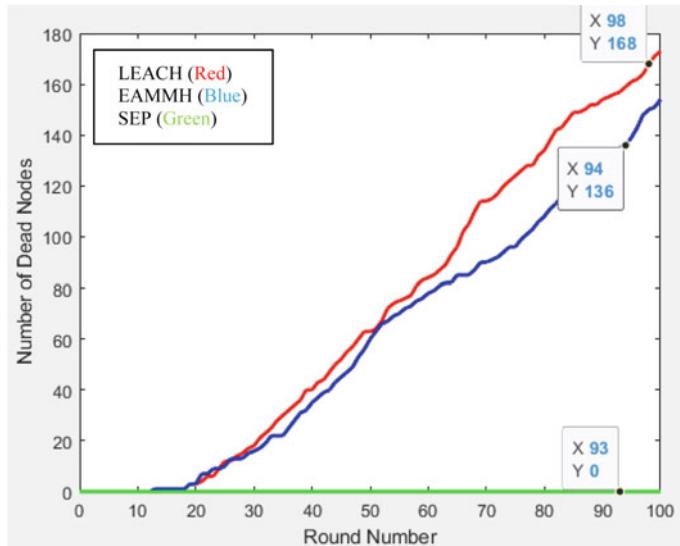
SEP protocol is best suited when the nodes are large in number.

- Number of dead nodes per round

Comparison between homogeneous protocols LEACH, EAMMH, and heterogeneous SEP is carried out for 250 nodes as shown in Fig. 7.

**Table 5** Average energy of each node per round comparison for LEACH, EAMMH, SEP

Round No rmax	CASE: P = 0.1, No. of nodes = 250		
	Average energy in sensor node (in J)		
	LEACH	EAMMH	SEP
0	0.09849	0.09849	0.1095
20	0.06663	0.06871	0.09937
40	0.04065	0.004417	0.08917
60	0.02315	0.02713	0.07908
80	0.01251	0.01515	0.06898
100	0.00715	0.00715	0.05883



**Fig. 7** Comparison between LEACH, EAMMH, and SEP protocols for no. of dead nodes per round

### Observation

As seen in Fig. 7 and Table 6, the simulation is carried out for a network field with 250 nodes. Also, probability  $P = 0.1$  is fixed. Thus, for same settings, the simulation results show that heterogeneous SEP protocol provides longer network lifetime when compared with homogeneous LEACH and EAMMH techniques. Only when, the number of nodes is still increased to 350, 500, we can observe dead nodes after 1000 rounds. Thus, again helps us to infer that with heterogeneity in the network, the protocols help to increase the network lifetime.

- EDEEC protocol is implemented with the following improved parameters compared to LEACH, EAMMH, and SEP are presented in Table 7

**Table 6** No. of dead nodes per round comparison for LEACH, EAMMH, SEP

Round No rmax	CASE: $P = 0.1$ , No. of nodes = 250		
	Average energy in sensor node (in J)		
	LEACH	EAMMH	SEP
30	18	16	0
40	40	35	0
70	114	90	0
80	134	108	0
90	154	131	0
100	173	154	0

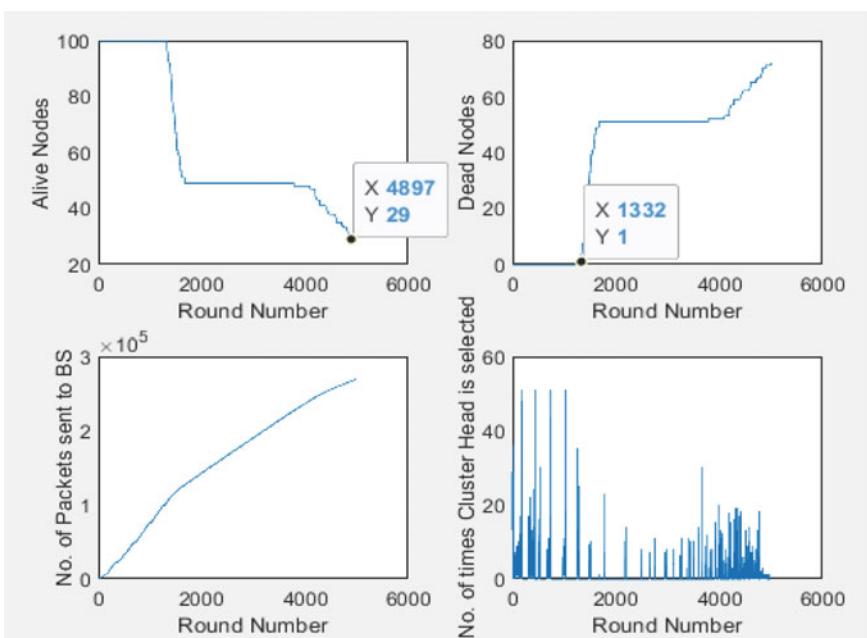
**Table 7** Simulation parameters for heterogeneous protocol: EDEEC

Simulation parameters	Values
Network area	100 m × 100 m
Number of nodes(varied)	500
Initial energy E0	0.5 J
Eelec = Etx = Erx	50 nJ
Probability of a node to become cluster head during election (varied)	0.1
Energy spent in transmit amplifier Types: Efs Emp	10 pJ 0.0013 pJ
Energy required for data aggregation: EDA	5 nJ
Maximum number of rounds: rmax	5000
Heterogeneity percentage	0.1

(i) Alive nodes per round; (ii) dead nodes per round; (iii) no. of packets transmitted to base station per round; (iv) no. of times the cluster head is selected per round.

### Observation

As seen in Fig. 8, EDEEC protocol provides longer lifetime for the network with the use of advanced and super advanced nodes.

**Fig. 8** EDEEC protocol simulation outputs

- Alive nodes are more even, when the protocol is run for 5000 rounds
- Results show that, at 1332 rounds, only 1 node is dead in the network.
- At 5000 rounds, the number of packets transmitted to BS is 0.26 Mb.

Conclusion for heterogenous protocols between SEP and EDEEC: From the tabulations for all the cases, it is clear that SEP is better than homogenous protocols LEACH and EAMMH because of heterogeneity of nodal energy. Heterogeneity provides longer network lifetime, and average energy per node is greater. Further EEDEC outperforms LEACH, EAMMH, and SEP when large dense network consisting of large number of nodes is involved.

## 6 Final Conclusion

Energy efficiency in sensor nodes for prolonged network lifetime is a basic requirement in WSN applications [16]. This can be effectively obtained by the use of energy efficient hierarchical clustering techniques. In this work, the comparative analysis between homogeneous and heterogeneous protocols is carried out. The MATLAB 2020a simulation tool was used. The previous work in this area compared the algorithms with fixed number of nodes and probability [17]. It did not consider EDEEC algorithm. So, in this work, an attempt is made to see the effects of change in these parameters to the energy efficiency and life time of the network and worked on LEACH, EAMMH, EDEEC, and SEP protocols.

- Firstly, the homogeneous protocols with nodes in the network with same capabilities for various cases were carried out. The results prove that EAMMH is more energy efficient as it consists of intra-clustering routing technique which is responsible for delaying the death of first node and hence contributes significantly in enhancing the lifetime of the network.
- However, LEACH continues to be the best choice when very few nodes are into consideration as its very simple to implement.
- As the number of nodes increase, EAMMH and LEACH provide similar results.
- Next, the heterogeneity in the nodes was introduced to enhance the lifetime of the network by introducing few nodes with more energy. The advanced and super nodes were initialized with more energy than the normal nodes.
- Results of SEP prove that where EAMMH ceases to be efficient when nodes in the network increase, SEP provides increased network lifetime as it prolongs the stability period when compared with other clustering techniques.
- The results of EDEEC protocol conclude that it is the most energy efficient protocol and suitable for dense and sizeable WSNs.

In future, the chain-based protocols can be implemented and performance evaluation with respect to different parameters can be carried out.

## References

1. Tagare, T.S., Narendra, R., Manjunath, T.C.: A GUI to analyze the energy consumption in case of static and dynamic nodes in WSN. In: Shakya, S., Bestak, R., Palanisamy, R., Kamel, K.A. (eds.) Mobile Computing and Sustainable Informatics. Lecture Notes on Data Engineering and Communications Technologies, vol. 68. Springer, Singapore (2021). [https://doi.org/10.1007/978-981-16-1866-6\\_40](https://doi.org/10.1007/978-981-16-1866-6_40)
2. Mugunthan, S.R.: Wireless rechargeable sensor network fault modeling and stability analysis. *J. Soft Comput. Paradigm (JSCP)* **3**(01), 47–54 (2021)
3. Zayed, H., Taha, M., Allam, A.H.: Performance evaluation of MODLEACH and MIEEPB routing protocols In WSN. In: 2018 International Conference on Electrical, Electronics, Computers, Communication, Mechanical and Computing (EECCMC); with catalog No. efp18037-PRJ:978-1-5386-430-7, during 28–29 January 2018 at Priyadarshini Engineering College, Vaniyambadi, India (2018)
4. Echoukairi, H., Bourgba, K., Ouzzif, M.: A Survey on Flat Routing Protocols in Wireless Sensor Networks Laboratory of Computer Networks, Telecommunications and Multi-media, Higher School of Technology, Hassan II University, Casablanca, Morocco, Springer Science+Business Media Singapore E. Sabir et al. (eds.), The International Symposium on Ubiquitous Networking, Lecture Notes in Electrical Engineering 366 (2016). [https://doi.org/10.1007/978-981-287-990-5\\_25](https://doi.org/10.1007/978-981-287-990-5_25)
5. Hiremani, N., Basavaraju, T.G.: Energy efficient routing protocols, classification and comparison in wireless sensor networks. In: International Conference on Current Trends in Computer, Electrical, Electronics and Communication, 978-1-5386-3243-7/17/IEEE (2017)
6. El-Sayed, H.H.: Performance evaluation of clustering EAMMH, LEACH SEP, TEEN protocols in WSN. *Inf. Sci. Lett.* **7**(2), 35–40 (2018). [www.naturalspublishing.com/Journals.asp](http://www.naturalspublishing.com/Journals.asp)
7. Ayoob, M., Zhen, Q., Adnan, S., Gull, B.: Research of Improvement on LEACH and SEP Routing Protocols in Wireless Sensor Networks (2016). <https://doi.org/10.1109/ICCRE.2016.7476141>
8. Liang, H., Yang, S., Li, L., Gao, J.: Research on routing optimization of WSNs based on improved LEACH protocol Liang et al. *EURASIP J. Wirel. Commun. Netw.* **2019**, 194 (2019). <https://doi.org/10.1186/s13638-019-1509-y>
9. Homaei, M.H.: Low Energy Adaptive Clustering Hierarchy protocol (LEACH) (2021). <https://www.mathworks.com/matlabcentral/fileexchange/44073-low-energy-adaptive-clustering-hierarchy-protocol-leach>. MATLAB Central File Exchange. Retrieved Sept 28 2021
10. Singh, H., Singh, D: Hierarchical clustering and routing protocol to ensure scalability and reliability in large-scale wireless sensor networks. *J. Supercomput.* **77**, 10165–10183 (2021). <https://doi.org/10.1007/s11227-021-03671-1>
11. Pitchaimickam, B.: Dragonfly algorithm for hierarchical clustering in wireless sensor networks. In: 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 192–197
12. Javaid, N., Qureshi, T.N., Khan, A.H., Iqbal, A., Akhtar, E., Ishfaq, M.: EDDEEC: enhanced developed distributed energy-efficient clustering for heterogeneous wireless sensor networks. Elsevier, *Procedia Comput. Sci.* **19**, 914–919 (2013). 1877–0509 © 2013 The Authors. Published by Elsevier B.V. Selection and peer-review under responsibility of Elhadji M. Shakshuki. <https://doi.org/10.1016/j.procs.2013.06.125> International Workshop on Body Area Sensor Networks (BASNet-2013). <https://doi.org/10.1016/j.procs.2013.06.125>
13. Tagare, T.S., Narendra, R.: Comparison of clustering techniques for reduction in energy consumption in wireless sensor networks. In: Conference Proceedings of NCRTE-2021, published in Shodhsamhita: J. Fundam. Comp. Res. **VII**(3), 2021 ISSN 2277-7067 (2021)
14. Homaei, M.H.: Low Energy Adaptive Clustering Hierarchy protocol (LEACH). <https://www.mathworks.com/matlabcentral/fileexchange/44073-low-energy-adaptive-clustering-hierarchy-protocol-leach>. MATLAB Central File Exchange (2021). Retrieved September 28 2021

15. Homaei, M.H.: SEP (A Stable Election Protocol) in Wireless Sensor Network. <https://www.mathworks.com/matlabcentral/fileexchange/44282-sep-a-stable-election-protocol-in-wireless-sensor-network>. MATLAB Central File Exchange. Retrieved Sept 28 2021
16. Tagare, T.S., Narendra, R., Manjunath, T.C.: A Brief Survey/Review of The Recent Advances in Wireless Sensor Networks Used in Communication Sectors, Conference Proceedings of NCRTE-2021, published in Shodhsamhita: Journal of Fundamental & Comparative Research Vol. VII, No. 3: 2021 ISSN 2277-7067 (2021)
17. Farooq, M. Y., Khan, K.B., Mohayy ud din, G., Rehman, E., Amin, S.: Energy Consumption of WSN Routing Protocols: LEACH, EAMMH and SEP, © Springer Nature Singapore Pte Ltd. 2020 I.S. Bajwa et al. (eds.) INTAP 2019, CCIS 1198, pp. 627–637 (2020)

# Academic Data Analysis and Projection Using Artificial Intelligence



K. Kanagaraj, Joyce R. Amirtharaj, and K. Ramya Barathi

**Abstract** Application of artificial intelligence (AI) in education is an emerging field of research. In a country like India with huge population, education cannot be provided without the participation of private institutions. However, many private institutions are unable to provide sustainable education as they lack the knowledge of the current industrial requirements and the quality of the students being admitted. AI can be used to build a complete road map for improving the performance of students in various directions using their existing academic data and to forecast the ways of improving the performance of the students for sustainable growth of the students and the institution. The proposed research applies AI methods to assist the academic institutes in formulating a necessary framework for making decisions towards sustainable education. The analysis involves different strategies to predict academic performance which comprises, collection of known data, data processing, generating training and testing datasets, building a model, and applying the model to the unknown data. Predictive algorithms are used to identify the most important attributes in academic data to suggest a prediction framework. The internal score and the health conditions are predicted based on the lunch provided to the students. Similarly, the AI model built on the collected data will be applied in the academic databases maintained by the NIC. As a result, more efficient measures can be identified to improve the academic performance of students in higher educational institutions and universities.

**Keywords** Artificial intelligence · Academic performance · Sustainable education · Self-learning

---

K. Kanagaraj (✉) · K. Ramya Barathi  
MEPCO Schlenk Engineering College, Sivakasi, India  
e-mail: [kanagaraj@mepcoeng.ac.in](mailto:kanagaraj@mepcoeng.ac.in)

J. R. Amirtharaj  
National Informatics Centre, Chennai, India  
e-mail: [joyce.tn@nic.in](mailto:joyce.tn@nic.in)

## 1 Introduction

AI and deep learning [1] comprise several algorithms and picking the right one or blend of algorithms, for the work is quite difficult for anybody working in this field. Be that as it may, before we inspect explicit calculations, comprehend the three general classifications of AI, the supervised learning, unsupervised learning, and reinforced learning. AI is totally subject to programming models.

Data mining is the method involved with finding and extracting information in the current datasets. It assists with finding stowed away examples in a huge dataset. Educational data mining (EDM) is the as of late developing discipline that assists with creating strategies in investigating particular sorts of information from training dataset and help to foresee students' scholarly performance. EDM is considered as learning science, just as a component of information mining. Foreseeing students' learning curve is a very difficult issue to address. Data mining provides an extremely essential support in foreseeing students' performance in their studies which will help in proposing upgrades to the existing system.

In this study, we have taken in to account the students undergoing their higher studies in universities. The academic performance of the students is predicted using neural network technique by analysing 27 relevant attributes such as the time spent on watching TV and Internet, time spent towards studying and sleeping, travel time, medium of education, and education level of their parents. Based on these parameters, we are able to recognize the influence of these parameters in their performance as well the most influencing parameter. Further, based on the values obtained, we can recommend the students to concentrate and improve their skills in the low-scoring parameters.

### 1.1 *Prediction Using Supervised Learning*

One of the most widely used and successful types of machine learning is supervised machine learning. When we wish to predict a specific outcome from a given input, we utilise supervised learning, and we have examples of input/output pairings. These input/output pairings, which make up our training set, are used to create a machine learning model. Our goal is to make precise predictions for novel, never-before-seen events. Supervised learning always necessitates more efforts from humans to create the training set, but it then automates and speeds up what would otherwise be a time-consuming or impossible activity.

## ***1.2 Unsupervised Learning Technique***

Unsupervised learning encompasses all types of machine learning in which the output is unknown, and there is no teacher to guide the learning algorithm. The learning algorithm is simply shown the input data and requested to extract knowledge from it in unsupervised learning.

## ***1.3 Reinforcement Learning Algorithm***

The latest and popular ML model is the reinforcement learning algorithm. Unlike the other two methods, the reinforcement learning uses input from earlier iterations to continuously improve its model. This makes it different and better than the other models in which no feedback is used for improvement. It is a very good example for learning models that improve through continuous learning. In a conventional reinforcement learning approach, outputs are graded rather than labelled, and performance requirements are quantified.

## **2 Literature Review**

Predicting students' academic performance using association rule mining was studied in [2] that talks about the application of rule mining in evaluating the performance of students. In [3], authors have used the decision tree algorithm that discovers the parameters that have a say in the academic performance of the students studying in universities. How Internet have influenced the other learning resources as well as the time. The influence of using the Internet as a learning resource and the effect of the social networks in sharing the students in learning time were discovered by [4] using artificial neural network (ANN). Predictive analytics using the decision tree along with the combination of neural networks to classify and group student learning patterns to describe their academic standard was explained in [5]. The correctness of various models which forecasts the students' educational performance and the one that contributes the most to better their e-learning outcome, taking into account the socio-economic and the demographic features of the students was studied by [6]. A useful study done by [7] to predict the chance of students being at risk (AR) or not 'Not at risk' (NAR) with respect to their degree is considered to be an informative work. However, there is a need to analyse the influence of other factors mostly indirect in nature that affects the performance of the students.

The effect of several data mining techniques and classifiers in forecasting the academic growth of the students was presented by [8, 9]. Baker et al. [10] analysed the contributions of the researchers towards educational data mining with respect to the present scenario. Identifying and selecting suitable features that are more

predominant in the evaluation process using machine learning were carried out by [4]. A good data mining approach based on lean production was done by [11]. The GALT proposed by [12] is a reasonable approach to predict the students' academic success level. A detailed overview about the role of emotional intelligence in transforming students from schools to colleges was presented by [13]. The works done by [14] and [15] deal with the frame work and an educational data mining approach, respectively.

An excellent review of the recent data mining techniques in predicting the academic performance of students was carried out by [16]. Reference [17] analysed the various factors that may put the students' performance at risk during their first year of graduation. The contribution of information technology towards identifying the upcoming performance of students and providing early recommendations was studied by [18]. The use of data mining algorithm in big data analysis and perturbations resented in [19] is a very useful work related to our research. The work proposed in [20] is very helpful to detect the possibilities of depression in students at early stage using social media. An improved sigmoid function presented in [21] is an important contribution to classify tasks in the ELM domain.

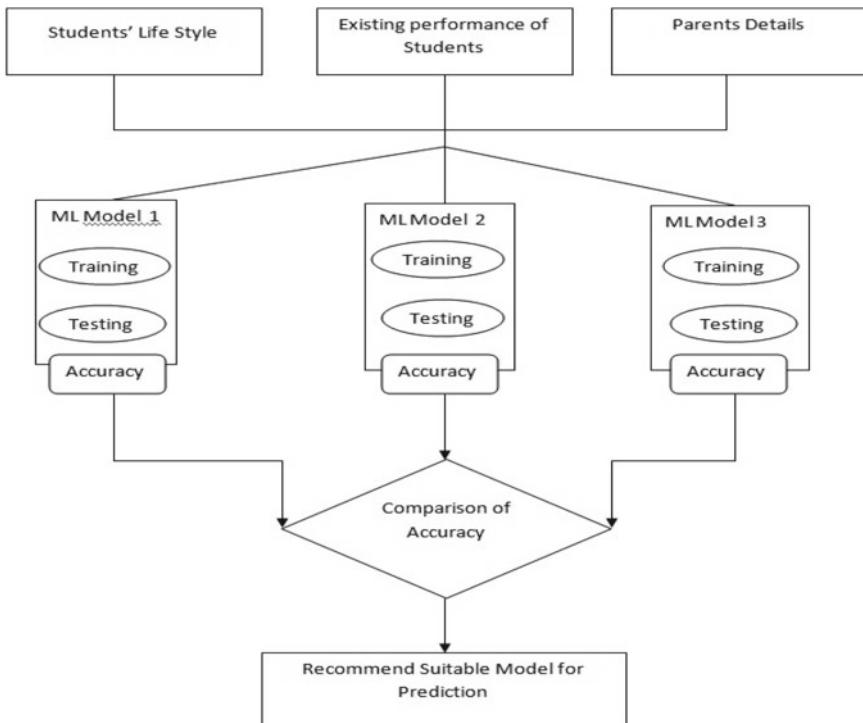
### 3 Overall Architecture of the Proposed Work

The proposed model contains three phases. At the first phase, data collection is carried out, in which data about the lifestyle of the students, existing students' performance, and the parent's details are collected. At the second phase, the gathered data is trained and tested against different ML models [1] and accuracy of the models is determined. At the final phase, the accuracy score of all the models are compared and best model is recommended as the suitable one to predict the performance of students. The overall architecture of the proposed work is presented in Fig. 1.

Data collection is done by requesting the students and other stake holders to fill the details in three different ways such as filling the form in person, filling a Google Form and filling the details in the Google Sheet. The form contains suitable questions to acquire data for all the required parameters. Also, the details of their parents such as their educational qualification, age, occupation, place of residence, and average monthly salary will be collected. After data collection, they are converted into .csv format for feeding to the ML models. The data is carefully split into train and test data usually 80% and 20%, respectively. The accuracy of all the ML models is recorded by varying the training and testing data split to obtain maximum accuracy. Finally, the maximum accuracy score obtained from various training models will be compared to select the best method for the prediction process.

The objectives of the proposed work are

- Gathering the living style and current performance of the students studying in universities.
- Applying different machine learning models and comparing their accuracies based on the most influencing parameter.



**Fig. 1** Overall architecture of the proposed work

- Recommending the model with high accuracy for predicting the student's academic performance and taking corrective actions at the early stage.

## 4 Implementation and Results

In the proposed work, artificial intelligence is a based predictive algorithm is used to identify the most important attributes in academic data. Prediction analysis of student performance and prediction analysis of institution/department/course performance are done. To predict academic performance, a strategy which comprises collection of known data, data processing, generating training and testing datasets, building a model, and applying the model to the unknown data is applied. Predictive algorithms are used to identify the most important attributes in academic.

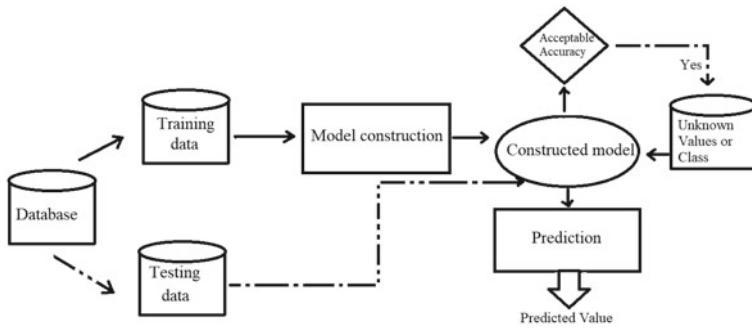
Among the several existing models, decision tree model, neural network, and naïve Bayes model are used in this research. A decision tree (DT) classifier brings out the association among the various nodes, branches, and children in the sense of the structure of a tree and association rules for natural language processing. The iterative dichotomiser or classification and regression trees algorithms can be used

by the DT classifier. It is a recursive algorithm that uses a top-down greedy search to scan the attribute space for feasible branches without going backwards. It builds the decision tree using a predetermined set of training samples and then prunes it. To classify the qualities, DT algorithms rely on entropy. Entropy is a mathematical computing technique that gives heterogeneous samples a binary numerical value of 0 and homogeneous samples a binary numerical value of 1. Random forest and K-nearest neighbour (KNN) algorithms give the best accuracy score for the collected student dataset when compared to the other competent algorithms in this field.

#### **4.1 Process Involved in Prediction**

- Data collection from institution and courses.
- AI-based predictive analysis of collected data using methods identified in literature survey.
- Creation of postgresql database with master tables for the collected dataset.
- Identification of predictive methods best suited for academic data available with National Informatics Centre (NIC).
- Improving prediction accuracy of the identified methods by attribute selection and parameter modifications.
- Integration of suitable AI-based predictive analysis methods.
- Creation of web application for the analysis using Windows-Apache-Postgresql-PHP/Linux-Apache-Postgresql-PHP based.
- Make the method as a module which can be integrated with NIC developed Web applications.

Artificial intelligence methods are chosen to apply and make modifications on the method according to the student dataset. Student data table is created with several master tables linked with it. Several machine learning methods were applied, and best method is chosen. According to the student dataset, the columns are categorized, and predictions are made on that dataset by using the method which gives the best accuracy of result. The predicted result for the student dataset is maintained in a website which can be referred later by the university authorities for improving the academic performance of each student in the institution. A Web application is created for the COE to look at the analysis made on the student data using Windows-Apache-Postgresql-PHP/Linux-Apache-Postgresql-PHP. The prediction result can then be imported as a module to the NIC developed websites for predicting students' performance. Figure 2 shows the data flow between various modules.



**Fig. 2** Data flow between modules

#### 4.2 Data Flow Between Modules

This study utilises a dataset from the National Informatics Centre, which contains 1440 student records. All of this information regarding student behaviour is used to build the set of features that make up the predictive system's input, which is based on a machine learning algorithm. There are 27 attributes in the dataset (Table 1).

#### 4.3 Training and Testing Data

- **Training set**—A training set is used to train a model and identify its optimal parameters, which are the parameters that the model must learn from data.
- **Test set**—A test set is required for evaluating the trained model's generalisation capability. The latter refers to a model's ability to spot patterns in previously unseen data after it has been trained on it.
- **Validation set**—Validation set is used to fine-tune a model's hyperparameters, which are higher-level structural settings that cannot be learned directly from data. These options can indicate how complicated a model is and how quickly it detects patterns in data.

As shown in Fig. 3, the proportions of the training and test sets are normally 80–20%. After that, a training set is split again, and the remaining 20% is utilised to create a validation set.

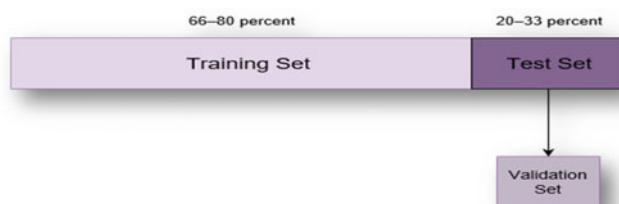
Splitting the data into training and testing sets

```

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33,
random_state=42)
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
  
```

**Table 1** Attributes of student dataset

S. No.	Attributes	Description
1	university_register_number	Assigns unique register number for each student
2	Section	Describes the students section
3	student_name	Stores the name of the student
4	Gender	Stores the gender of the student
5	Age	Stores the age of the student
6	Program	Describes the program the student is undergoing E.g., B.Sc., B.C.A., M.Sc., M.C.A., M.E.,
7	physical_activities	List of physical activities the students is participating
8	history_of_arrears	The details of arrears of the student
9	race_ethnicity	The
10	Lunch	The choice of students lunch E.g., Veg, non-veg
11	internal_marks	The score of the student in internal assessment
12	external_marks	The score of the student in external assessment
13	assignment_score	The mark scored by the student in assignment
14	tv_usage	The time spent by the student in watching TV/ day
15	internet_usage	The time spent by the student in surfing Internet/day
16	sleep_hours	The number of hours the student sleeps
17	study_hours	The number of hours spent in studying subjects
18	dayScholar_hosteller	Whether hosteller or day scholar
19	travel_time	The time spent in travelling
20	Health	The health condition of the student
21	Attendance	Regularity level in attending classes
22	parent_guardian	Name of the parent or guardian
23	parent_level_of_education	The education level of the parent or guardian
24	Medium	The medium of education of the parent
25	home_ownership	Whether the parents are staying in own/rental house
26	marital_status	Marital status of the student
27	student_class	Year of study of the student

**Fig. 3** Training and testing data split percentage

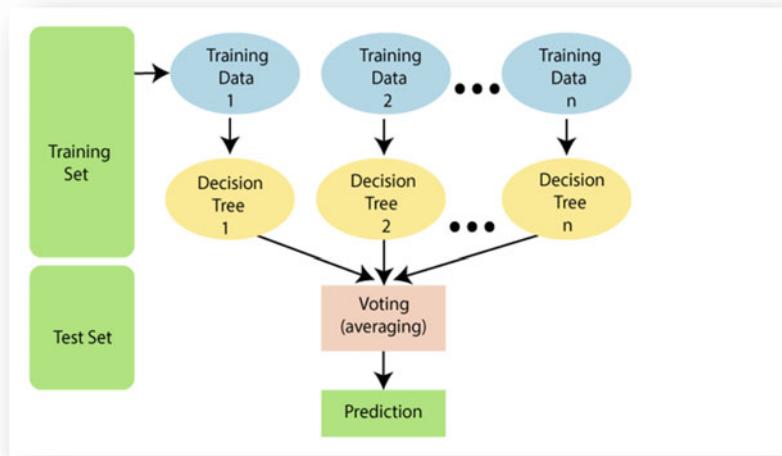
#### 4.4 The Methodology

A brief overview of the machine learning algorithms employed in this study is presented in this section.

#### 4.5 Random Forest Algorithm Implementation

Random forest is a well-known machine learning algorithm that uses the supervised learning method. In machine learning, it can be utilised for both classification and regression issues. It is based on ensemble learning, which is a method of integrating several classifiers to solve a complex problem and increase the model's performance. There are a number of benefits to go for random forest algorithm, but one of the most focal points is that it diminishes the risk of over fitting and the specified preparing time. Also, it offers elevated level of exactness. It runs proficiently in huge databases and produces exceedingly exact forecasts by evaluating lost information.

- Random forest is a classifier that combines a number of decision trees on different subsets of a dataset (as illustrated in Fig. 4) and averages the results to increase the dataset's predictive accuracy. Instead of relying on a single decision tree, random forest collects the forecasts from each tree and predicts the final output based on the majority votes of predictions. Let's have a look at the steps in the random forest algorithm:



**Fig. 4** Random forest algorithm implementation

- **Data Processing**

In data processing, the student data is read, and the preparation of the student data is done.

```
"Read the data"
df = pd.read_csv(r'C:\dataset\student-dataset.csv')
# Any results you write to the current directory are saved as output.
df.head(10)
print("the shape of our dataset is {} ".format(df.shape))
#Data preparation
df.info()
df.describe()
df.nunique()
```

- **Fitting Random Forest Algorithm to Training set**

The random forest algorithm will now be fitted to the training set. We'll use the random forest classifier class from the `sklearn.ensemble` package to make it fit. The code can be found below.

```
from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier(n_estimators = 50)
classifier.fit(X_train, y_train)
```

- **Predicting test set result**

We can now anticipate the test outcome because our model has been fitted to the training data. We'll make a new prediction vector, `y pred`, for prediction. The code for it is as follows:

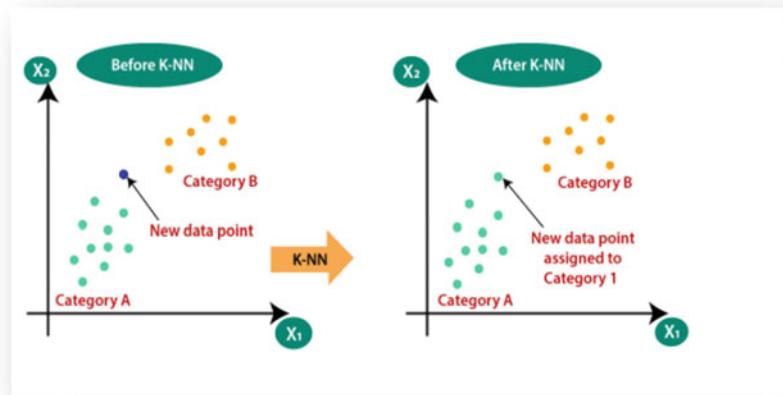
$$y\_pred = classifier.predict(X\_test)$$

- **Creating Confusion Matrix**

Now, we'll make the confusion matrix to figure out which forecasts are true and which are incorrect. The code for it is as follows:

```
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
result = confusion_matrix(y_test, y_pred)
print("Confusion Matrix:")
print(result)
```

The accuracy score of random forest algorithm is: ***0.9684873949579832***.



**Fig. 5** K-nearest neighbour (KNN) algorithm implementation

#### 4.6 K-Nearest Neighbour (KNN) Algorithm Implementation

The K-nearest neighbour algorithm is based on the supervised learning technique and is one of the most basic machine learning algorithms. The KNN architecture is seen in Fig. 5. It assumes that the new case/data, and existing cases are similar and places the new case in the category that is most similar to the existing categories.

It's also known as a lazy learner algorithm since it does not learn from the training set right away; instead, it saves the dataset and performs an action on it when it comes time to classify it. The steps involved in KNN algorithm are,

##### Data Processing

In data processing, the student data is read, and the preparation of the student data is done.

```
"Read the data"
df = pd.read_csv(r'C:\dataset\student-dataset.csv')
# Any results you write to the current directory are saved as output.
df.head(10)
print("the shape of our dataset is {}".format(df.shape))
#Data preparation
df.info()
df.describe()
df.unique()
```

- **Fitting KNN Algorithm to Training set**

We'll now apply the KNN algorithm to the training data. We'll use the K neighbour classifier class from the sklearn neighbours library to make it work. The code can be found below:

```
from sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors = 8)
classifier.fit(X_train, y_train)
```

- **Predicting test set result**

We can now predict the test outcome because the suggested model is fitted to the training data. For prediction, we'll make a new prediction vector called *y pred*, with the following code:

$$y\_pred = classifier.predict(X\_test)$$

- **Creating Confusion Matrix**

Using the code provided below, we will now generate the confusion matrix to identify the accurate and erroneous predictions.

```
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
result = confusion_matrix(y_test, y_pred)
print("Confusion Matrix:")
print(result)
```

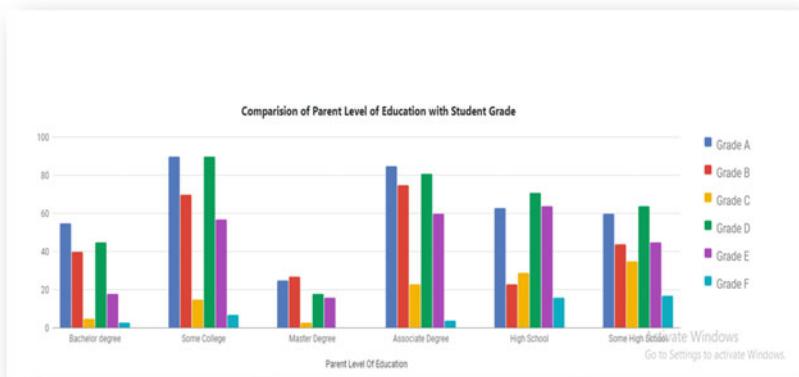
The accuracy score of KNN **Algorithm is: 0.8676470588235294**.

The accuracy of the different algorithms for the same set of data is compared and plotted in Table 2. The random forest algorithm is found to have better performance than the other algorithms and is recommended as the best technique for predicting student's academic performance.

Several interesting observations are made from the acquired result. Figure 6 shows the impact of parent's level of education in the grade obtained by the students. It is

**Table 2** Accuracy score of different prediction algorithms

Name of the algorithm	Number of samples used	Percentage used for training (%)	Percentage used for testing (%)	Accuracy score
Decision tree classifier	1440	80	20	<b>0.72</b>
Random forest algorithm	1440	80	20	<b>0.96</b>
KNN algorithm	1440	80	20	<b>0.86</b>

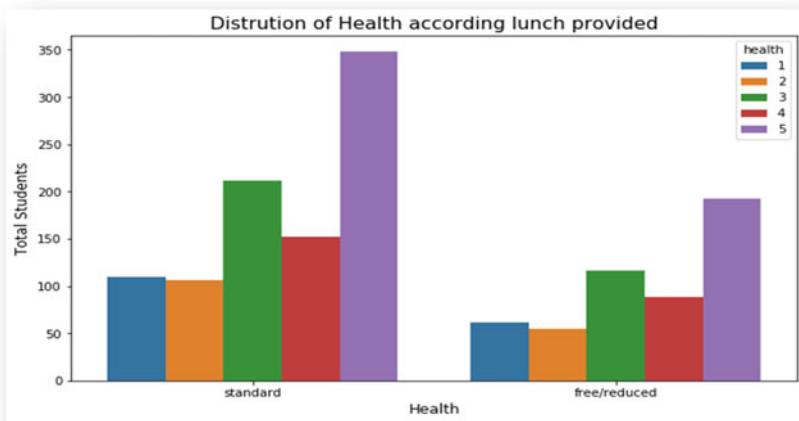


**Fig. 6** Comparison of parent's level of education with the student's grade

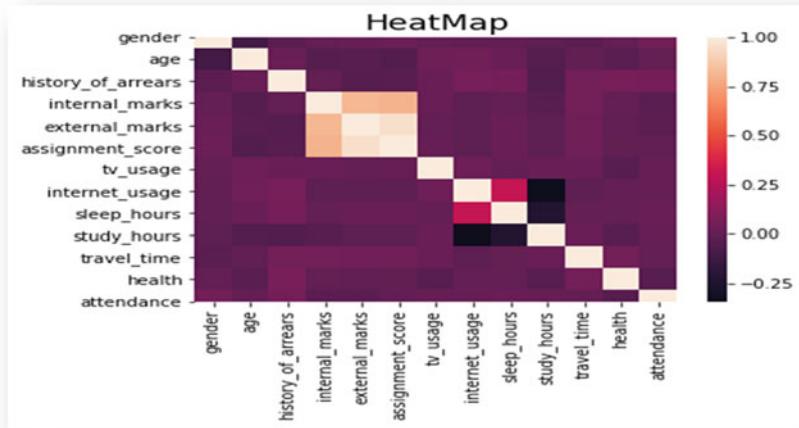
also found that there is even variation in the student's grade with the difference in the nature of job of the parents having similar educational level.

From 10, we found that the wards of the parents who had some collegiate education secure higher grades and the wards of the parents having master's degree score lesser grades.

The impact of the difference in lunch provided to the students in their grade is presented in Fig. 7. This indicates that students taking standard lunch are performing better than the students having lunch under reduced or free meals scheme (Fig. 8).



**Fig. 7** Distribution of student health according lunch provided to them

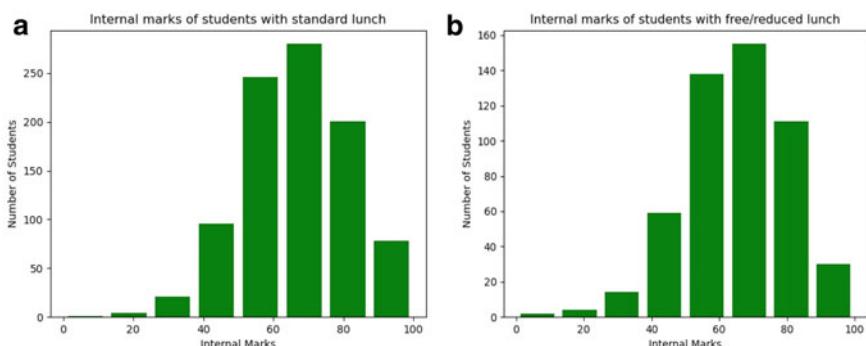


**Fig. 8** The graphical representation of student data and behaviour

The influence of lunch provided to the students at standard rate as well as subsidized rate are studied and presented in Figs. 9a, b.

From Figs. 9a, b it is observed that the number of students scoring less than fifty belongs to the lunch with free/reduced cost.

Educational institutes' major goal is to deliver high-quality education to its students and to increase the quality of the students. The proposed system helps to discover knowledge from educational dataset to examine the primary factors that may affect student performance. This technique helps higher educational institutions to take corrective actions to overcome the disruptive factors that influence the pupil's education at the earlier stage. The AI model adopted in this system uses the knowledge gained from various stake holders to provide helpful and constructive



**Fig. 9** **a** Distribution of internal marks of 927 students taking lunch at standard rate, **b** distribution of internal marks of 513 students taking lunch at free/reduced rate

recommendations to academic planners in higher educational institutions in order to improve their decision-making process, improve student academic performance, reduce failure rates, better understand students' behaviour, assist instructors, improve teaching, and many other benefits.

## 5 Conclusion and Future Enhancements

Machine learning techniques for predicting student performance have proven to be useful for identifying poor performers and enabling tutors and administrators to take remedial measures at an earlier stage. Predicting performance using the recent AI-based technologies as proposed in this work would allow institutions to focus more on students who are more likely to perform poorly in examinations. The ability to forecast a student's progress aids the educational organization in providing timely support to the student and also helps the growth of the educational system in educational institutions. In future, the system can be improved by increasing the number of records in the dataset to get more accuracy. The training can also be extended by applying other popular machine learning models to explore the influencing parameters unidentified by the current system. Also, more parameters can be added to the existing dataset to help administrators and professors to intervene early to help and assist students in the bad and average categories to improve their grades. This study will also be enhanced in the future by adding more data from various years and different universities.

## References

1. Klaise, J., Vacanti, G.: Alibi explain: algorithms for explaining machine learning models. *J. Mach. Learn. Res.* **22**, 1–7 (2021)
2. Borkar, S., Rajeswari, K.: Attributes selection for predicting students' academic performance using education data mining and artificial neural network. *Int. J. Comput. Appl.* **86**(10), 25–29 (2014). <https://doi.org/10.5120/15022-3310>
3. Kuye, G., Adeyemo, A.B.: Mining students' academic performance using decision tree algorithms. *J. Inf. Technol. Impact* **6**(3), 161–170 (2006)
4. Altabrawee, H., Ali, O.A.J., Ajmi, S.Q.: Predicting students' performance using machine learning techniques. *J. Univ. BABYLON Pure Appl. Sci.* **27**(1), 194–205 (2019). <https://doi.org/10.29196/jubpas.v27i1.2108>
5. Alloghani, M., Al-Jumeily, D., Hussain, A., Aljaaf, A.J., Mustafina, J., Petrov, E.: Application of machine learning on student data for the appraisal of academic performance. In: Proceedings—International Conference on Developments in eSystems Engineering, DeSE, vol. 2018-Septe, no. September, pp. 157–162, 2019. <https://doi.org/10.1109/DeSE.2018.00038>
6. Ofori, F., Maina, E., Gitonga, R.: Using machine learning algorithms to predict students' performance and improve learning outcome: a literature based review Francis Ofori, Dr. Elizabeth Maina and Dr. Rhoda Gitonga. *J. Inf. Technol.* **4**(1), 33–55 (2020). ISSN: 2617-3573
7. Ahmad, Z., Shahzadi, E.: Prediction of students' academic performance using artificial neural network. *Eric* **40**(3), 157–164 (2018)

8. Ahmad, F., Ismail, N.H., Aziz, A.A.: The prediction of students. *Acad. Perform. Using Classif. Data Min. Tech.* **9**(129), 6415–6426 (2015)
9. Almarabeh, H.: Analysis of students' performance by using different data mining classifiers. *Int. J. Mod. Educ. Comput. Sci.* **9**(8), 9–15 (2017)
10. Baker, R.Y.A.N.S.J.D., Blum, A.L., Langley, P.: The state of educational data mining in 2009: a review and future visions. *J. Educ. Data Min.* **5**(8), 3–16 (2009)
11. Bragança, R., Portela, F., Santos, M., Bramer, M.: A regression data mining approach in lean production. *Concurr. Comput. Pract. Exp.* **31**(22), 4449 (2019)
12. Bunce, D.M., Hutchinson, K.D.: The use of the GALT (group assessment of logical thinking) as a predictor of academic success in college chemistry. *J. Chem. Educ.* **70**(3), 183 (2009)
13. Parker, J.D., Hogan, M.J., Eastabrook, J.M., Oke, A., Wood, L.M.: Emotional intelligence and student retention: predicting the successful transition from high school to university. *Pers. Individ. Dif.*, **41**(7), 1329–1336 (2006)
14. Peng, Y., Kou, G., Shi, Y., Chen, Z.: A descriptive framework for the field of data mining and knowledge discovery. *Int. J. Inf. Technol. Decis. Mak.* **7**(4), 639–682 (2008)
15. Romero, C., Ventura, S.: Educational data mining: a review of the state of the art. *IEEE Trans. Syst. Man, Cybern. Part C (Appl. Rev.)* **40**(6), 601–618 (2010)
16. Shahiri, A.M., Husain, W., Rashid, N.A.: A review on predicting Student's performance using data mining techniques. *Procedia Comput. Sci.* **72**, 414–422 (2015)
17. Willems, J., Coertjens, L., Tambuyzer, B., Donche, V.: Identifying science students at risk in the first year of higher education: the incremental value of non-cognitive variables in predicting early academic achievement. *Eur. J. Psychol. Educ.* **34**(4), 847–872 (2019)
18. Yassein, N.A., Helali, R.G.M., Mohomad, S.B.: Information technology & software engineering predicting student academic performance in KSA using data mining techniques. *J. Inf. Technol. Softw. Eng.* **7**(5), 1–5 (2017)
19. Haoxiang, W., Smys, S.: Big Data analysis and perturbation using data mining algorithm. *J. Soft Comput. Paradig.* **1**(3), 18–28 (2021)
20. Smys, S., Jennifer, S.R.: Analysis of deep learning techniques for early detection of depression on social media network-a comparative study. *J. trends Comput. Sci. Smart Technol.* **1**(3), 24–39 (2021)
21. Mugunthan, S.R., Vijayakumar, T.: Design of improved version of sigmoidal function with biases for classification task in ELM domain. *J. Soft Comput. Paradig.* **2**(3), 70–82 (2021)

# Analysis of Factor Verification Affecting Recruitment Process Through Social Dynamics



Krishna Kumar Singh and Priyanka Srivastava

**Abstract** Verification of candidate profile is most challenging job for HR professional as it involves a lot of time and other resources. Verification of credentials needs to be done before the job offer is made, and if any discrepancy is found at a later stage, the entire process of recruitment loses its credibility. This can result in a loss of money, time and other resources for the organizations. Graph-based analysis using NoSQL software has proved to be extremely helpful in the verification of the credentials at the very beginning of the recruitment process, thereby saving a lot of resources and enhancing the credibility of the process. Graph databases help in identification of patterns in semi-structured and unstructured data. By analysing the data gathered from social media platforms, with the help of graph-based tools, deeper and newer insights about candidate's profile can be generated. Nodes of the graph-based NoSQL can store various heterogeneous data for map, analyse it and create connection between datasets. HR analytics can further help recruiters in intelligent decision-making and in improving the efficiency and quality of talent acquisition process. Various factors like demographical information, educational qualification, employment history, behaviour, communication records, etc. are taken into consideration for deciding the candidature and suitability of the prospective employee. Objective of the paper is to provide a model to verify and validate candidate's profile before recruitment process begins. Keeping this objective in mind, authors propose a graph-based model for analysing this heterogeneous data used during recruitment process with the help of nodes for better, quick and authentic verification of data given by the candidate. Authors used algorithms like clustering and association to establish relationship among nodes and develop various insights about the suitability and fit of the candidate. This model would be helpful for recruiters in differentiating and segregating candidates based on graphical prediction and would also come

---

K. K. Singh   
Symbiosis Centre for Information Technology, Pune, India  
e-mail: [krishnakumar@scit.edu](mailto:krishnakumar@scit.edu)

P. Srivastava  
Amity University, Noida, India  
e-mail: [psrivastava7@amity.edu](mailto:psrivastava7@amity.edu)

handy in sorting the huge list of applicants quickly and accurately. At last, authors also discuss the limitations of the model.

**Keywords** Graph · Social connectivity · Verification · Predictive modelling · NEO4J · Cybervetting · HR analytics · Analytics

## 1 Introduction

According to Marc Benioff, ‘Acquiring the right talent is the most important key to growth. Hiring was—and still is—the most important thing we do’. Attracting and recruiting talented candidates who are culturally fit with the organization is very important for the success of any organization. As social media has become a significant part of people’s lives, it is also emerging as a relevant tool to not only connect with potential candidate but also to ease out the screening and background checks. The pervasiveness and power of social media networks is more than obvious. According to ‘DIGITAL 2021: GLOBAL OVERVIEW REPORT’, the number of active social media users worldwide is 4.20 billion which is more than 53% of the total population of the world. The number has gone up by 13% as compared to the last year. The average time that a user spends on social media every day is 2 h and 25 min. Global Web Index(Q3,2020) says that a millennial or Gen Z-er has, on an average, 8.8 social media accounts (which is up 83.33% from 4.8 accounts in 2014) and India topped the list with 11.5 social media accounts per person, in the survey of Internet users of 46 countries. The information available on social media is easily accessible and visually appealing. A glance through the social media engagement of the individuals is enough to provide important insights into his/her beliefs, behaviour and reputation. Because of these factors, social media has assumed greater relevance in various aspects and stages of HR functions especially with respect to recruitment and background verification process. Background verification through social media involves analysing the profile information, social media posts and studying his/her social network and engagements on social media platforms to check the authenticity of the information provided in the resume and also to understand if he/she is culturally fit with the organization. For example, a huge number of users share the information about the places they visited or their travel details on Facebook. Their ‘Likes’, ‘Dislikes’, pages or groups—they are member of—everything gives an idea about the personality, hobbies and interests of an individual and can be very useful for HR in assessing his/her candidature objectively. Although using social media data for background checks is not without challenges and obstacles. Graph theory-based analytics is helpful in establishing relationship in the nodes created with the heterogeneous dataset acquired from different sources including social media. However, the number of organizations using online information for the purpose of recruitment and verification is bound to surge as social media is increasingly becoming an integral part of our social lives and is intricately woven into the phenomenon of social networking. A survey by CareerBuilder says that 70% of the employer’s research

about candidates using social networking sites during hiring process and about 48% of the employers engage in continuous checking up of the current employees on social platforms; 34% of the employers accepted that they have fired or reprimanded the employees on the basis of content found online [1]. Clustering based on the similarities is helpful in the character analysis and psychological analysis. Clustering also used to verify data and credentials of the candidates as per requirement. It can be assumed that social media verification can contribute hugely to a comprehensive background verification process by corroborating the findings in other checks, as well as, by providing new information.

## 2 Literature Review

With the enhanced emphasis on hiring the right kind of employees who are culturally fit with the organization, a comprehensive, accurate and objective assessment of prospective employees has also assumed greater significance. Internet gives individuals access to huge amount of information on any topic they desire and provides them a tool to get insights into their friends, family members, prospective employees, etc. [2]. The practice of utilizing social media data for enhancing the quality of recruitment has been prevalent among organizations [3]. Few other studies have also corroborated rapid growth in the use of social media in staffing and emphasized the increasing relevance of platforms like Facebook, LinkedIn and Twitter in recruitment [4, 5]. According to a study by Narvey, social media has become the lifeline of how communities engage and create meaningful relationships [6]. It shows how deeply social media is entrenched in our lives. Social network of individuals can include interpersonal ties with friends, family members, colleagues, professional contacts, etc. These interpersonal connections come handy while deciding if there is a fit between job description and the profile of the candidate [7]. Bartakova et al. have also acknowledged the huge potential of social media for recruitment [8]. An earlier study by Gundecha and Liu stated that organizations, trying to leverage social media for recruitment, face challenges because of the dynamic, unstructured and scattered nature of the data available on social media [9].

You are responsible for everything you post and everything you post will be a reflection of you. [Social Media]

—Germany Kent

A number of visualization approaches and tools are now being used for exploring the hidden connections and revealing the insights which may be difficult to understand from the traditional methods. Graph representation is one of them. It helps in analysing in huge datasets by visualizing and studying them as connected graph [10]. The network graph or node-link diagram, where nodes refer to the elements in the network and edges refer to connections between these nodes, helps in visualizing the data in totality thereby bringing more clarity about the relationship between categories and element; some of the nodes in a network play more decisive role than

others [11]. These are known as central and are characterized by certain metrics referred as centrality measures. These centrality measures indicate the relative relevance of nodes and edges in their study titled, ‘Using graph theory and social media data to assess cultural ecosystem services in coastal areas: Method development and application’, observe that relative ease of application and shorter processing time makes it a very cost-effective method which can smoothly be applied to an expanded geographical scale [12]. Few recent studies assert the importance of knowledge graph in instant screening and significant reduction in the time and effort involved in manual screening of the profiles for finding the best fit with the job, during a recruitment process and provide relevant insights for the HR specialists [13]. A lot of personal information is shared by people on social media which gives a glimpse into their personalities, hobbies, values, religious beliefs and political ideologies [14]. In another study, 89% of the recruiters said that if they come across any proof of ‘unprofessional behaviour’ on social media profile of the prospective employees, it can lead to the person not getting hired [15]. It has also been established that connectivity and interfaces on social media platforms impact shaping of the online identities of the people [16]. Research shows that reliability and accuracy of the information related to the personality of people, which has been filtered out from their social media profiles, using the data analytics tools. The most important use of social media for background verification is to validate the credentials and identify the misleading information in the resumes. The organizations tend to access the information available on the social media to evaluate if the candidate is a suitable to be hired or if there is any ‘red flag’ about the individual that makes him unfit for the organization in the long run, like any act of vulgarity, evidence of sexually implicit actions, consumption of prohibited drugs or alcohol, etc. [15, 17]. In a study, few of the employers surveyed, found ‘cybervetting’ to be transformational because it helped them to find the real fit with the organization and authenticate the ‘real person’, they were hiring ‘providing the opportunity to piece together exactly who the applicants are, cutting through layers of impression management and possible deceit’ [18]. In the same study, thirty-five of the total forty-five participants opined that ‘cybervetting’ is more efficient, less time-consuming, and less expensive than other conventional tools.

The goal is to turn data into information, and information into insight.

—Carly Fiorina, former executive, president, and chair of Hewlett-Packard Co.

There has also been a lot of development in the field of predicting the personality of an individual on the basis of social media content analysis and then evaluating his/her suitability for the job. A software created by Leite, Palgon and Vila gathers the information, about a profile, from the social networking connections, creates verification score and relays the information along with verification score to a set of profile consumers, thereby helping them in validating the information provided by candidates [19]. Another application developed by Khan, ‘Perfect match’, allows jobseekers access to a ranked list of matching jobs and recruitment professionals, a ranked list of matching resumes. Zide et al. studied the relationship between the content people share on their LinkedIn accounts and their occupations for three

	employee_id	city	city_development_index	relevant_experience	enrolled_university	education_level	major_discipline	experience	company_size	company_type	last_new_training_hours
2	32403	city_41	0.827	Male	Has relevant experience Full time course	Graduate	STEM	9 <10		1	21
3	9858	city_103	0.92	Female	Has relevant experience no_enrollment	Graduate	STEM	5	Pvt Ltd	1	98
4	31806	city_21	0.624	Male	No relevant experience no_enrollment	High School		<1	Pvt Ltd	never	15
5	27385	city_13	0.827	Male	Has relevant experience no_enrollment	Masters	STEM	11	Oct-49	Pvt Ltd	1
6	27724	city_103	0.92	Male	Has relevant experience no_enrollment	Graduate	STEM	>20	10000+	Pvt Ltd	>4
7	21734	city_23	0.899	Male	No relevant experience Part time course	Masters	STEM	10			2
8	21465	city_21	0.624		Has relevant experience no_enrollment	Graduate	STEM	<1	100-500	Pvt Ltd	1
9	27302	city_160	0.92	Female	Has relevant experience no_enrollment	Graduate	STEM	>20			>4
10	12994	city_173	0.878	Male	Has relevant experience no_enrollment	Graduate	STEM	14			4
11	16287	city_21	0.624	Male	Has relevant experience Full time course	Graduate	STEM	3-50-99	Funded Startups	1	4
12	10856	city_103	0.92	Male	Has relevant experience no_enrollment	Masters	Other	>20			>4
13	9272	city_90	0.698	Male	Has relevant experience no_enrollment	Graduate	STEM	20	Oct-49	Pvt Ltd	2
14	14249	city_46	0.762	Male	Has relevant experience no_enrollment	Graduate	STEM	8-100-500	Other	never	48
15	24372	city_98	0.949		Has relevant experience no_enrollment	Masters	STEM	4-100-500	Pvt Ltd	1	134
16	14070	city_103	0.92		No relevant experience no_enrollment	Graduate	STEM	5		never	10
17	24914	city_21	0.624		Has relevant experience Full time course	Graduate	STEM	13-1000-4999	Pvt Ltd	1	125
18	7865	city_21	0.624	Male	Has relevant experience no_enrollment	Masters	STEM	4-100-500	Pvt Ltd	1	4
19	7463	city_13	0.827	Male	Has relevant experience no_enrollment	Masters	Business Design	2-50-99	Pvt Ltd	1	31
20	21514	city_21	0.624		Has relevant experience no_enrollment	Graduate	STEM	6	Pvt Ltd	4	23
21	29033	city_21	0.624	Male	No relevant experience Full time course			2		never	110
22	15359	city_103	0.92		No relevant experience Full time course	Graduate	STEM	2		never	74
23	16001	city_103	0.92		Has relevant experience no_enrollment	Graduate	STEM	7-10000+		1	44
24	25202	city_21	0.624	Male	Has relevant experience no_enrollment	Graduate	STEM	8-1000-4999	Pvt Ltd	3	33

**Fig. 1** Data after pre-processing

domain groups—HR, sales/marketing, and industrial-organizational psychologists and concluded that out of the three groups, people belonging to sales/marketing profiles were highly into networking [20]. Few other researches also point out towards an approach of predicting the personality of an individual on the basis of the analysis of the content available on his/her social media account. Few HR practitioners find it inappropriate because of the lack of standardization [21].

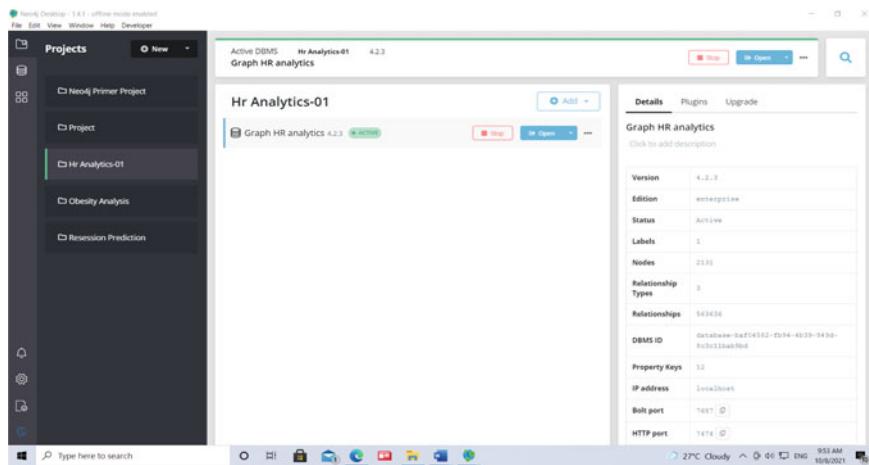
### 3 Pre-processing of Data

Recording of data has been done through survey from employees who applied for the job and sourcing data from company HR. As per company policies, authors are restricted from revealing company identity. As per requirements, authors collected data related to employee's professional life only. Data contains employee id, city, city development indexed, gender, experience, university they enrolled, education level, experience, company size, domain, company type, recent job, training hours, etc. Initially, data had some ambiguities and was available, both in qualitative and quantitative nature (Fig. 1). Authors removed all possible ambiguities by following data pre-processing methods with the help of Python and final results after pre-processing of data is given below.

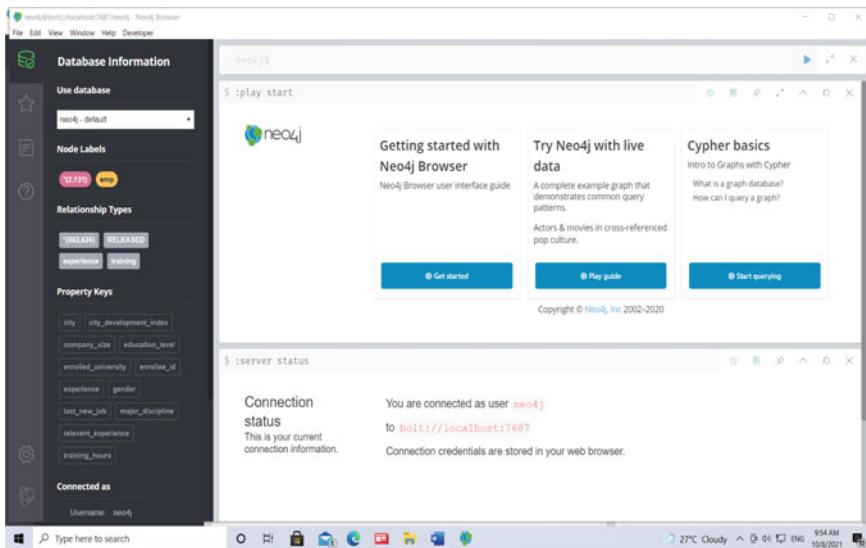
### 4 Model Building and Implementation

Flow of the model is being describe in five stages with the help of NoSQL tool called Neo4J. First step is to collect data from organization which are interested to mine various insights of the prospected applicants applied for the job and data from social media related to the prospective employees. There are various limitations of

the data collections as law of the land and may create hindrance. Second step is to pre-process data according to the requirement of tools and analysis. This is explained in Figs. 2 and 3. Authors imported data in Neo4J and created nodes in it. Each row in the table will become one node which contains all parameters (employee id, city, city development indexed, gender, experience, university they enrolled, education level, experience, company size, domain, company type, recent job, training hours, etc.) of a candidate. Third step is to finalize various parameters (conditions) of the candidate to discover various insights from his/her profile to judge the suitability for the job. Fourth step involves the implementation with the help of tools. In this model, authors have used Neo4J. It is most advance tools to work with graph-based model with heterogeneous datasets. Fifth and the last step is to get analytical results and then interpreting those results. Neo4J cypher code of Neo4J, link and relationship between nodes were created which satisfied desired conditions like friendships. With the help of connectivity of nodes and establishing relationship among people, authors were able to establish core relationship among people's credentials and organizational goals. There should be an alignment between the prospective employees' credentials and the organizational goals and vision. Benchmarking of the whole HR process is equally important for HR professionals. Graph helps in understanding skillsets of the prospective candidate with respect to the other professionals having same experience in the open market. Social media data and its connectivity is very helpful in that process. Past behaviour of the prospective employees is being verified with the help of relationship with other employee in that organization. Friendships are formed between people with similar mindsets as like personalities attract each other, and HR professional can have better understanding and can make more accurate predictions about the psychological behaviour of the person with the help of relationships established through the data. Prediction of longevity and ROI will become easy



**Fig. 2** Model formation in Neo4J



**Fig. 3** Model formation in Neo4J

with the patterns identified from the data through graphical networks. By keeping these theoretical beliefs, authors built graphical model with the help of NoSQL.

Nodes with headers and conditions were created with the help of cypher query (Figs. 4 and 5). All cypher queries are visible on the top of working area. Number of total possible nodes with the conditions mentioned in queries are visible at the bottom of the working areas. Explanation of nodes and heterogeneous data stored by nodes are visible in Figs. 6 and 7.

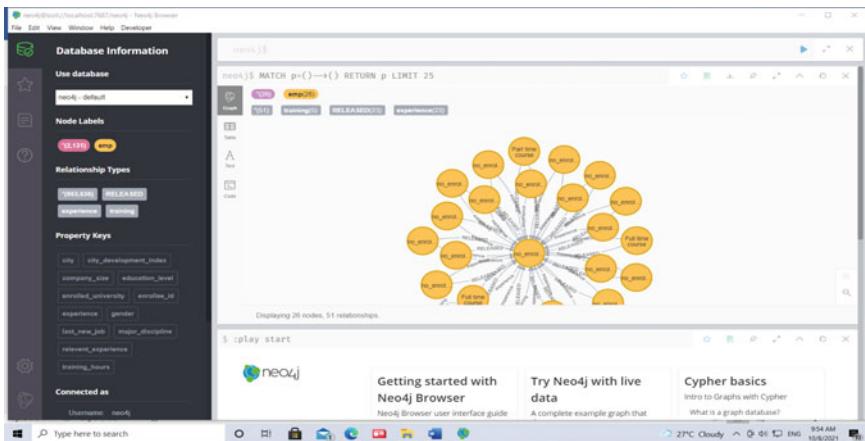
The screenshot shows the results of a Cypher query run in the Neo4j Browser. The query is:
`match (n) WHERE EXISTS(n.city) RETURN DISTINCT "node" as entity, n.city AS city LIMIT 25 UNION ALL MATCH ()->()->()`

The results table lists 25 entities and their corresponding cities:

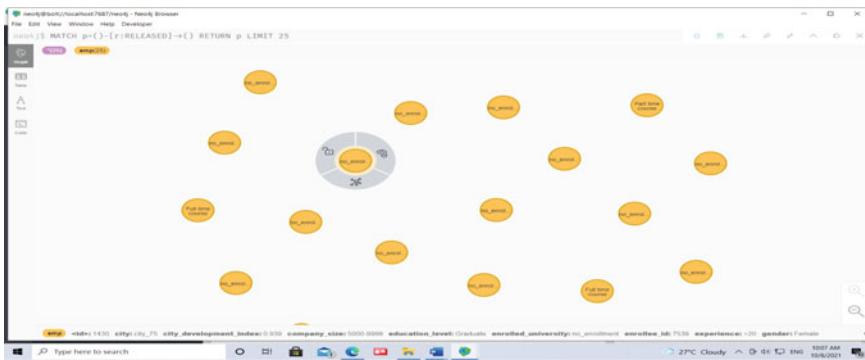
entity	city
"node"	++
"node"	"city"
"node"	"city_41"
"node"	"city_103"
"node"	"city_21"
"node"	"city_13"
"node"	"city_23"
"node"	"city_160"
"node"	"city_173"
"node"	"city_90"

At the bottom, it says 'Started streaming 25 records after 79 ms and completed after 611 ms.'

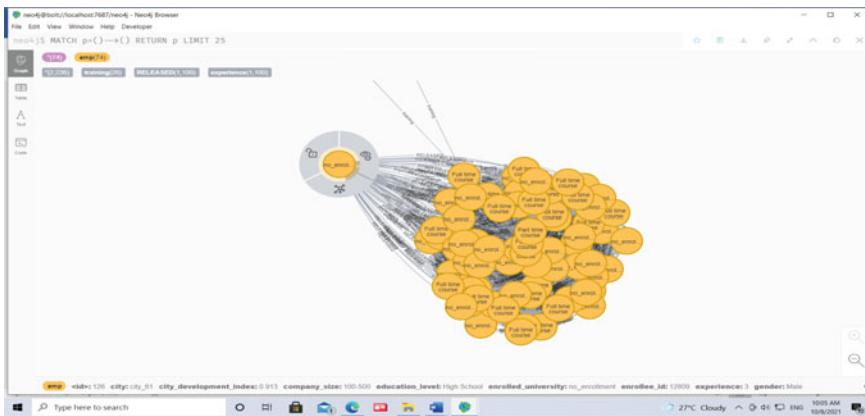
**Fig. 4** Creating nodes with data



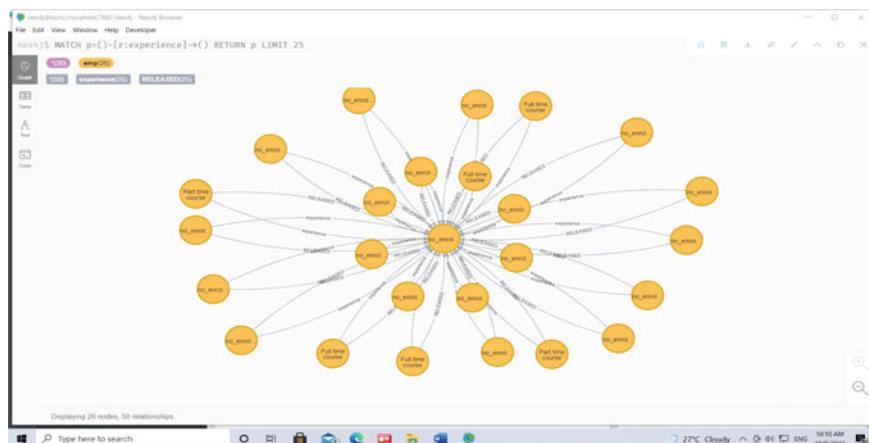
**Fig. 5** Nodes with headers



**Fig. 6** Nodes based on data



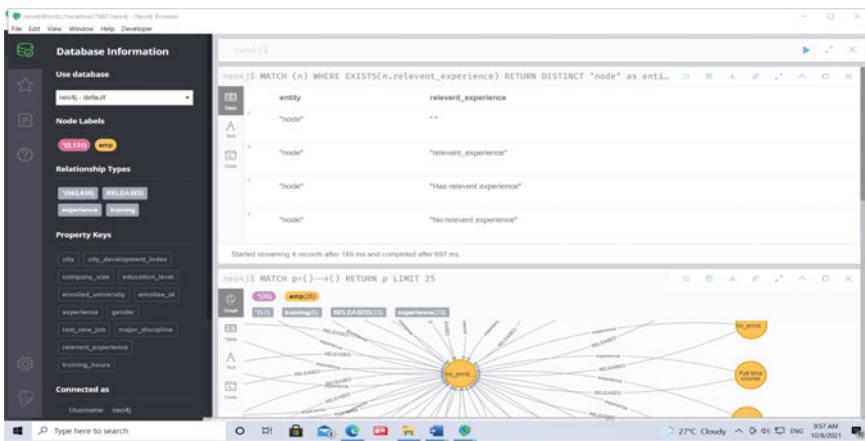
**Fig. 7** Node description and data details



**Fig. 8** Graph formation by applying conditions

To find out various nodes, combinations with people having some similarities are being clustered with the help of cypher language (Fig. 8). Graph-based clustering of prospective candidates according to the skill sets proves to be an effective way to find cluster of people with similar skillsets. All nodes having similar properties on one account are placed together. To create any number of clusters with the dataset, implementor have flexibility of customize queries and gets desired results. Authors fetched clusters having similar experience and it is visible in Fig. 6. Then centrality nodes will be discovered based upon most efficient employees or maximum value among all nodes.

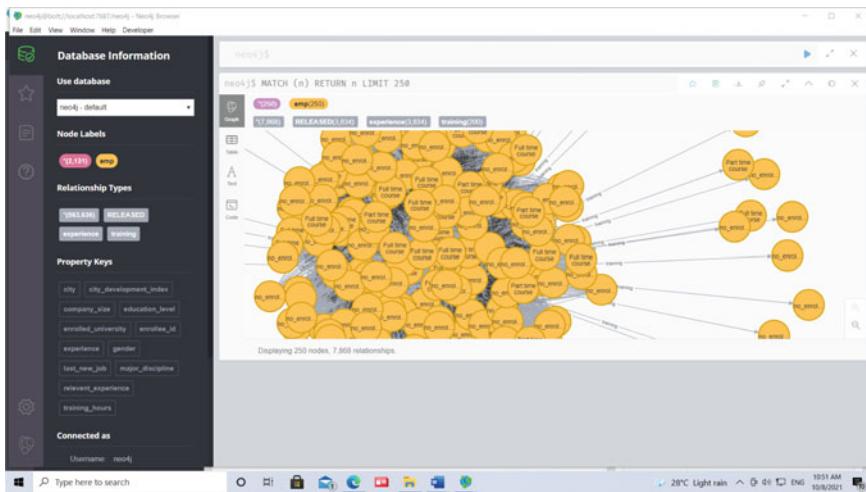
Verification of employees is the most critical and costly process for every organization. Many small- and medium-scale companies are not able to conduct verification process because of the cost. Data from social media and other sources would be helpful in the verification of employees during recruitment and selection process. Authors used graph-based tool called NEO4J to analyse social interaction-based data. These types of data are available on social media account or socio-relational data available from personal and professional background. Educational qualification of the candidate is one of the key parameters during recruitment process. Authors have created cluster of nodes whose one of the values under educational qualification is same as prospective candidate for recruitment. Researchers try to find out ranking of the institute and working status of the other nodes including quality of job. If ranking of educational institute is good and other students of the institute have good placements in better companies, educational standards of the institute are satisfactory for the recruitment. Second factors of analysis is to verify educational qualification of the candidate. If we are able to create cluster of nodes whose educational institute are the same and they are connected for long, then we can be sure that his/her degrees are genuine. So educational verification through graphical nodes can be done for better understanding of candidate. HR professionals have flexibility to see details



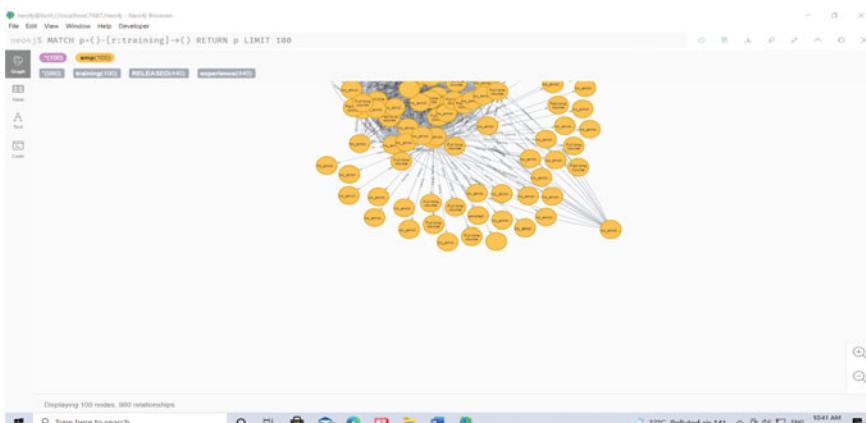
**Fig. 9** Query based on condition

of nodes and set conditions as per requirements of recruitment process and suitable candidate required as in Fig. 9. Verification of behaviour of employee can be done based on employee's friends and their behaviour. Skillsets of friends reflect in the skills of employees. Society where employees have been living plays a vital role in the mindset of the people. So, cultural aspects of the employee will be understood by verifying locality and its social impact. Verification of educational background of the employee can be done by establishing connectivity of nodes among people who had educational degrees from same institute. In spite of many skills, employee may not be able to perform unless they spent quality time during job. Nodes containing details of communication and digital media can throw light on the time spent by employees, on these mediums. If employees spend very less time on digital media, it may imply they are not tech savvy, and if they spend a lot of time, it may be concluded that they are investing much of their time and energy on these platforms. Financial health of the employees is also a critical factor to judge their mental status. If they are more in debt, they may not be able to focus on their job. Nodes of bank loans, credit cards as well as credit score and its connectivity with employees' nodes will present a picture of their financial health.

Clustering based on the data will help in the analysing and getting insights of the most suitable employee in the list of prospective candidates (Figs. 10, 11 and 12). Authors formed clusters based on the various parameters. Connectivity of data through daily communication with mobile phone is one of the key datasets for the graph analysis. Time spent on mobile phone and number of times calls made on particular phone number revealed various psychological behaviour about candidate which plays key role in recruitment or selection of candidate and is one of best-known parameters to check suitability of candidate for the particular post. More time spent on official phone numbers means he/she is busy to perform his/her duties, and more time on personal phone number may mean candidate is less dedicated towards current job. This behaviour might affect the performance of the candidate



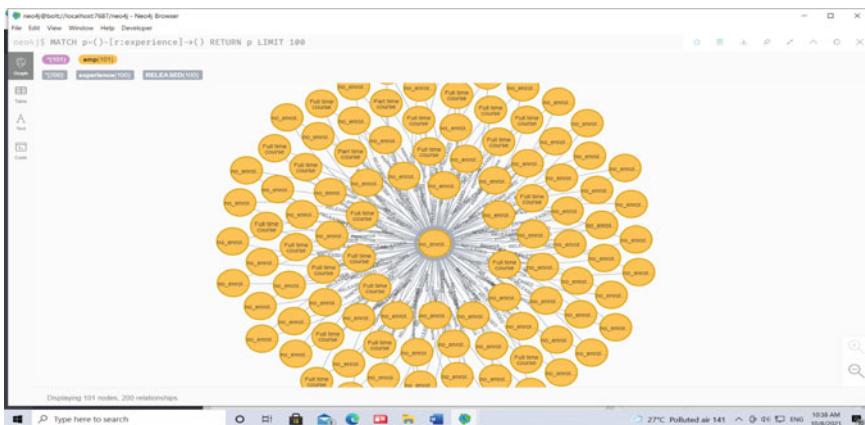
**Fig. 10** Formation of clustering based on certain similarity



**Fig. 11** Formation of clustering based on certain similarity

after selection. This psychological aspect of the candidate is also a key consideration during selection process.

Authors drawn centrality of graph which provides most prominent node in the graph to influence all other nodes. If candidate is in central nodes and other relative nodes are not having much influence, it can mean candidate has leadership position in the society and may have leadership quality. This method will be more viable to understand leadership quality of the person and connectivity of the person in the society. This analysis will also highlight societal behaviour of a person, i.e. whether a person is social or not.



**Fig. 12** Formation of clustering based on certain similarity

Clustering of nodes was formed to find similarity in the nature among people. It is hard fact that people having similar interest are connected with each other. Cluster on Web is being verified by nodes in graph formed over a long-time during communication. Interest of candidate can be known by the fact of known clusters. So, authors created clusters by clubbing nodes having similar properties. Employment verification of the candidate will be done by clustering of nodes having same employment history in the data of nodes. Number of nodes connected with each other in cluster and if candidate's node has connection with other nodes in the cluster verifies its employment status. More number of edges between the nodes increased confidence level of recruiters about employment history of the candidate. Output of these is the data and information based upon relationship among different entities. On the basis of these outputs, authors concluded various insights as discussed in the next section of the paper. Outputs of the graph-based analysis tools are different cluster of people with same attitude and same factors which plays vital role in exploration of their psychology.

## 5 Conclusion and Limitations

After getting different output values like clustering of nodes, etc., analysis and interpretation of those results are based upon situation. It also depends upon the positions that organizations are going to hire for. Importance and number of factors considered for the analysis may vary from case to case. For example, experience verification will be based upon changes in positional coordinates of the person in the duration in which candidates are claiming experience. Similarly, connectivity is the most effective way of establishing relationship among entities and graph-based methods are most effective in it. With the help of graph, HR professionals are able to explore many

hidden patterns of the prospective employees in the organization. Social bonding and behaviour of people speak louder about their sociological and psychological behaviour, which can be helpful in recruitment process of the organization. Each node contains all vital details of the prospective employees and their societal relationship and reveal various key aspects of a person like nature, attitude, etc. Changes in the data like location and movement will talk about continuity of job in the previous organizations. Authors used NoSQL (Neo4J) tool to explore various patterns to know different aspects of the people during recruitment process. NoSQL (Neo4J) tools are effective way to draw model and verify credentials of prospective candidates with their heterogeneous and unstructured nature of data. Model, proposed here, shows verifications as well as validation process of prospective employees during recruitment process. This paper also makes an attempt to explore psychological aspects of personality with graph theoretical concepts. Model and results of the paper can be used during recruitment process to validate and verify various claims made by candidate in the resume and evaluate his/her candidature during the process. This model can be integrated with existing system of HR process to enhance its credibility and efficiency. Although model is meant for the recruitment and selection process based on verification of data entered by candidates, there are some limitations of the model also. Graph-based model is based upon data on social media and data provided by candidates but if candidate has no social media account, access of social media data is denied or there is any other legal hindrance of data, it may restrict the model may not work effectively. There is a possibility of cooking data for long term, and in that case, results of the model may not be reliable and authentic.

## References

1. Career Builder (2021) More than half of employers have found content on social media that caused them NOT to hire a candidate, according to recent Career Builder survey. Accessed July 7 2021
2. Kraut, R., Kiesler, S., Boneva, B., Cummings, J., Helgeson, V., Crawford, A.: Internet paradox revisited. *J. Soc. Issues* **58**(1), 49–74 (2002)
3. Davis, D.C.: MySpace isn't your space; Bepress Legal Series; p. 1943 (2019). Available online: <https://asu-ir.tdl.org/bitstream/handle/2346.1/30045/Fowler%20Thesis.pdf?sequence=1>
4. Archana, L., Nivya, V.G., Thankam, S.M.: Recruitment through social media area: human—resource. *IOSR J. Bus. Manag.*, 37–41 (2012). [www.iosrjournals.org](http://www.iosrjournals.org)
5. Yokoyama, T.T., Okada, M., Panacea, T.T.: Visual exploration system for analyzing trends in annual recruitment using time-varying graphs. *PLoS ONE* **16**(3), e0247587 (2021). <https://doi.org/10.1371/journal.pone.0247587>
6. Narvey, J.: Let's get social. *BC Bus.* **37**(5), 35 (2009)
7. Marsden, P.V., Gorman, E.H.: Social networks, job changes, and recruitment. In: *Sourcebook of Labor Markets*, pp. 467–502. Springer, Boston, MA (2001)
8. Bartakova, G.P., Brtkova, J., Gubiniova, K., Hitka, M.: Actual trends in the recruitment process at small and medium-sized enterprises with the use of social networking. *Econ. Ann.-XXI*, **164**(3–4), 80–84 (2017). <https://doi.org/10.21003/ea.V164-18>
9. Gundecha, P., Liu, H.: Mining social media: a brief introduction. In Mirchandani, P.B. (ed.) *TutORials in Operations Research*, pp. 1–17. INFORMS (2012). <https://doi.org/10.1287/educ.1120.0105>

10. Liu, J., Tang, T., Wang, W., Xu, B., Kong, X., Xia, F.: A survey of scholarly data visualization. *IEEE Access* **6**, 19205–19221 (2018). <https://doi.org/10.1109/ACCESS.2018.2815030>
11. Sharma, R., Sungheetha, A.: An efficient dimension reduction based fusion of CNN and SVM model for detection of abnormal incident in video surveillance. *J. Soft Comput. Paradigm (JSCP)* **3**(02), 55–69 (2021)
12. Ujlayan, A., Sharma, M.: An emergent role of knowledge graph and summarization methodology to simplify recruitment for the Indian IT industry. In: Solanki, A., Sharma, S.K., Tarar, S., Tomar, P., Sharma, S., Nayyar, A. (eds.) *Artificial Intelligence and Sustainable Computing for Smart City. AIS2C2, Communications in Computer and Information Science*, vol. 1434. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-82322-1\\_5](https://doi.org/10.1007/978-3-030-82322-1_5)
13. Shahbaz, U., Beheshti, A., Nobari, S., Qu, Q., Paik, H.Y., Mahdavi, M.: iRecruit: Towards automating the recruitment process. In: Lam, H.P., Mistry, S. (eds) *Service Research and Innovation. ASSRI 2018, Lecture Notes in Business Information Processing*, vol. 367. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-32242-7\\_11](https://doi.org/10.1007/978-3-030-32242-7_11)
14. Böhmová, L., Malinová, L.: Facebook user's privacy in recruitment process. In: *Proceedings of the IDIMT 2013—Information Technology Human Values, Innovation and Economy*, pp. 159–166. Trauner Verlag, Linz (2013)
15. Grasz, J.: Forty-Five Percent of Employers Use Social Networking Sites to Research Job Candidates, Career Builder Survey Finds. Career Builder Press Releases. Available online: <http://www.careerbuilder.co.uk> (2009)
16. Smys, S., Wang, H.: Security enhancement in smart vehicle using blockchain-based architectural framework. *J. Artif. Intell.* **3**(02), 90–100 (2021)
17. Kasper, K.: JobviteInfographic: Watch What You Post on Social Media. Jobvite (2015). Available online: <http://www.jobvite.com/blog/jobvite-infographic-watch-post-social-media/>
18. Berkelaar, B., Buzzanell, P.: Cybervetting, Person–environment fit, and personnel selection: employers' surveillance and sensemaking of job applicants'. *Online Information. J. Appl. Commun. Res.* **42**(2014)
19. Leite, L.T., Palgon, J., Vila, R.: Hiring process by using social networking techniques to verify job seeker information. U.S. Patent Application 11/950,458, filed June 11, 2009
20. You, W., Kosinski, M., Stillwell, D.: Computer-based personality judgments are more accurate than those made by humans. *Proc. Natl. Acad. Sci.* **112**, 1036–1040 (2015)
21. Annisette, L.E., Lafreniere, K.D.: social media, texting, and personality: a test of the shallowing hypothesis. *Pers. Individ. Differ.* **115**, 154–158 (2017). <https://doi.org/10.1016/j.paid.2016.02.043>
22. Manoharan, J.S.: A novel user layer cloud security model based on Chaotic Arnold transformation using fingerprint biometric traits. *J. Innov. Image Process. (JIIP)* **3**(01), 36–51 (2021)
23. Federico, P., Heimerl, F., Koch, S., Miksch, S.: A survey on visual approaches for analyzing scientific literature and patents. *IEEE Trans. Visual Comput. Graphics* **23**, 2179–2198 (2017). <https://doi.org/10.1109/TVCG.2016.2610422>

# Routing Protocols in an Opportunistic Network: A Survey



Sushil Kumar Mishra and Ruchika Gupta

**Abstract** Opportunistic network (OppNets) is a derivative of delay tolerant network (DTN), mainly composed of mobile connected nodes and communication range among nodes are not more than 100 m. These mobile nodes are temporarily connected, and network topology in an opportunistic network is dynamic due to mobility of all nodes. In an opportunistic network, there is no end to end link exists between source and destination node, and message is transmitted from source to destination in hop by hop manner. Opportunistic network works using store, carry, and forward mechanism. A node receiving message from another node, wait for other node to come in its range to be forwarded the message. In this article, surveys of the various routing strategies are discussed along with available tools for simulation in an opportunistic network. The main objective of this article is to deal with current challenges in routing and to provide a future direction for the same.

**Keywords** Opportunistic network (OppNets) · Node · Delay tolerant network (DTN)

## 1 Introduction

Due to the rapid rise of Internet, new applications such as such email and social networking sites have changed the way of accessing and creating information over Internet. In last two decades, electronic gadgets such as smartphone and tablets have played a significant role as communication devices have been expanded rapidly by increasing their storage, processing capacities with wireless communication technology, ranging from Bluetooth to 5G. Unlike a traditional network, opportunistic network does not have a direct route between nodes because of frequent disconnections, which makes it a vulnerable network, which requires a security mechanism to identify legitimate and malicious nodes. Opportunistic networks engage a store,

---

S. K. Mishra (✉) · R. Gupta

Department of Computer Science and Engineering, Chandigarh University, Mohali, Punjab 140413, India

e-mail: [sushiljune@gmail.com](mailto:sushiljune@gmail.com)

carry, and forward mechanism to forward message to all the nodes including intermediate nodes. Intermediate nodes store message and wait for suitable opportunity to forward other node within its range. Every node in an opportunistic network having high degree of mobility, limited battery backup, short range, limited storage, and computational capacities. Due to these qualities, opportunistic network has obtained a tremendous attention for research toward privacy, security, trust, and authentication challenges.

The objective of this research is to give methodological reviews of an opportunistic network. The various routing algorithms are explained along with mobility models for implementation. This study is able to provide a combined data on an existing research and direction for upcoming work.

In this research paper, we come up with a detailed overview about different existing research papers regarding routing algorithms in an opportunistic network.

## 2 Routing Protocols for an Opportunistic Network

Message transmission in opportunistic network is done in hop by hop manner from source node to destination node. The node participating in opportunistic network is responsible to forward message closer to destination node. If intermediate node does not exist, then message is stored by source node and wait for opportunity for transmission. In an opportunistic network, participating nodes have no information about other nodes; they only can communicate if they are in communication range. The following characteristics play a major role in an opportunistic network.

- Store, carry and forward mechanism
- Node collaboration
- Node mobility.

In addition, only one copy of the message is transmitted throughout routing. The message is kept by one node, and other nodes wait for the message to be transferred closer to the destination. In multicasting routing, numerous messages are sent to a many destinations in an opportunistic network.

A network is often divided into many partitions called sections in an OppNets. Conventional applications are ineffective in this context because they expect that the possibility for end to end connection between source node and the destination node exist. By operating messages in a store, carry, and forward mode, an OppNets allow devices in different areas to connect. The store, carry forward message swapping method is implemented by the intermediate nodes by adding bundle layer as shown in Fig. 1. Each node in an OppNets is a separate entity with bundle layer that can function as a router or gateway.

There are multiple overhead while performing multicast routing. These overhead is known as an action or activities to meet the aims. The following metrics are used to calculate overhead during routing.

**Fig. 1** Protocol stack for an OppNets routing

Application Layer	
Bundle Layer	
Transport Layer 1	Transport Layer 2
Network Layer 1	Network Layer 2
Link Layer 1	Link Layer 2
Physical Layer 1	Physical Layer 2

- **Extra Data:** During message transmission in an opportunistic network, node receives multiple duplicate data or undesired data. It causes a network overhead.
- **Buffer overflow and energy consumption:** The duplicate or irrelevant data stored and forwarded by node and wait for suitable node to transfer. It is a network overhead. The unnecessary forwarding of the message also causes an energy consumption.

The routing protocols are categorized into the various following types.

## 2.1 Direct Protocol

Direct protocols represent a concept of message delivery. In which, only one message is transmitted in the network. The protocols proposed described in Table 1.

## 2.2 Flooding-Based Protocols

In this approach, broadcast the message in an opportunistic network to attain the highest degree of delivery ratio. The message is forwarded to all nodes that come in the communication range of nodes carrying the message. This approach having a high network overhead. The protocols proposed described in Table 1.

### ***2.3 History-Based Prediction Protocols***

The decision to send the message to specific nodes is made based on previous history in this technique. It calculates the possibility of a node delivering the message successfully. The protocols proposed described in Table 1.

### ***2.4 Context-Based Protocol***

History-based prediction protocol is further improved for better message delivery ratio by applying context information. Context information is like node location, node speed, and other local information to enhance the delivery ratio. The protocols proposed described in Table 1.

### ***2.5 Social Awareness-Based Protocols***

This routing strategy is based on human behaviors like social interest, contact, and popularity of the node. This information plays a vital role in predicting future behavior of the node. The protocols proposed described in Table 1.

### ***2.6 Timing Protocols***

Waiting time, message transmission, and time to live (TTL) are all taken into account for each message while it is being transmitted in this method. The protocols proposed described in Table 1 (Fig. 2).

## **3 Literature Survey on Routing Protocols in an Opportunistic Network**

Routing protocol in an opportunistic network is based on store, carry forward mechanism, and there is no permanent path existing between source and destination node. In this section, a detailed description about existing routing protocols has been highlighted with its strength and weakness.

**Table 1** Literature survey on routing protocols

References year	Approach	Description	Strengths and weaknesses
[1] (2000)	Based on flooding approaches, the author presented an epidemic routing scheme	For node buffer management, epidemic routing employs the FCFS method. If a node's buffer is full, it won't be able to receive fresh messages from other nodes	The suggested routing scheme provides 100% message delivery with higher congestion and network overhead
[2] (2004)	The author proposed a spray and wait routing is based on flooding technique	The spray phase and the wait phase have been introduced in the current study, in which the node sprays the message first and then waits for it to reach the destination node	The proposed routing scheme phase made significant success in message delivery, but it requires unlimited storage and bandwidth
[3] (2013)	The author proposed an agent-based multicast opportunistic routing protocol	By monitoring successful transmission, the agent in the current study decreases retransmission in the network	The proposed routing scheme improves throughput and require limited bandwidth
[4] (2013)	The author proposed a history-based predictions routing	History-based prediction routing takes decision for node selection based on its previous characteristics	The proposed routing protocol is able to detect reliable node for message forwarding and require limited storage for processing
[5] (2006)	The author proposed a routing protocol called H-EC	In this routing protocol, message is encoded into small message at sender side and send it to next hop	The suggested routing is helpful in successful delivery of the message
[6] (2004)	The author proposed a routing scheme based on time factor	In this routing protocol, path selection is purely based on shortest time taken by node for message transmission	In this technique, scalability, storage, and bandwidth is low
[7] (2008)	The author proposed a routing scheme for message prioritization	In this routing scheme, higher and lower priority is assigned to a message based on delegation number	The main drawback of this scheme is low scalability and storage

(continued)

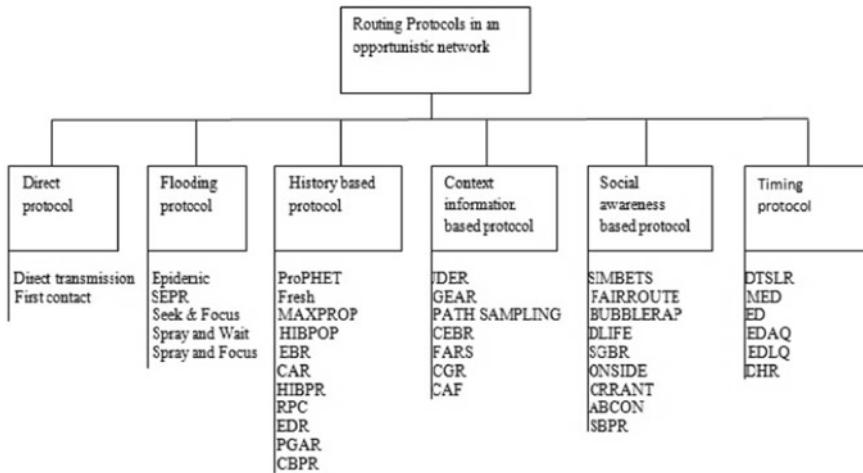
**Table 1** (continued)

References year	Approach	Description	Strengths and weaknesses
[8] (2006)	The author proposed a routing protocol for evaluation of best available path	Each node has a vector assigned to it. When two nodes meet each other. They share a vector to find the shortest way to their goal	The suggested scheme is able to provide shortest path in the network with limited bandwidth
[9] (2008)	The author described about a delegation forwarding routing scheme	The node has a quality metric. If a node meets another node with higher quality metrics than other nodes. It transmits message to the concern node	The suggested scheme has higher network overhead
[10] (2009)	The author proposed a routing scheme to control the transmission of message	This scheme utilizes the concept of spray and wait scheme, in which it uses previous history to transmit the message having higher a higher encounter rate in the network	This scheme has a counter measures against black hole attack
[11] (2004)	The author suggested a direct transmission scheme for reducing number of hop between the encountered nodes	In this scheme, source node directly sends a message to destination nodes. If source node is not in the range of destination node, then it tries to come close to destination node to transmit the message	The suggested scheme has minimum network overhead
[12] (2003)	The author proposed an EASE algorithm for selection of nodes in the network	According to this study, node routing is based on the environmental location of the destination node, while each node maintains a records for the encountered node	The efficiency and efficacy of the proposed scheme is above average, but network overhead is high
[13] (2008)	The author proposed a hybrid approach	This method combines a single-copy flooding protocol and a utility function	Delay in the network is observed due to single transmission but having low network overhead

(continued)

**Table 1** (continued)

References year	Approach	Description	Strengths and weaknesses
[14] (2013)	The author suggested a prediction-based theory	Identification of future encounters of a node based on mobility pattern and social attributes	Better result is expected in case of regular encounters
[15] (2016)	The author proposed a new protocol based on quality metrics	In this approach, a cost-effective routing is done based on quality metrics	It is applicable to only small-scale network
[16] (2018)	Proposed a rule-based expert system for routing	Prediction for routing and next node is done based on past contact history	No restriction for energy consumption. It is best suited to small-scale network
[17] (2014)	Proposed an approach Jaccard distance and encountered ration	The method is based on two theories: first, a new node is selected from equivalent type network based on prediction (based on history); next, a new node is selected from a heterogeneous type of network based on history	Low latency is observed as comparing it with other approach
[18] (2008)	Proposed an approach for opportunistic network routing with window-aware adaptive replication	This strategy is based on resource-conserving replication routing. The decision to route is dependent on parameters such as transmission speed, direction, and so on	In this approach, low latency is observed
[19] (2019)	The author proposed a scheme for interest-based epidemic routing	Interest-based epidemic routing depends upon three factors contents, interest and social-based network	Message delivery ratio is better as compared to spay and wait but poor as compared to ProPHET
[20] (2020)	The author proposed an energy efficient routing scheme	The routing scheme is based on idle extent, network energy efficiency ratio, and efficiency factor	In this approach, high latency is observed but restriction for energy consumption
[21] (2021)	The author worked on ACO multi-cast routing	The routing strategy is based on multi path selection with carry forward mechanism	In this strategy, high latency is observed



**Fig. 2** Routing protocol of opportunistic network

## 4 Comparative Analysis

Performance is examined by applying the various ways on one simulator in order to deliver the effectiveness of routing protocol. According to the findings of the study, most opportunistic network routing protocols are compared to epidemic routing, ProPHET, and spray and wait routing. For performance analysis, the following parameters are taken into account for various routing protocols (Tables 2 and 3).

Message delivery probability is being represented for the various routing protocol with respect to direct contact routing protocol in Table 3 and graphically in Fig. 3.

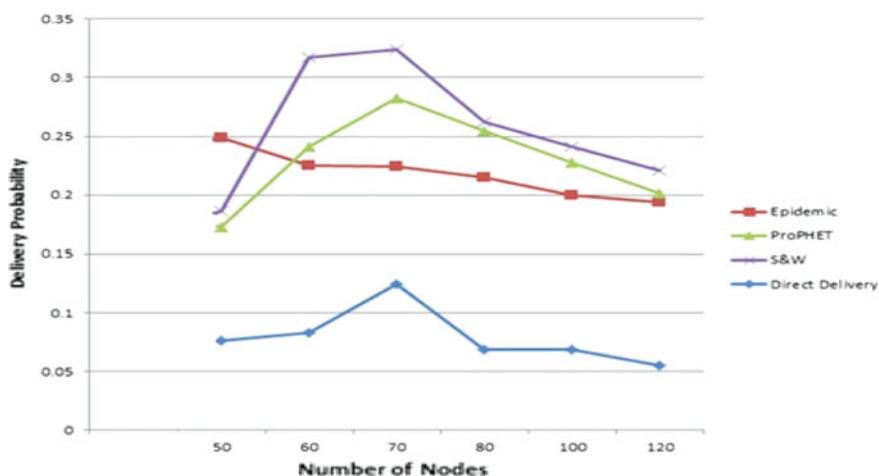
When compared to epidemic routing, ProPHET, and spray and wait routing protocols, performance analysis demonstrates that direct contact routing has the lowest message delivery probability. When compared to ProPHET and spray and wait routing systems, epidemic routing has a higher hop count due to uncontrolled flooding routing.

**Table 2** Fixed parameters

Parameter	Range in one simulator
Message size	250–500 KB
TTL (message)	60 min
Transmission speed	5 Mbps
Transmission range	500 m
Total number of nodes	50–120

**Table 3** Message delivery probability versus number of present nodes

Total no of nodes	ProPHET	Epidemic	Spray and wait	Direct contact
50	0.1723	0.2400	0.1789	0.0713
60	0.2313	0.2300	0.3225	0.0823
70	0.2915	0.2256	0.3289	0.1325
80	0.2616	0.2213	0.2615	0.0693
100	0.2300	0.2030	0.2456	0.0699
120	0.1980	0.1930	0.2234	0.0595

**Fig. 3** Graph representing message versus number of present nodes

## 5 Conclusion

Wirelessly connected nodes communicate with each other opportunistically using a store, carry and forward method in an opportunistic network. In order to build connectivity between mobile nodes, opportunistic networks also leverage social relationships, human behavior, or characteristics. Although there are other routing protocols for opportunistic networks, spray and wait, epidemic routing, and ProPHET routing are considered standard routing strategies. These routing algorithms were implemented with direct contact, but not suitable for OppNets due to the store, carry, and forward mechanism, as seen in the results.

## References

- Vahdat, Becker, D.: Epidemic Routing for Partially Connected Ad Hoc Networks. Tech. Rep., CS-200006, Duke University (2000)
- Spyropoulos, T., Psounis, K., Raghavendra, C.: Multiple-Copy Routing in Intermittently Connected Mobile Networks. Tech. Rep., CENG-2004-12, USC (2004)
- Krishna, M., Barman, D.: Agent-based multicast opportunistic routing protocol for wireless networks. In: HP-MOSys'13 Proceedings of the 2nd ACM Workshop High Performance Mobile Opportunistic Syst., Barcelona, Nov 2013, pp. 1–8
- Dhurandher, S., Sharma, D., Woungang, I., Bhati, S.: HBPR: History based prediction for routing in infrastructure-less opportunistic networks. In: 2013 IEEE 27th International Conference on Advanced Information Networking and Applications (AINA), Barcelona, March 2013, pp. 931–936
- Chen, L., Yu, C., Sun, T., Chen, Y., Chu, H.H.: A hybrid routing approach for opportunistic networks. In: Proceedings of the 2006 SIGCOMM Workshop on Challenged Networks, Pisa, Sept 2006, pp. 213–220
- Jain, S., Fall, K., Patra, R.: Routing in a delay tolerant network. *SIGCOMM Comput. Commun. Rev.* **34**(4), 145–158 (2004)
- Erramilli, V., Crovella, M.: Forwarding in opportunistic networks with resource constraints. In: Proceedings of the Third ACM Workshop on Challenged Networks, San Francisco, CA, Sept 2008, pp. 41–48
- Burgess, J., Gallagher, B., Jensen, D., Levine, B.: Maxprop: routing for vehicle-based disruption-tolerant networks. In: Proceedings of the INFOCOM 2006. Twenty Fifth IEEE International Conference on Computer Communications, Barcelona, April 2006, pp. 1–11
- Erramilli, V., Crovella, M., Chaintreau, A., Christophe, Delegation forwarding. In: Proc. Ninth ACM International Symposium on Mobile Ad Hoc Networking and Comput., Hong Kong, May 2008, pp. 251–260
- Nelson, S., Bakht, M., Kravets, R.: Encounter based routing in DTNs. In: Proceedings of the INFOCOM 2009, New York, NY, April 2009, pp. 846–854
- Spyropoulos, T., Psounis, K., Raghavendra, C.: Singlecopy routing in intermittently connected mobile networks. In: First Annual IEEE Communications Society Conference on Sensor and Ad Hoc Communications and Networks (SECON), Santa Clara, CA, Oct. 2004, pp. 235–244
- Grossglauser, M., Vetterli, M.: Locating nodes with EASE: Last encounter routing in ad hoc networks through mobility diffusion. In: INFOCOM 2003, Twenty-Second Annual Joint Conference of the IEEE Computer and Communications Societies, San Francisco, CA, vol. 3, Mar. 2003, pp. 1954–1964
- Spyropoulos, T., Psounis, K., Raghavendra, C.S.: Efficient routing in intermittently connected mobile networks: the single-copy case. *IEEE/ACM Trans. Netw.* **16**(1), 63–76 (2008). <https://doi.org/10.1109/tnet.2007.897962>
- Ciobanu, R.I., Dobre, C., Cristea, V.: SPRINT: social prediction-based opportunistic routing. In: 2013 IEEE 14th International Symposium on “A World of Wireless, Mobile and Multimedia Networks” (WoWMoM). 2013. <https://doi.org/10.1109/wowmom.2013.6583442>
- Wu, J., Wang, J., Liu, L., Tanha, M., Pan, J.: A data forwarding scheme with reachable probability centrality in DTNs. In: 2016 IEEE Wireless Communications and Networking Conference. 2016. <https://doi.org/10.1109/wcnc.2016.7565160>
- Sajid, A., Hussain, K.: Rule based (forward chaining/data driven) expert system for node level congestion handling in opportunistic network. *Mob. Netw. Appl.* **23**(3), 446–455 (2018)
- Ciobanu, R.I., Reina, D.G., Dobre, C., Toral, S.L., Johnson, P.: JDER: a history-based forwarding scheme for delay tolerant networks using Jaccard distance and encountered ration. *J. Netw. Comput. Appl.* **40**, 279–291 (2014). <https://doi.org/10.1016/j.jnca.2013.09.012>
- Sandulescu, G., Nadim-Tehrani, S.: Opportunistic DTN routing with window-aware adaptive replication. In: Proceedings of the 4th Asian Conference on Internet Engineering, AINTEC’08. 2008. <https://doi.org/10.1145/1503370.1503397>

19. Ayu, V., Soelistijanto, B., Sijabat, J.: Interest-based epidemic routing in opportunistic mobile networks. In: 2019 7th International Conference on Information and Communication Technology (ICoICT). 2019. <https://doi.org/10.1109/icict.2019.8835355>
20. Tang, L., Lu, Z., Fan, B.: Energy efficient and reliable routing algorithm for wireless sensors networks. *Appl. Sci.* **10**, 1885 (2020). <https://doi.org/10.3390/app10051885>
21. Dhaya, R., Kanthavel R.: Bus-based VANET using ACO multipath routing algorithm. *J. Trends Comput. Sci. Smart Technol. (TCSST)* (2021). <https://doi.org/10.36548/jcsst.2021.1.004>

# K-Splits: Improved K-Means Clustering Algorithm to Automatically Detect the Number of Clusters



Seyed Omid Mohammadi, Ahmad Kalhor, and Hossein Bodaghi

**Abstract** This paper introduces  $k$ -splits, an improved hierarchical algorithm based on  $k$ -means to cluster data without prior knowledge of the number of clusters. K-splits starts from a small number of clusters and uses the most significant data distribution axis to split these clusters incrementally into better fits if needed. Accuracy and speed are two main advantages of the proposed method. This research experiments on six synthetic benchmark datasets plus two real-world datasets MNIST and Fashion-MNIST, to show that the proposed algorithm has excellent accuracy in automatically finding the correct number of clusters under different conditions. The experimental analysis also indicates that  $k$ -splits is faster than similar methods and can even be faster than the standard  $k$ -means in lower dimensions. Furthermore, this article delves deeper into the effects of algorithm hyperparameters and dataset parameters on  $k$ -splits. Finally, it suggests using  $k$ -splits to uncover the exact position of centroids and then input them as initial points to the  $k$ -means algorithm to fine-tune the results.

**Keywords** Data clustering · Initialization · K-means algorithm · Number of clusters

---

S. O. Mohammadi (✉) · A. Kalhor · H. Bodaghi

School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran

e-mail: [S.OmidMohammadi@alumni.ut.ac.ir](mailto:S.OmidMohammadi@alumni.ut.ac.ir)

A. Kalhor

e-mail: [AKalhor@ut.ac.ir](mailto:AKalhor@ut.ac.ir)

H. Bodaghi

e-mail: [Hossein.Bodaghi@ut.ac.ir](mailto:Hossein.Bodaghi@ut.ac.ir)

## 1 Introduction

Recent advances in scientific data collection technologies and the ever-growing volume of complex and diverse data make it harder for us to extract useful information. Moreover, most of this data is unlabeled, as finding suitable labels can be costly and time-consuming. Here is where unsupervised clustering methods come to assist. Clustering is the act of grouping items together which have similar characteristics and features. This way, each group (called a cluster) consists of similar items dissimilar to the other clusters' items.

K-means is a popular unsupervised clustering algorithm widely used due to its simple implementation and reasonably good results. However, this algorithm has its downsides, including high execution time and the dependency of final results on initial configurations. We also need to know the exact number of clusters ( $k$ ) before proceeding with  $k$ -means, which can be a tricky task, especially for high-dimensional data. Setting a large  $k$  can cause many dead clusters (clusters with very few items), whereas a small  $k$  forces the items into insufficient clusters, leading to poor results. Multiple researchers tried to address this problem and introduced methods to estimate the number of clusters, including the Elbow method [1], the Silhouette method [2], and numerous variations of Subtractive clustering [3]. However, some of them require an exhaustive search or have a high computational cost.

This paper proposes a hierarchical algorithm wrapped around  $k$ -means to systematically and automatically determine the best number of clusters. The novelty of the proposed algorithm is that it uses the most significant data distribution axis to split the clusters incrementally into better fits if needed, causing a significant boost to the accuracy and speed. Unlike previous methods, this algorithm also tries to limit the over-/under-splitting of the clusters. First, Sect. 2 presents a short review of the  $k$ -means algorithm and discusses its limitations and related works. Then, Sect. 3 explains the proposed algorithm, and finally, experimental results are presented in Sect. 4, followed by a conclusion.

## 2 Related Works: K-Means Algorithm and Limitations

K-means is one of the widely used algorithms in data mining. Scalability and simplicity of implementation make the  $k$ -means algorithm a perfect candidate for many practical applications ranging from optimization, signal processing, and big data analysis to face detection and emotion recognition [4–6].

First published in 1956 [7],  $k$ -means soon gained popularity using Lloyd's algorithm [8]. The algorithm starts by breaking the data into  $k$  clusters. As this initialization heavily impacts the final results, numerous researchers such as [9] and [10] suggested different initialization schemes during the years. A straightforward method

is Random initialization, which sets  $k$  random data points as cluster centers (called centroids). Next, the distances (usually Euclidean distance) between centroids and all data points are calculated, and each data point is assigned to the nearest cluster. The next step calculates the mean of all items in each cluster and sets new centroids. All data points are then checked against new centroids, and any relocation of data points to another cluster is done if needed. By repeating these steps, eventually, these once random centroids move step by step until we meet the convergence criterion, which is when all centroids are stable, and no change is needed. This process is shown in Algorithm 1. However, the  $k$ -means algorithm has several significant limits which  $k$ -splits tries to overcome.

The main limitations of  $k$ -means are as follows:

1. The number of clusters ( $k$ ): The standard  $k$ -means algorithm needs a predefined  $k$  to start, obtaining the exact value of which can be challenging in complex and multidimensional data. Multiple methods exist to estimate this setting, but the computational overhead is too much for complex and large data sets [11]. Wrapper methods like x-means and g-means also try to find the best value of  $k$  by splitting or joining clusters. X-means uses the Bayesian Information Criterion penalty for model complexity [12], and g-means uses a statistical test to find Gaussian centers [13]. However, both these methods lack accuracy.
2. Centroids initialization: Solving this issue is very important because initialization seriously affects the final results. Reference [14] proposes a method of finding these centroids, which leads to better accuracy.
3. Time consumption: Calculating the distance between all data points and centroids repeatedly through iterations makes it a time-demanding process, especially for large amounts of data. Researchers usually try to tackle this problem by implementing  $k$ -means on parallel platforms [15, 16] or optimizing the algorithm itself [17, 18]. Even small reductions in the run-time can highly increase temporal efficiency when dealing with large amounts of data.

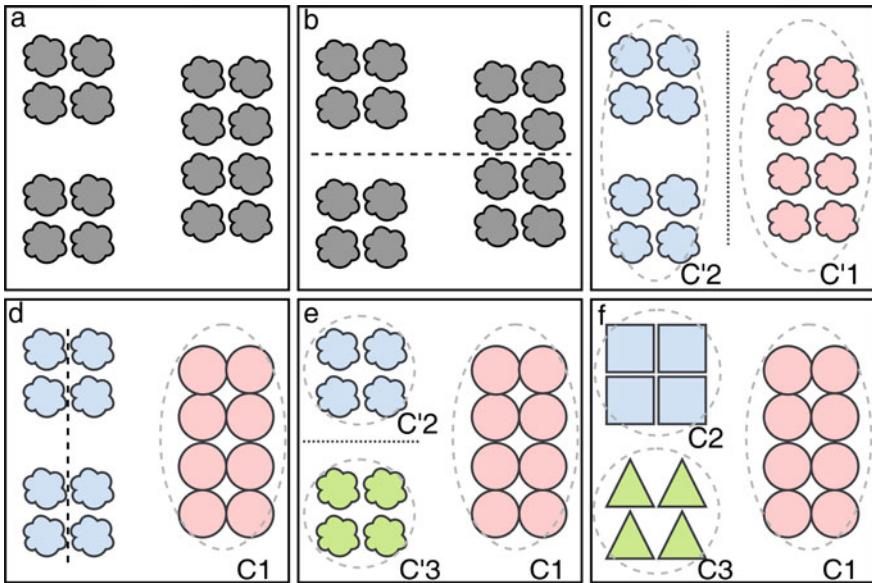
<b>Algorithm 1:</b> The Standard K-Means Algorithm ( $X, k$ )	<b>Algorithm 2:</b> The proposed K-splits Algorithm ( $X, \beta$ )
<p>1. Choose <math>k</math> data points as cluster centroids.</p> <p>2. Calculate the distance of all data points from centroids and assign each data point to the nearest cluster.</p> <p>3. Calculate the mean value of items in each cluster and recalculate new centroids of clusters.</p> <p>4. If none of the centroids changed, proceed to step 5; otherwise, go back to step 2.</p> <p>5. Output the results.</p> <p>End of the algorithm.</p>	<p>1. Start with <math>k=1</math>.</p> <p>2. Calculate <math>I^C</math> for each cluster using (10) and <math>J_k</math> using (9), then find the worst cluster (<math>C^w</math>) using (12).</p> <p>3. Split the worst cluster into two clusters by assigning items to each cluster based on (5), then run Algorithm 1 (standard k-means) for these two clusters to obtain centroids (<math>c_{kw_1}, c_{kw_2}</math>).</p> <p>4. If this is the first iteration; Use (8) to calculate the reference distance <math>d_{base}</math> between these two centroids.</p> <p>5. Else:</p> <p>6. Calculate centroid distances between every two clusters using (13) and set the minimum distance as <math>d</math>.</p> <p>7. If condition (14) is not satisfied:</p> <p>8. Update <math>k \leftarrow k + 1</math> and go to step 2.</p> <p>9. Else:</p> <p>10. Discard centroids (<math>c_{kw_1}, c_{kw_2}</math>) and output the results of the iteration, which satisfies condition (15).</p> <p>End of the algorithm.</p>

### 3 Proposed Algorithm

This section proposes a new method of  $k$ -means based clustering algorithm called “ $k$ -splits” to systematically and automatically find the correct number of clusters ( $k$ ) along the way. This hierarchical algorithm starts from a small number of clusters and splits them into more clusters if needed. The main advantage of  $k$ -splits, making it superior to others mentioned before, is that other methods split the clusters then run multiple tests to understand if it was the right decision, but we do the opposite. As a result, the algorithm intelligently chooses the right cluster to break (called the worst cluster) and only focuses on one cluster at a time, which saves much computational effort.

The main idea comes from the tendency of the  $k$ -means algorithm to find clusters of spherical shapes. Hopefully, the data can be separated into more relevant clusters by finding the axes of variance and data density in different areas. A perfect example is a cluster in a dumbbell shape. One can easily split this cluster by a hyperplane perpendicular to its most significant variance axis orientation (Fig. 1).

The complete algorithm of the proposed  $k$ -splits is shown in Algorithm 2. The algorithm starts with one cluster (or more, based on prior knowledge), assuming all



**Fig. 1** A simplified example of  $k$ -splits: **a** Raw data, **b** find the most significant variance axis, **c** assign sides of the perpendicular hyperplane to clusters, **d** verify using  $k$ -means, choose the worst cluster ( $C'2$ ), and prepare for the next split, **d-f** repeat the procedure

the data belong to one cluster, then it splits this massive cluster into two smaller clusters. It does this separation by finding the axis with the most significant variance. This axis has the same orientation as the eigenvector relevant to the data points' covariance matrix's maximum eigenvalue.

For data  $X^C$  in cluster  $C$  with centroid  $c$ , the covariance matrix  $\Sigma^C$  is calculated from

$$\Sigma^C = \frac{1}{Q_C} \tilde{X}^{CT} \tilde{X}^C \quad (1)$$

where  $Q_C$  is the number of items in cluster  $C$ , and

$$\tilde{X}^C = (X^C - c), \quad c = \bar{X}^C \quad (2)$$

with  $c$  being the centroid,  $\bar{X}^C$  being the mean value of all items in each cluster  $C$  and also

$$X^C = \begin{bmatrix} x_{C1} \\ \vdots \\ x_{CQ_C} \end{bmatrix} \quad (3)$$

Then eigenvalues ( $\lambda$ ) and eigenvectors ( $v$ ) are extracted from this matrix. The method of accomplishing this goal can significantly impact the final computational complexity and run-time of this algorithm. However, for simplicity, one can use the SVD (Singular Value Decomposition) method [19] to obtain these values and then sort them in descending order.

$$\Sigma^C \xrightarrow{SVD} [\lambda_1^C \dots \lambda_n^C], [v_1^C \dots v_n^C] \quad \lambda_1^C \gg \lambda_2^C \gg \dots \gg \lambda_n^C \quad (4)$$

After that, it is time to find the hyperplane perpendicular to the most significant variance axis and assign data on different sides of it to two separate clusters:

$$\forall x^q : \begin{cases} \text{if } \tilde{x}^q V^1 \geq 0 \rightarrow \tilde{x}^q \in \text{First cluster} \\ \text{if } \tilde{x}^q V^1 < 0 \rightarrow \tilde{x}^q \in \text{Second cluster} \end{cases} \quad (5)$$

where:

$$V^1 = \begin{bmatrix} v_1^1 \\ \vdots \\ v_n^1 \end{bmatrix}_{n \times 1} \quad (6)$$

and

$$\text{For } q = 1, \dots, Q_C : \quad x^q - c = \tilde{x}^q \quad (7)$$

This step assigns data points to these two clusters. An estimation of two initial centroids ( $\tilde{c}_1, \tilde{c}_2$ ) is obtained by averaging data points in each cluster. Now to fine-tune the centers, the  $k$ -means algorithm is applied as in Algorithm 1 on these two clusters until convergence to find the final centroids ( $c_1, c_2$ ). Then the distance between these two centroids is calculated using (8). Any distance can be used here, but this article proceeds using  $l_2$ -norm.

$$d_{\text{base}} = \| c_1 - c_2 \|_2 \quad (8)$$

The  $d_{\text{base}}$  distance is the longest distance between two centroids and will be used as a stop condition in future steps to determine when to end the process.

By now, the algorithm has formed only two clusters. From this point forward, in each iteration ( $k$ ), the algorithm finds the worst cluster  $C^w$ , the best candidate for separation, then repeats all these mentioned steps and splits these worst clusters into two smaller ones. This process adds one cluster ( $k \leftarrow k + 1$ ) with each iteration.

In each iteration, the ratio of total items in each cluster to its covariance matrix's greatest eigenvalue is also checked, and the average value  $J_k$  from (9) is saved for later use.  $J_k$  shows the density of clusters in each iteration and makes sure the algorithm is not over-splitting.

$$\text{For } C = 1, \dots, k : \quad J^C = \frac{Q_C}{\lambda_1^C}, \quad J_k = \frac{\sum_{C=1}^k J^C}{k} \quad (9)$$

Testing the clusters, finding the worst cluster, and splitting it into two, increases the algorithm's time efficiency. Although computationally complex, running this test bypasses multiple unnecessary and exhaustive iterations of  $k$ -means, which leads to better run-time, especially on large data sets.

This work introduces  $I^C$ , a criterion to help us determine the worst cluster (the cluster which needs further splitting). Choosing the worst cluster is an essential part of the process; that is why it takes finesse. The algorithm checks multiple elements to guarantee that we are splitting the right cluster. One of them is  $\lambda_1^C$ , giving us an idea of diversity and data distribution in that cluster. Another insight can come from  $thr$ , a threshold made of a combination of  $Q$ ,  $Q_C$ , and  $k$  as in (11), which shows the density of data in that cluster with a hint of the whole clustering situation the algorithm is in now, taking into account the within distance of each cluster. These information sources are combined to calculate  $I^C$ . However, one last detail remains to complete this process.

While splitting clusters, especially with unbalanced data, the algorithm might encounter massive clusters with slight variations that do not need separation. Thus, given the unbalance of data, the algorithm might get stuck in splitting these clusters (called Black Holes). These Black Holes take much effort from the algorithm causing it to overlook other significant clusters and waste time and computational effort. This issue is solved by applying tanh to create a soft saturation as in (10).

$$I^C = \tanh\left(\frac{Q_C}{thr}\right)\lambda_1^C \quad (10)$$

where:

$$thr = \frac{Q}{k}, \quad \sum_{C=1}^k Q_C = Q \quad (11)$$

Calculating  $I^C$  from (10), helps determine the worst cluster. The cluster with the highest value of  $I^C$  according to (12) is the candidate for further splitting.

$$C^w = \arg \max I^C \quad (12)$$

After pinpointing the worst cluster, the algorithm repeats all previous splitting steps precisely, with one difference. This time after finding the centroids of sub-clusters ( $c_{kw_1}, c_{kw_2}$ ), we calculate the distance of all centroids two by two and set the minimum distance as  $d$  according to (13).

The final step in the algorithm is checking the stop condition (14). If the ratio of minimum cluster distance to maximum cluster distance is smaller than a threshold  $\beta$ , then no further separation is needed. This condition considers the intra-class distance

of clusters, and  $\beta$  can be set according to our prior knowledge of the data. Smaller  $\beta$  values lead to more clusters, and larger values do the opposite. For very dense data, larger values of  $\beta$  are suggested, and for more sparse data, vice versa.

For  $i = 1, \dots, k + 1$ ,  $j = 1, \dots, k$ ,  $i \neq j$ :

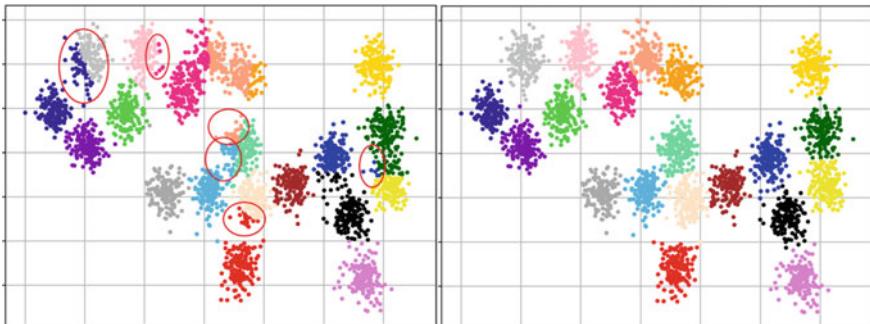
$$d_{ij} = \|c_i - c_j\|_2, \quad d = \min_{\substack{i = 1, \dots, k + 1 \\ j = 1, \dots, k \\ i \neq j}} d_{ij} \quad (13)$$

$$\frac{d}{d_{\text{base}}} \leq \beta, \quad 0 < \beta < 1 \quad (14)$$

After reaching the stop condition, the optional step (15) can be used to choose the best iteration and discard later results to reduce redundant clusters and sensitivity to hyperparameter  $\beta$ , which controls the separation condition. This step makes the algorithm easier to use. However, experiments show that it is best to skip this step to get better results for very dense or highly overlapped data ( $> 50\%$ ). So it is always good to run some tests on the density of data before starting the algorithm.

$$\text{iter} = \arg \max J_k \quad (15)$$

K-splits efficiently and automatically finds the number of clusters and their borders. However, the accuracy of assigning items to clusters might not be the highest because the primary focus is on finding centroids as fast as possible. This problem is solved via an optional fine-tuning step. Thus, it is advisable to use final centroids from  $k$ -splits and then run the conventional  $k$ -means on the data using these known points as initialization. This simple step can fine-tune  $k$ -splits and highly improve the final results, as shown in Table 2 and Fig. 2.



**Fig. 2** Clusters before and after fine-tuning results of  $k$ -splits on the A1 dataset. The first graph highlights some incorrectly labeled data points and fine-tuning corrects them, as shown in the second graph

## 4 Experiments

### 4.1 Accuracy and Run-Time Comparison

For experimental analysis, this essay chooses synthetic benchmark datasets carefully to show different challenging aspects of the data. Detailed information about the datasets can be found in [20]. Each dataset has unique properties, including varying cluster size (A), dimension (Dim), overlap (S), structure (Birch), balance (Unbalance), and a combination of dimension and overlap (G). Real-world datasets MNIST [21] and Fashion-MNIST [22] are also used to demonstrate the method's applicability further. Table 1 presents specifications for each dataset.

Although many advanced methods exist to find the correct number of clusters, the focus here is only on  $k$ -means based algorithms. Thus, this work compares the proposed algorithm and its fine-tuned version with  $g$ -means and  $x$ -means, which are most similar and comparable methods to  $k$ -means.  $K$ -means (with the correct number of  $k$  as input) is used as the baseline. A tricky part of the standard  $k$ -means is knowing the exact number of clusters. Here, much better results are expected as the correct number of  $k$  (which is not available in real-world problems) is used as input to  $k$ -means.  $K$ -means is highly affected by initialization, and wrapper methods might find local optimum, so it is known to run these algorithms multiple times and report the best results. Therefore, it is customary to report the results of a ten repeat (10R) of each of these three algorithms. Although comparing the run-time of a 10R algorithm with one repeat of the proposed algorithm might seem unfair, bear in mind that most of the time, a single run of those other algorithms does not lead to acceptable results. In contrast,  $k$ -splits needs no repetition, leading to deterministic results that remain unchanged through reruns.

Experiments are conducted using Python 3.7.4 on a system with Intel Core i5-4200 M CPU@ 2.50 GHz, 4 GB memory, and 1 T hard disk space. Scikit-learn package implementation of  $k$ -means and PyClustering package implementation of  $g$ -means and  $x$ -means are used. Furthermore, numpy.linalg, implemented using LAPACK routines, is used to obtain eigenvalues and eigenvectors.  $\beta = 0.01$  is

**Table 1** Datasets specifications

Dataset	N	C	D	Overlap	Dataset	N	C	D	Overlap
<i>A1</i>	3000	20	2	20%	<i>G2-2-30</i>	2048	2	2	15%
<i>A2</i>	5250	35	2	20%	<i>G2-2-50</i>	2048	2	2	43%
<i>A3</i>	7500	50	2	20%	<i>G2-128-10</i>	2048	2	128	13%
<i>S1</i>	5000	15	2	9%	<i>Birch1</i>	100k	100	2	52%
<i>S2</i>	5000	15	2	22%	<i>Birch2</i>	100k	100	2	4%
<i>Dim32</i>	1024	16	32	0%	<i>MNIST</i>	60k	10	87	-
<i>Dim1024</i>	1024	16	1024	0%	<i>F-MNIST</i>	60k	10	84	-
<i>Unbalance</i>	6500	8	2	0%					

N—number of data points, C—number of clusters, D—dimensions

**Table 2** Comparison results of the algorithms

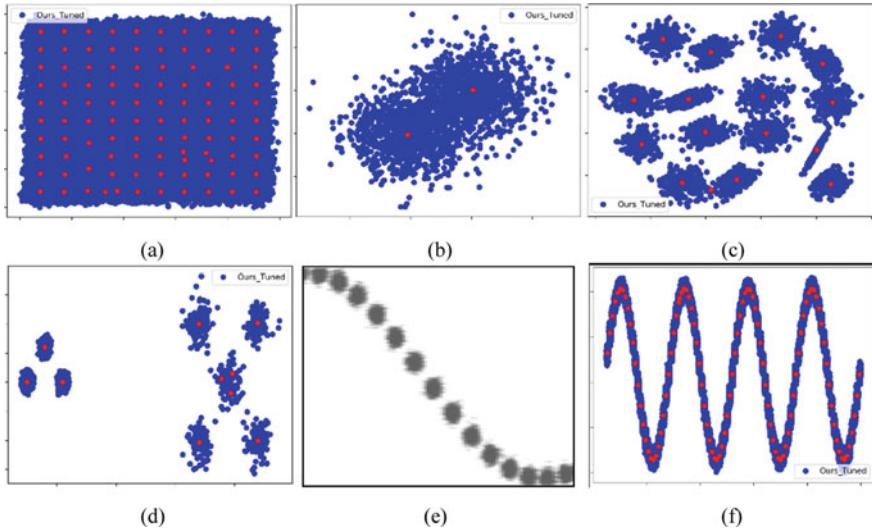
Dataset	K-splits			Fine-tuned k-splits		10R G-means			10R X-means			10R K-means	
	k	t (s)	ARI	t (s)	ARI	k	t (s)	ARI	k	t (s)	ARI	t (s)	ARI
A1	<b>20</b>	0.10	0.80	<b>0.14</b>	<b>1.0</b>	27	0.17	0.89	<b>20</b>	0.16	<b>1.0</b>	0.18	<b>1.0</b>
A2	<b>35</b>	0.22	0.87	<b>0.23</b>	<b>1.0</b>	41.2	0.37	0.94	20	0.27	0.62	0.54	0.99
A3	<b>50</b>	0.33	0.90	<b>0.38</b>	0.97	57.8	0.86	0.95	4	<u>0.18</u>	0.11	0.86	<b>0.99</b>
S1	<b>16</b>	0.27	0.96	0.27	0.98	233.4	1.73	0.19	20	0.40	0.94	<b>0.14</b>	<b>0.99</b>
S2	<b>16</b>	0.29	0.87	0.31	0.92	194.2	1.93	0.20	18.6	0.37	0.90	<b>0.20</b>	<b>0.94</b>
Dim32	<b>16</b>	0.13	<b>1.0</b>	0.13	<b>1.0</b>	649.6	1.42	0.07	16.8	<b>0.07</b>	<b>1.0</b>	0.08	<b>1.0</b>
Dim1024	17	310	<b>1.0</b>	313	<b>1.0</b>	827.4	36.3	0.03	<b>16</b>	0.91	<b>1.0</b>	<b>0.79</b>	<b>1.0</b>
G2-2-30	<b>2</b>	0.01	<b>0.96</b>	<b>0.01</b>	<b>0.96</b>	<b>2</b>	0.06	0.95	<b>2</b>	0.06	0.95	0.02	<b>0.96</b>
G2-2-50	<b>2</b>	0.01	<b>0.70</b>	<b>0.01</b>	<b>0.70</b>	<b>2</b>	0.06	<b>0.70</b>	<b>2</b>	0.04	<b>0.70</b>	0.03	<b>0.70</b>
G2-128-10	<b>2</b>	0.07	<b>1.0</b>	<b>0.09</b>	<b>1.0</b>	779	6.60	0.00	<b>2</b>	0.24	<b>1.0</b>	0.11	<b>1.0</b>
Birch1	<b>101</b>	2.36	0.66	<b>4.85</b>	0.94	2562	741	0.08	4	<u>3.66</u>	0.06	48.32	<b>0.95</b>
Birch2	<b>101</b>	2.04	0.98	<b>2.58</b>	<b>1.0</b>	126.6	8.51	0.93	20	4.77	0.30	15.44	<b>1.0</b>
Unbalance	<b>10</b>	0.28	1.0	0.30	<b>1.0</b>	14.2	0.46	1.0	4	0.24	0.99	<b>0.08</b>	<b>1.0</b>
MNIST	<b>19</b>	3.80	0.20	<b>7.16</b>	<b>0.36</b>	11.9 k	16 h	0.00	20	71.6	<b>0.36</b>	25.54	<b>0.36</b>
F-MNIST	<b>7</b>	1.96	0.21	<b>2.98</b>	<b>0.37</b>	14.2 k	21 h	0.00	20	57.5	0.34	15.63	0.35

k—number of predicted clusters, t(s)—execution time in seconds

set for half the experiments, and for the other half containing medium density and overlap such as A and S,  $\beta = 0.1$  is used. For pictorial datasets MNIST and Fashion-MNIST, first PCA (Principal Component Analysis) with 0.9 variance is applied to reduce the original dimension size of 784, to 87 and 84 respectively, then  $\beta = 0.95$  is set due to the density of data points and distances, and then the algorithm is applied.

The results are summarized in Table 2. Each value is the mean across five validation folds. Table 2 reports the number of detected clusters ( $k$ ), execution time (t reported in seconds), and an external validity measure called the *adjusted rand index* (ARI). ARI is a score that shows the similarity between two sets of clustering results and is equal to 0 for random results and 1 for exact clustering matches. For each dataset, the best instances of the number of predicted clusters ( $k$ ), execution time (t), and ARI are in bold font with some exceptions. One exception is that execution time is comparable only if both methods provide “acceptable” results; thus, lower run-time with terrible results in Table 2 are only underlined and not bolded. The other exception is that fine-tuned  $k$ -splits is the improved version of  $k$ -splits; thus, it is enough to compare the fine-tuned version’s execution time.

In predicting the number of clusters,  $k$ -splits’ predictions are the closest to reality, except for Dim1024, and even there, the error is minimal. Whereas g-means almost always over-splits the data, with many unacceptable results. One extreme example is the MNIST and F-MNIST datasets, for which the algorithm predicts more than ten thousand clusters! This situation might be an excellent example of the black hole problem mentioned before. Getting stuck in very dense clusters misguides the



**Fig. 3** Predicted centroids using  $k$ -splits on benchmark datasets. **a** Birch, **b** G2-2-50, **c** S1 and **d** unbalance datasets. **e** and **f** Zoomed structure of Birch2 and our results

algorithm, leads to false results, and wastes time and effort, the exact thing this research avoids by implementing multiple stop conditions in  $k$ -splits. X-means, on the other hand, seems to under-split in many cases. K-splits accurately predicts the number of clusters in almost all cases. Some examples of clustering results using  $k$ -splits are shown in Fig. 3.

Regarding the execution time, g-means is always slower, especially for large (and dense) datasets; the black hole problem completely ruins the process leading to extreme examples like the Fashion-MNIST dataset, which took 21 hours to be processed. The  $x$ -means algorithm takes an acceptable amount of time to finish, but we should consider that under-splitting of data might be a factor in that. K-splits (even the fine-tuned version) is faster than g-means and  $x$ -means under almost all situations, with few exceptions. It is also faster than the conventional  $k$ -means in many cases.

The time efficiency of  $k$ -splits is the result of three decisions in each iteration:

- It only splits the so-called “the worst cluster” and keeps other clusters intact.
- It only uses the  $k$ -means algorithm to separate only two clusters in each step, which is very simple to solve.
- It gives the  $k$ -means algorithm accurately calculated initial centers leading to much faster convergence and better results than random initialization.

High dimensionality harms the time efficiency of the  $k$ -splits algorithm. The type of computations used here grows in difficulty and negates the aforementioned positive effects in higher dimensions. Although the algorithm yields good results, the

time consumption in dimensions higher than 1000 is high. Therefore, a dimension reduction before using this kind of data can be helpful.

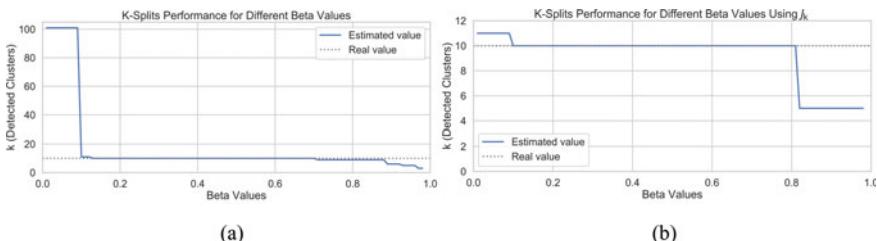
Comparing the ARI values in Table 2 indicates that  $k$ -splits is always more accurate than both  $g$ -means and  $x$ -means in assigning each data point to the correct cluster. It is also clear that it is always a good idea to fine-tune  $k$ -splits which leads to much better results, some even better than the standard  $k$ -means. That is because  $k$ -splits pinpoints the centroids, which is an excellent boost for the  $k$ -means algorithm. The time cost of fine-tuning is also acceptable. In some cases, like A1 and A2 datasets, the fine-tuned  $k$ -splits is still more time-efficient than the standard  $k$ -means, and in some unique structures like Birch, this time gap is more significant.

One should bear in mind that the main superiority of  $k$ -splits is its ability to automatically find the correct number of clusters. Hence, better accuracy or faster results than the original  $k$ -means with the right  $k$  as input is not expected. However, as shown in Table 2, the execution time of  $k$ -splits is faster than  $k$ -means in some cases, and the performance is acceptable. Part of this performance gap is due to the rigidness of  $k$ -splits. Once a cluster is separated into two, those data points cannot move to other clusters. The combination of  $k$ -splits and  $k$ -means, introduced as fine-tuned  $k$ -splits, solves this problem. Fine-tuned  $k$ -splits benefits from the speed and accuracy of  $k$ -splits in finding the right  $k$  and the high performance of standard  $k$ -means in assigning each data point to the clusters.

## 4.2 Effect of Hyperparameter $\beta$

This section experiments with different values of hyperparameter  $\beta$  on a synthetic dataset to investigate the sensitivity of the algorithm to this hyperparameter. Two tests are conducted, one without applying condition (15) and another with this condition and using  $J_K$  as a second stop criterion, the results are shown in Fig. 4.

The synthetic dataset used in these experiments has a size of  $N = 10,000$  data points distributed in ten clusters and ten dimensions. It is evident from Fig. 4a that  $k$ -splits successfully predicts the correct number of clusters in a wide range of  $\beta$ , but outside this range, the algorithm fails. However, Fig. 4b shows that using  $J_K$  as



**Fig. 4** K-Splits performance in finding ten clusters: **a** using only  $\beta$  as stop condition, **b** using  $J_K$  as stop condition

a secondary stop condition helps decrease this sensitivity and provides acceptable results for almost the whole range. Nonetheless, the problem with  $\beta$  is not entirely solved. Firstly, as mentioned in Sect. 3,  $J_K$  cannot be used for highly overlapped and dense data, and secondly, although  $\beta$  has an acceptable range for each dataset, this range changes as the structure of the dataset, especially the density, changes.

### 4.3 Effects of Dataset Parameters

Several other experiments were also conducted to further analyze the effects of dataset parameters on  $k$ -splits. The datasets used in this section share the specifications mentioned above, and in each experiment, only one parameter of the dataset is analyzed. These parameters are dataset size, number of clusters, and dimension size. Changing the parameters leads to a different set of data with a new structure; thus, a fivefold scheme was utilized to calculate each point, and regression plots were created using central tendency and confidence intervals.

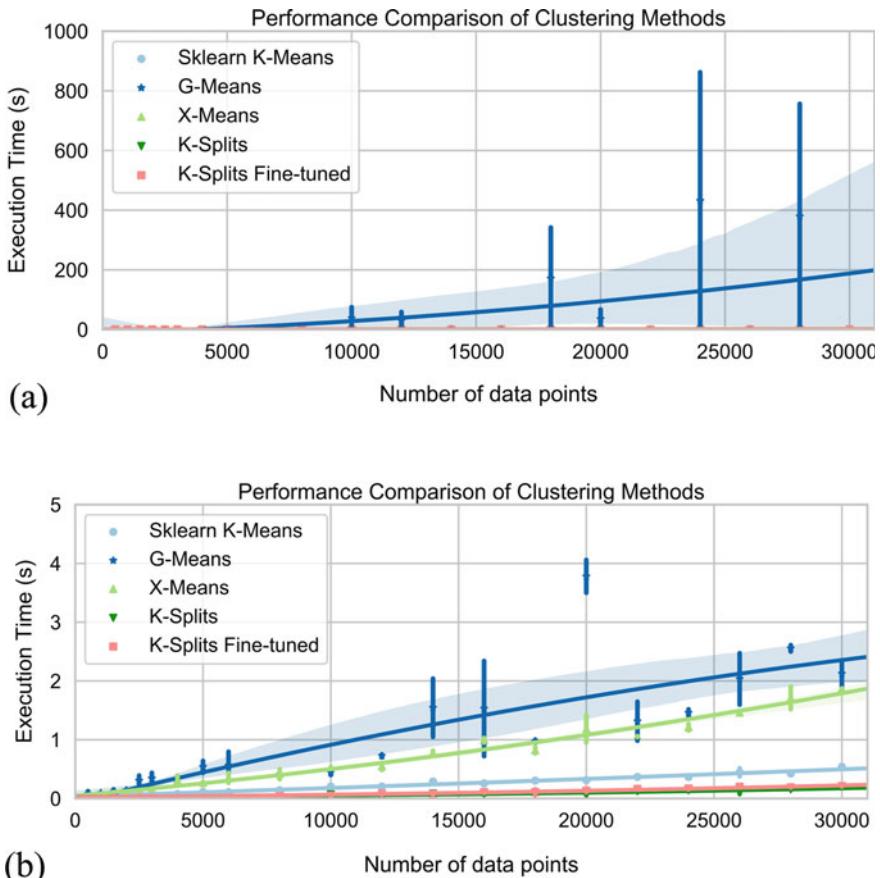
**Dataset Size (N).** For this experiment, dataset sizes in a range of [500, 30,000] are analyzed. This test provides a good sense of the time complexity of methods; that is why it is conducted on  $k$ -splits and 10R versions of  $k$ -means,  $x$ -means, and g-mean simultaneously. The results are shown in Fig. 5.

As mentioned earlier, the g-means algorithm experiences critical problems, getting stuck in local optimum and black holes. These problems are readily detectable in Fig. 5a as in some points, abnormalities can be seen in the form of strong peaks in execution time. In this example, these peaks get as high as 800 s, while execution time for other methods and even other points of the same algorithm is well below 5 s. In conclusion, g-means is the slowest method among these methods.

To further compare these methods, in the next test, it is assumed that g-means never encounters the black hole problem, and all these problematic points are discarded; the result is Fig. 5b. It is evident that regardless of the problems caused by getting stuck, g-means is still the slowest algorithm. After that,  $x$ -means and  $k$ -means come second and third, respectively, and both  $k$ -splits and its fine-tuned version execute faster than all three methods.

**Number of Clusters (C).** The number of clusters is another critical parameter in each dataset. Therefore, this section tests the  $k$ -splits algorithm under different numbers of clusters in the range of [2, 100]. The results are presented in Fig. 6. Each dataset instance has a different structure; moreover, changing the number of clusters in a limited space changes the density and overlap. Thus, as presented in Fig. 6a, the algorithm cannot accurately predict larger cluster numbers using one constant  $\beta = 0.5$ . However, if  $\beta$  is scheduled to change through the experiment based on the prior knowledge of the structure (similar to real applications), the performance will increase according to Fig. 6b.

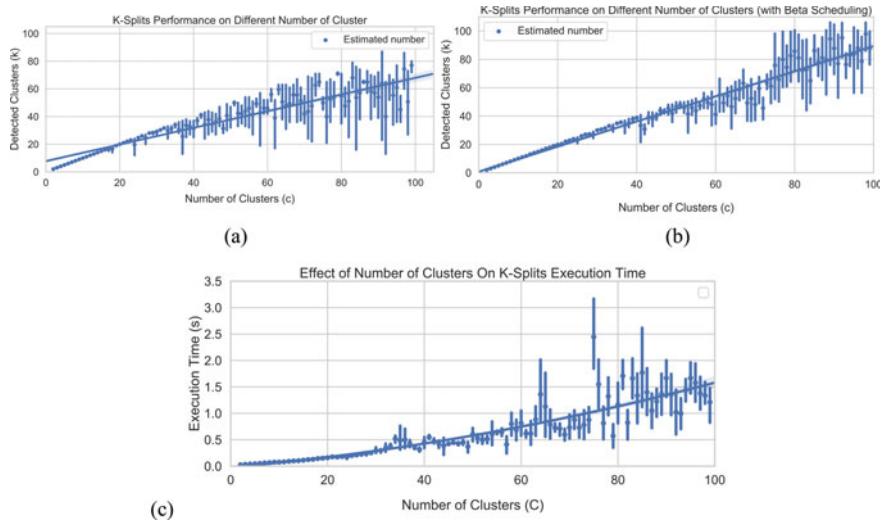
Needless to point out that although one can achieve great results using a good guess about the range of  $\beta$ , it is still one of the main downsides of the  $k$ -splits algorithm. Different cluster sizes require a different range of  $\beta$ , a good guess of which is needed



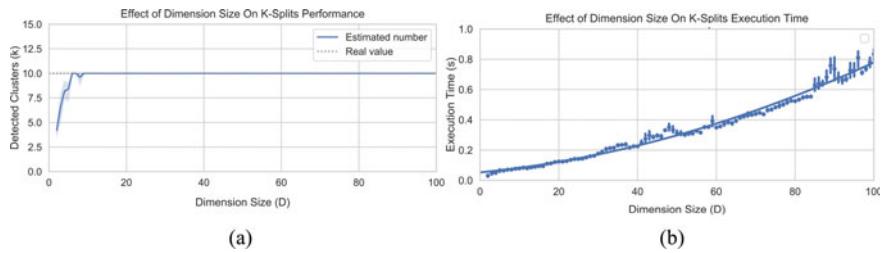
**Fig. 5** Comparison of methods regarding execution time. **a** Actual values, **b** values in case g-means never encounters the black hole problem

to predict the number of clusters in turn! That is why checking the density of data is a good step in setting hyperparameter  $\beta$ . Finally, the effect of the number of clusters on execution time is analyzed, and the results are of order two and presented in Fig. 6c.

**Data Dimension Size (D).** As the last experiment, the effect of dataset dimension on  $k$ -splits is checked. A constant  $\beta = 0.5$  is used throughout the test. The effects of dimension size on the proposed algorithm's accuracy and execution time are shown in Fig. 7. It can be seen that  $k$ -splits accurately estimates the number of clusters in different dimensions, except in about ten first instances, which have lower accuracy. This low accuracy might occur because of the data's higher density due to very low dimensions. Figure 7 shows that dimension has minimal effect on the range of suitable  $\beta$ , but should not be neglected.



**Fig. 6** Effect of the number of clusters on  $k$ -splits. **a** Performance using a constant  $\beta = 0.5$ . **b** Performance using a variable  $\beta$ . **c** Effect of the number of clusters on execution time



**Fig. 7** The effect of data dimension on  $k$ -splits: **a** effect on cluster prediction accuracy, **b** effect on execution time

## 5 Conclusion

This article proposed  $k$ -splits: an incremental  $k$ -means based clustering algorithm to automatically detect the number of clusters and centroids. It also introduced a fine-tuned version of  $k$ -splits that uses these centroids as initialization for the standard  $k$ -means and dramatically improves the performance. Moreover, this work used six synthetic datasets and also MNIST and Fashion-MNIST datasets to show that  $k$ -splits can accurately find the correct number of clusters and pinpoint each cluster's center under different circumstances.  $K$ -splits is faster than  $g$ -means and  $x$ -means and, in some cases, even faster than the standard  $k$ -means. The accuracy of the results and the performance are also higher than these methods. Furthermore,  $k$ -splits needs no repetition as it leads to deterministic results.

K-splits starts from a small number of clusters and further splits each cluster if needed. The starting point does not have to be one cluster and, if known, can be set by the user leading to an even faster result. The algorithm takes one threshold parameter  $\beta$  as input, which controls separation condition and should be set based on our prior knowledge of the data density. Clustering in very high-dimensional spaces with  $k$ -splits is time-consuming, so it is preferable to use dimension reduction techniques before applying the algorithm. Future works will focus on two improvements to this research. First, to optimize the calculations needed to obtain similar results, especially in higher dimensions, and second, to eliminate the need for any extra hyperparameters if possible.

## References

1. Thorndike, R.L.: Who belongs in the family? *Psychometrika* **18**(4), 267–276 (1953). <https://doi.org/10.1007/BF02289263>
2. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987)
3. Chen, J., Qin, Z., Jia, J.: A weighted mean subtractive clustering algorithm. *Inf. Technol. J.* **7**(2), 356–360 (2008)
4. Ahmed, M., Seraj, R., Islam, S.M.S.: The k-means algorithm: a comprehensive survey and performance evaluation. *Electronics* **9**(8), 1295 (2020). <https://doi.org/10.3390/electronics9081295>
5. Haoxiang, W., Smys, S.: Big Data analysis and perturbation using data mining algorithm. *J. Soft Comput. Paradig.* **3**(1), 19–28 (2021)
6. Smys, S., Raj, J.S.: Analysis of deep learning techniques for early detection of depression on social media network—a comparative study. *J. Trends Comput. Sci. Smart Technol.* **3**(1), 24–39 (2021)
7. Steinhaus, H.: Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci., C1. III* **IV**, 801–804 (1956)
8. Lloyd, S.: Last square quantization in PCM's. Bell Telephone Laboratories Paper (1957). Published in journal much later: S. P. Lloyd. Least squares quantization in PCM. Special issue on quantization. *IEEE Trans. Inform. Theory* **28**, 129–137 (1982)
9. Kaufman, L., Rousseeuw, P.J.: *Finding Groups in Data: An Introduction to Cluster Analysis*, vol. 344. Wiley, Hoboken, NJ, USA (2009)
10. Peña, J.M., Lozano, J.A., Larrañaga, P.: An empirical comparison of four initialization methods for the K-means algorithm. *Pattern Recognit. Lett.* **20**(10), 1027–1040 (1999). [https://doi.org/10.1016/S0167-8655\(99\)00069-0](https://doi.org/10.1016/S0167-8655(99)00069-0)
11. Yuan, C., Yang, H.: Research on K-value selection method of K-means clustering algorithm. *J—Multi. Sci. J.* **2**(2), 226–235 (2019). <https://doi.org/10.3390/j2020016>
12. Pelleg, D., Moore, A.: X-means: extending K-means with efficient estimation of the number of clusters. In: *Proceedings of the 17th International Conf. on Machine Learning*, pp. 727–734. Morgan Kaufmann, San Francisco, CA (2000)
13. Hamerly, G., Elkan, C.: Learning the k in k-means. *Adv. Neural. Inf. Process. Syst.* **16**, 281–288 (2004)
14. Yuan, F., Meng, Z.H., Zhang, H.X., Dong, C.R.U.: A new algorithm to get the initial centroids. In: *Proceedings of 2004 International Conference on Machine Learning and Cybernetics*, 2004, vol. 2, pp. 1191–1193
15. Zechner, M., Granitzer, M.: Accelerating K-means on the graphics processor via CUDA. In: *2009 First International Conference on Intensive Applications and Services*, April 2009, pp. 7–15. <https://doi.org/10.1109/INTENSIVE.2009.19>

16. Zhang, J., Wu, G., Hu, X., Li, S., Hao, S.: A parallel K-means clustering algorithm with MPI. In: 2011 Fourth International Symposium on Parallel Architectures, Algorithms and Programming, Dec 2011, pp. 60–64
17. Poteraş, C.M., Mihăescu, C., Mocanu, M.: An optimized version of the K-means clustering algorithm. In: 2014 Federated Conference on Computer Science and Information Systems, FedCSIS 2014, Sept 2014, pp. 695–699
18. Nazeer, K.A.A., Sebastian, M.P.: Improving the accuracy and efficiency of the k-means clustering algorithm. Proc. World Congr. Eng. **I**(July), 1–3 (2009)
19. Golub, G.H., Reinsch, C.: Singular value decomposition and least squares solutions. In: Linear Algebra, pp. 134–151. Springer, Berlin Heidelberg (1971)
20. Fränti, P., Sieranoja, S.: K-means properties on six clustering benchmark datasets. Appl. Intell. **48**(12), 4743–4759 (2018)
21. LeCun, Y., Cortes, C.: MNIST handwritten digit database. AT&T Labs, 2010 [Online]. Available: <http://yann.lecun.com/exdb/mnist>
22. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arxiv, pp. 1084–1091, Aug. 2017, [Online]. Available: <http://arxiv.org/abs/1708.07747>

# Implementation of IoT-Based Intelligent Patient Healthcare Monitoring System Using KNN Algorithm



G. Sreenivasulu and T. P. Anithaashri

**Abstract** Healthcare is an effective factor and one of the most important primary capabilities of a human life. Nowadays, Internet of Things (IoT) is important in all fields, including healthcare and medical industries due to the increased use of Wi-Fi sensors. The implementation of new machine intelligence-based technological e-healthcare development and e-healthcare parameters will now involve and include Internet and Wi-Fi sensors, which are commonly referred to as Internet of Things (IoT). Machine intelligence enables and activates a global approach to the development of healthcare monitoring-based systems. Machine intelligence is the development of machines, and intelligence necessitates a significant amount of relevant knowledge. Manual health monitoring is becoming more difficult in today's busy world. This system establishes a real-time approach and provides a set of specifications. This information is relevant to the health of the monitored patients. Many countries are reporting an increase in the number of deaths as a result of a lack of certain factors, such as timely medical treatments and a monitoring system. This paper has proposed a smart healthcare methodology for monitoring the patients by involving healthcare sensors to reduce doctor's workload in order to detect the heartbeat and body temperature and inform the doctor as well as attending hospital staff about the patients' status by providing accurate results. The proposed system employs the KNN algorithm to forecast the patient's condition in able to prevent the patient from becoming further ill. The proposed system also includes a buzzer beeping situation, which means that if the beeping occurs, the nurse must be notified of the patient's emergency situation. The existing system did not include a predicting component. As a result, the proposed system provides accurate results for patients, allowing them to easily handle the emergency situation.

---

G. Sreenivasulu (✉)

Department of CSE, Audisankara College of Engineering and Technology, ASCET, Gudur, Nellore District, Andhra Pradesh, India

e-mail: [gudalirinivas@gmail.com](mailto:gudalirinivas@gmail.com); [srinivas.g@audisankara.ac.in](mailto:srinivas.g@audisankara.ac.in)

T. P. Anithaashri

Department of CSE, Saveetha School of Engineering, SIMATS, Chennai, India

e-mail: [anithaashript.sse@saveetha.com](mailto:anithaashript.sse@saveetha.com)

**Keywords** Healthcare · IoT · Machine intelligence · KNN · Methodology · Emergency · Confusion matrix · Sensors

## 1 Introduction

This main objective of the proposed healthcare machine intelligent system is the adoption, integration, and utilization of all the communication systems. IoT-based healthcare monitoring system will effectively monitor the patients and health status. IoT is implemented by the application of machine intelligence system in order to improve the user behavioral patterns and gain knowledge of context action rules for its relation with the user's behavioral patterns etc. The term Internet of Things (IoT) define the services to deliver health assistance for people is ambient machine intelligence and particularly dealing with healthcare monitoring system of admitted patients with the critical conditions. The main goal of this paper is to define a healthcare model based on Internet of Things (IoT) to reduce the doctor's workload and tedious process of the nurses to check the admitted severely suffering patients during certain emergency situations. Firstly, to read the body temperature, heartbeat, and pulse rate of the patient using healthcare sensors with the patients and their reports. Secondly, the patient buzzer buzzes to indicate the patient emergency situation to the nurse and present in that particular patient wards. Also, the nurse will be equipped with healthcare sensors to monitor the patient's critical condition at regular intervals. Thirdly, a message is sent to the doctor's mobile in case of any critical emergency situations, where the presence of ward nurse is insufficient to treat the patients.

## 2 Existing System

Smart healthcare system is one of the most basic requirements to everyone. The healthcare products such as Tonic water with Quinine, skin collagen stimulation derma roller-540 micro-needles, 3 in 1 Infrared light ultrasonic sound electro-matic stimulation, Egyptian magic skin cream olive oil, beeswax, honey, bee pollen, propolis, electroporation, radio frequency, LED photon, high-frequency vibration, HAELO frequency therapy, hyaluronic acid serum skin care, hustle drops respiratory performance supplement, iron off stubborn fat and stretch marks, peak low testosterone treatment, Hydroderm abrasion skin care device, signal relief nanotech pain relief patch, hydrogen water generator, wearable Taopatch acupuncture, light nanotechnology, neorhythm neurostimulation headband, back support posture corrector, OrCam MyEye for blind visually impaired, vibrating plaque remover, LED light teeth whitener, dental teeth aligner Orthadonic retainer, Silko 12,000 bristles toothbrush, sonic brush vibrating toothbrush, and jaw teeth bite exerciser are used as monitoring instruments.

The literature review reveals that there around 180 number of research articles published in this topic over the past 5 years. W. Y. Chung and S. J. Jung proposed “The Flexible and scalable patient’s health monitoring system in 6LoWPAN.” The Proposed System enabling factor, some technologies and communications, Internet of Things telecommunications, informatics and electronics. M. J. Mao Kaiver and K. S. Shin proposed “A Cell phone based health monitoring system with self-analysis.” The proposed system uses smart objects interacting with the physical world. Tabilo Paniclo and Gennaro tartarisco proposed “A Maintaining sensing coverage and connectivity in large sensor networks.” The proposed system includes the information about how to develop a new computational technology, information processing, and wireless communication personal healthcare.

The analysis on healthcare sensors, irregular food habits, no nutrition diet, environmental pollutions etc., has been interpreted and arrived with less efficiency in execution. The main advantage of this enabling unhealthy human life style. Khan et al. [1] included non-physician health workers. The main advantage of this is tele-consulting and video conferencing. Healthcare system [2], which is capable of measuring different physiological parameters like temperature sensor, humidity sensor, pulse sensor, WSN, WDM, UV, CO<sub>2</sub> sensor, ECG, pulse sensor, temperature, and camera and are used to design a system for headache and rapid pulse rate detection.

The analysis to detect ear diseases [3] system has been developed, which is capable of measuring different physiological parameters like temperature sensor, gas sensor, heartbeat sensor, and Raspberry Pi and are used to design a system for pulse temperature and smoke detection. Using smartphone and laptop, VGA display developed a system [4, 5], which is capable of measuring different physiological parameters like ECG, Bluetooth, temperature sensor, heart rate sensor, Arduino, and biosensor and are used to design a system for detecting the abnormalities in heart. To measure a different physiological parameters detection instruments like blood pressure sensor, body weight sensor, pulse oximeter, glucometer, accelerometer etc., and are used to design a system for chronic disease [6, 7] progression. Healthcare applications a solution based on the Internet of Things (IoT). This survey aims to implement quality of advanced technology in medicine and nursing and patient care and safety can be used to develop wearable sensors, medical equipment, and implantable [8] devices.

The healthcare monitoring system in the IoT by using KNN [1] technology is used to take care of the patients of the part of Wi-Fi, wearable, and healthcare sensing the desirable values through the healthcare sensors and displaying patients data in their respective Web page [9] and their results. The continuous checking the patients of the value from the healthcare sensors and the relative Web page in one of the most heavy works. Human is always intended to do certain mistakes and produces errors, so if the person checking the patient diverts the health condition, some [10] mishappenings and misresults may occur.

The limitation of the existing system is more manual work, sensor failures, and power failures. Also, requires latest values from the sensors will show so that no prediction can be done with their [11, 12] monitoring values. The objective of the proposed system is to provide enhanced healthcare monitoring system [13] consists

of various doctor's monitoring devices that are used to monitor and assist the patients and also provide alerts, if the patient gets into a critical state such as a slowdown of heartbeat and blood pressure.

### 3 Proposed System

The proposed system provides a smart system which can be used by people to solve various healthcare issues. This paper also presents healthcare diagnosis treatment and prevention of disease, illness, and injury in human. This domain is automatically learning some task of healthcare information, medical management, patient health information etc.

This proposed work's main aim is to design and implement a hospital monitoring system using machine intelligence algorithms. The machine methodology parameters of the patients to collect in real time by the health sensor networks in the normal environmental patient conditions, and it stores the patient data in the Web server where it is patient-predicted by the health prediction algorithm, so, it displays the current patient data in the health monitoring webpage, and it sends the signals according to the health prediction in case of emergency cases. It can be classified into three main parts

- (i) Healthcare sensing system
- (ii) Patient health prediction system
- (iii) Patient emergency alert system (Fig. 1).

#### 3.1 *Healthcare Sensing System*

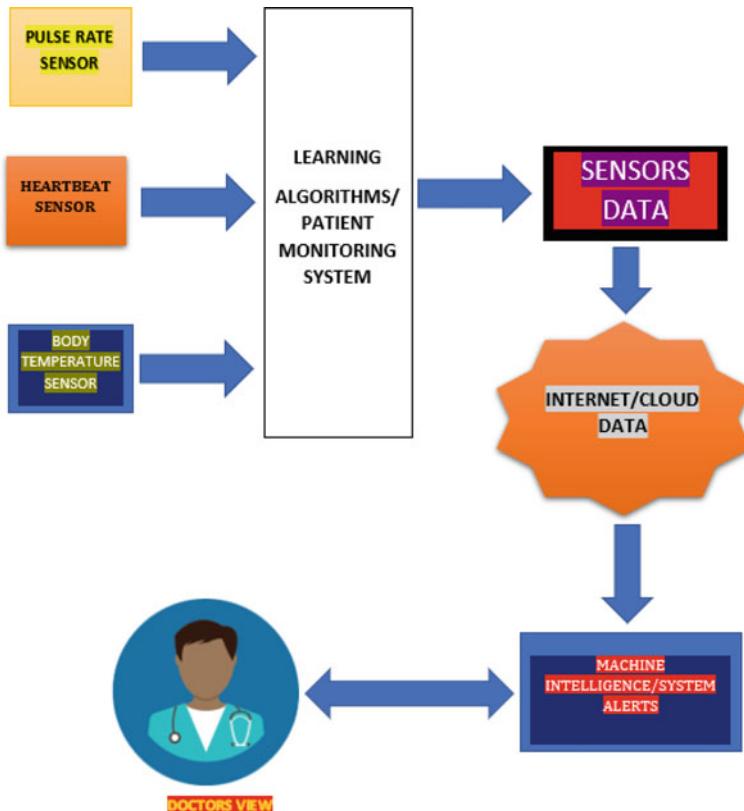
##### 3.1.1 **Body Temperature Sensor**

Body temperature sensor is used to check whether patient condition normal or abnormal. It indicates high body temperature or a fever, and it is used to convert Celsius by the analog to digital. The body temperature sensor is attached to the patient bed, and then the temperature is collected continuously.

##### 3.1.2 **Heartbeat Sensor**

It is an electronic device, and it is used to measure speed of heartbeat rate. It converts values from analog to digital. It is used to measure heartbeat rate in two ways.

- (a) **Manual way:** In this way, to check radial pulse and carotid pulse manually.
- (b) **Using a sensor:** It can be measured light scattered or absorbed based on optical power variation.



**Fig. 1** Architecture of proposed system

### 3.2 Patient Health Prediction System

Even though, the patient prediction value collected from the healthcare sensor is printed on the webpage, and it will be more efficient and also more accurate. When the healthcare system predicts the normal or abnormality of the patient, then it gives an alert message. Machine learning becomes more efficient and also effective in action using KNN. KNN algorithm has been used for the patient prediction of the healthcare sensor predicted values collected. K-nearest neighbor (KNN) algorithm is the simplest nearest neighbor algorithm and works efficiently in practice.

This algorithm can be classified into two ways

- (a) **Supervised algorithm**
- (b) **Unsupervised algorithm.**

It uses patients' predicted data and classify the new data predicted points based on the similar measurements. The predicted data is assigned to the class which has the patients nearest neighbors. KNN algorithm uses all the predicted training datasets to

predict the patient result based on the best subset of the predicted training dataset. This algorithm is implemented in Python which contains several methods.

### **3.3 Patient Emergency Alert System**

The patient predicted data gives the alert signal to the Raspberry Pi where an emergency buzzer is connected to it the emergency buzzer cannot work alone; it should be connected to patient bed sensor.

The patient emergency buzzer has three pins (a) **ground**, (b) **power supply, and** (c) **a value pin**.

## **4 Results and Discussion**

This system is used to test various persons with normal to abnormal healthcare monitoring conditions and then produces results with minimal error rate such as temperature findings with minimal error of + or - 5, ECG findings to measure pulse rate in the error range of + or - 6, and EEG sensor is used to describe error rate (Tables 1, 2, and 3).

### **Classification Accuracy**

Classification accuracy is the percentage of correct healthcare predictions. It is for evaluating the different models. Accuracy is the fraction of predictions where the model got right (Fig. 2; Table 4).

$$\text{Accuracy} = \frac{\text{Number of exact predictions}}{\text{Total number of predictions}}$$

Thus, the algorithm used is comparatively better than the other algorithms used. It also works best for real-time data processing and analyzing of data.

**Table 1** Patient temperature sensor

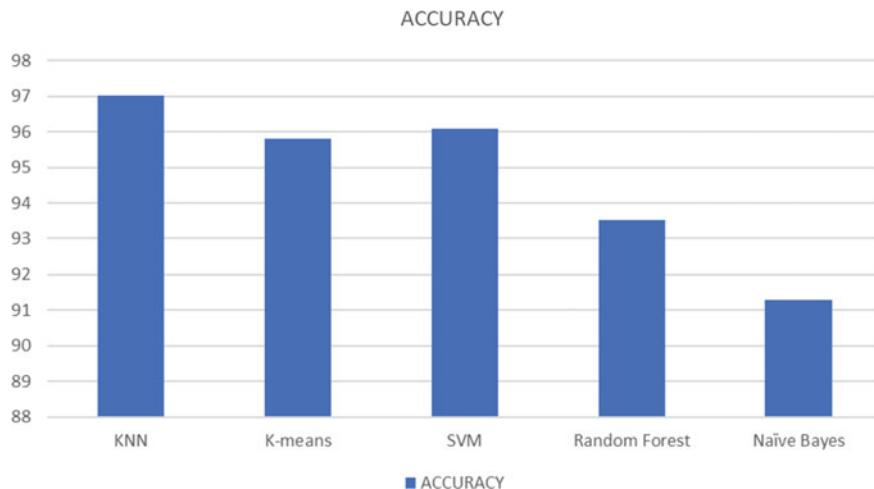
Patient testings	Patient normal value	Patient observed value	Patient error rate
Person 1	24	29	+5
Person 2	24	28	+4

**Table 2** Patient ECG measurement sensor

Patient testings	Patient normal value	Patient observed value	Patient error rate
Person 1	74–78	71	-3
Person 2	74–78	83	+5

**Table 3** Patient EEG measurement sensor

Patient testings	Patient normal value	Patient observed valued	Patient status
Person 1	100–300	231	Active
Person 2	100–300	365	InActive

**Fig. 2** Accuracy comparison**Table 4** Accuracy comparison values

Algorithms	Accuracy
KNN	97.02
K-means	95.8
SVM	96.1
Random forest	93.52
Naïve Bayes	91.3

## 5 Conclusion and Future Enhancement

The proposed patient healthcare monitoring system can be mostly used in emergency situations, and it daily monitored the patients, recorded the patients' data, and stored the daily patient as a database. In future, the IoT and machine learning technologies have provided great opportunities with the cloud computing so that some challenges are to be overcome, and the patient database can be easily shared in all the different hospitals for healthcare monitoring and treatment.

## References

1. Khan, S.F.: Health care monitoring system in the Internet of Things (a lot) by using RFID. In: 6th International Conference on Industrial Technology and Management 2017
2. Zamora, A., Skarmeta, A.F.G., Jara Miguel, A.J.: An architecture based on Internet of Things to support mobility and security in medical environments. In: IEEE 2010
3. Khan, I., Zeb, K., Mahmood, A., Uddin, W., Khan, M.A., Ul-Islam, S., Kim, H.J.: Healthcare monitoring system and transforming monitored data into real time clinical feedback based on IoT using Raspberry Pi. In: ICOMET 2019
4. Geman, O., Costin, H.-N., Chiuchisan, I.: Adopting the Internet of Things technologies in health care systems. In: International Conference and Exposition on Electrical and Power Engineering 2014
5. Ganorkar, S., Koshti, M.: IoT based health monitoring system by using Raspberry Pi and ECG signal. *Int. J. Innov. Res. Sci., Eng. Technol.* **5**(5) (2016)
6. Gope, P., Hwang, T.: BSN-care: a secure IoT-based modern healthcare system using body sensor network. *IEEE 1530-437X* (c) 2015
7. Hassanaliagh, M., Page, A., Soyata, T., Sharma, G., Aktas, M., Kantarci, G.M.B., Andreescu, S.: Health monitoring and management using Internet-of-Things (IoT) sensing with cloud-based processing: opportunities and challenges. In: IEEE International Conference on Services Computing, 2015
8. Shamim Hossaina, M., Muhammad, G.: Cloud-assisted industrial Internet of Things (IoT)—enabled framework for health monitoring. *Comput. Netw.* (2016)
9. Rahmani, A.-M., Thanigaivelan, N.K., Gia, T.N., Granados, J., Negash, B., Liljeberg, P., Tenhunen, H.: Smart e-health gateway: bringing intelligence to Internet-of-Things based ubiquitous healthcare systems. In: 12th Annual IEEE Consumer Communications and Networking Conference (CCNC), 2015
10. Hassanaliagh, M., Page, A., Soyata, T., Sharma, G., Aktas, M., Mateos, G., Kantarci, B., Andreescu, S.: Health monitoring and management using Internet-of-Things(IoT) sensing with cloud-based processing:opportunities and challenges, IEEE 2015
11. Catarinucci, L., De Donno, D., Mainetti, L., Palano, L., Patrono, L., Stefanizzi, M.L., Tarricone, L.: An IoT-aware architecture for smart healthcare systems. *IEEE Internet of Things J.* (2015)
12. Mieronskia, R., Azimi, I., Rahmani, A.M., Aantaa, R., Terava, V., Liljeberg, P., Salanterä, S.: The Internet of Things for basic nursing care. *Int. J. Nurs. Stud.* **69**, 78–90 (2017)
13. Natarajan, K., Prasath, B., Kokila, P.: Smart health care system using Internet of Things. *J. Netw. Commun. Emerg. Technol.* **6**(3), (2016)

# Comprehension of Ultra-Wideband Transceiver for Wireless Communication System based on Code-Shifted Reference



R. Santosh Kumar, Rajashree Narendra, and R. Devaraju

**Abstract** In this paper, research on the recent developments in the insights of ultra-wideband transceiver for wireless communication system based on code-shifted references is presented. There have been extensive studies carried out individually towards ultra-wideband (UWB) transceiver and security-based approaches. However, none of them have jointly studied securing the ultra-wideband transceiver performance of a communication system. The proposed study presented is a novel framework of ultra-wideband transceiver architecture that uses a digital code-shifted reference (DCSR) scheme and the security of the UWB transmission which is based on changing the physical properties of transmission and does not rely on higher-level security. Our goal is to find a solution to provide an additional level of security and hardware architecture for impulse radio ultra-wideband (IR-UWB) transmitter and non-coherent receiver systems using pulse modulation and decoding scheme for efficient communication performance. Recently, it has been proposed that physical characteristics of UWB signals can be used to improve the cryptographic security of the system. We consider both coherent and reference-based UWB schemes to enhance security and optimize the implementation for the impulse radio ultra-wideband (IR-UWB) system when an eavesdropper is observing communications over multipath channels between two legitimate partners who share a secret key of a limited length using novel digital code-shifted reference (DCSR). A deep survey on the schemes and security of UWB signals is presented in this research article, and thus, it serves as a ready reckoner for all the researchers who want to pursue their research in this exciting field of communications.

---

R. Santosh Kumar (✉)

Department of Electronics Engineering, School of Engineering, Dayananda Sagar University, Bangalore, Karnataka, India

e-mail: [rsanthosh.kumar665@gmail.com](mailto:rsanthosh.kumar665@gmail.com)

R. Narendra

Department of Electronics & Communication Engineering, School of Engineering, Dayananda Sagar University, Bangalore, India

R. Devaraju

Department of Electronics & Telecommunication Engineering, Dayananda Sagar College of Engineering, Bangalore, India

**Keywords** Ultra-wideband (UWB) · Impulse radio ultra-wideband (IR-UWB) · Digital code-shifted reference (DCSR)

## 1 Introduction

UWB programmes must adhere to the strict rules of the federal communications commission (FCC) which limit ultra-wideband bandwidth, spectral density emissions and data levels to reduce interference. Similarly, UWB systems are shown to use very low power due to the power limits imposed by the FCC. Excessive bandwidth of UWB signals makes UWB transmissions resistant interference rather than small band transfers. In addition, UWB transfers have been widely accepted over the years due to the growing demand for portable devices that provide high data rates with less power [1–4].

Similarly, UWB communication systems have fascinated a lot of attention due to the very low power output, thus avoiding interference by ordinary receivers, as well as the protection of dynamic body layers as a result of their large bandwidth provided [5–8].

UWB signing models with a standard frame structure mean preventing body transfers. In order to convey pieces of symbols, there are many frames in a single symbol that have one touch on each frame. For UWB programmes, performance, acquisition difficulties, power consumption and costs will be considered to determine whether to use parallel or non-compliant adoption. UWB compatible applications require a complex host configuration to balance channel information. Therefore, the complexity of UWB compatible communication systems is often increased to attain robust performance [9–13].

In general, compatible UWB systems are considered superior to non-UWB operating systems. In contrast, UWB incompatible systems can provide a simple host structure by avoiding complex channel measurements from excessive bandwidth. The functionality of UWB compliant and non-compliant systems has been evaluated [14–23].

## 2 Existing Approaches and Review of Literatures

Extensive literature survey on the topic, "Comprehension of ultra-wide band transceiver for wireless communication system based on code shifted reference". In this category, the review work done by other authors is being presented with their advantages and limitations. In this section on literature review, an exhaustive survey [1–23] of the research works achieved by other authors till date is being presented w.r.t. the work taken up in this exciting and application-oriented field of communications and is being presented along with their limitations, advantages, etc.

Similar to the works presented by a large no. of researchers, authors, engineers, scientists, students, etc., in the preceding paragraphs, there was still a large amount of work done by many researchers across the world till date in the field of wireless communications and its applications. But, here, we have considered only the significant ones [1–23], which are being referred by us. The review starts as follows.

Here, we discuss the existing work being carried out towards ultra-wideband (UWB) transceiver design using digital code-shifted reference and its corresponding contribution to securing the communication systems. Ko and Goeckel [1] have presented ultra-wideband transmitted reference (UWB-TR) systems for enhancing the security in wireless physical layer under IEEE 802.15.4a. The UWB-TR systems have best security trade-off than the conventional UWB systems. Nie and Chen [2, 5] have presented the concept digital code-shifted reference (DCSR) transceiver for IR-UWB, which is compared with frequency-shifted reference (FSR) transceiver.

Digital code-shifted reference (CSR) transceiver improves the BER performance and low complexity than frequency-shifted reference (FSR) transceiver. Nie and Chen [3, 19] have presented a performance analysis in terms of low complexity, BER rate of digital code-shifted reference (CSR) UWB radio with respect to transmitted reference (TR) and FSR transceivers. The UWB radio systems implement no delay element and no analog carrier while designing the CSR-based IR-UWB transceiver system. Nguyen and Tran [4] have presented a Simulink model and system generator base HDL model of transmitted-reference UWB receiver on FPGA.

The receiver includes the practical synchronization algorithm to calculate the bits quantization and sampling rate on BER. Nie and Chen [20] have presented to improve the BER rates to reduce the power spent to transmit the reference pulse sequence in CSR-UWB transceiver, and the design includes the differential encoding and decoding to enhance the performance in communication system. Sedaghat and Nasiri Kenari [6] present internally coded time-hopping UWB communication system using CSR to avoid the channel estimation in transceiver systems, the super orthogonal encoder and Viterbi decoder are coding schemes used, and performance results are compared with encoded CSR-based UWB communication system.

Strackx et al. [7] presented the electromagnetic (EM) subtraction technique for Gaussian pulse generation in UWB transmitter, which is highly flexible and rapid prototyping alternative compared with CMOS designs. Olonbayar et al. [8] present the testing of IR-UWB baseband transceiver for IEEE802.15.4a on both FPGA and ASIC implementation, baseband transceiver system includes the synchronization, and data detection performances are robust. Shah et al. [9] present UWB receivers including a performance comparison between transmitted reference (TR), multi-differential frequency shift reference (MD-FSR) and code shift reference (CSR) which are the three primary pulse sending techniques in UWB communications. In [11], the authors worked on the low power embedded high-precision and low latency UWB localization and produced very good significant results which were later used by the team of researchers in [12] and developed an IEE IR-UWB CMOS transceiver for very high data rate, low power and the short-range communications. In [13], the authors worked on the transceivers and produced significant results. They used the differential CSR algorithm to accept the results.

In [5, 20], the researchers worked on the BER performance on the differential code-shifted reference (DCSR) transceiver for impulse radio ultra-wideband (IR-UWB). The correlation results showed that the DCSR transceiver has the better BER performance and comparatively low system complexity, but the authors did not work on the high power aspects as such it was a drawback to the wireless communication nets.

In [14], the authors worked on the various ultra-wide signals in opaque multipath environments and obtained significant results, which was extended by the group of researchers in [15] to the work on delay-hopped transmitted-reference RF communications. Similarly, the authors in [16] researched upon the slightly frequency-shifted reference ultra-wideband (UWB) radios and did some contributory works. In [17], the work on multi-differential slightly frequency-shifted reference ultra-wideband (UWB) radios was developed, which led to a significant revolution in the wireless communication world. Work on code-shifted reference ultra-wideband (UWB) radio was developed in [18]. The performance analysis of code-shifted reference UWB radios was shown in [19]. The work on differential code-shifted reference ultra-wide band (UWB) radios was developed in [20]. Some novel digital communication techniques in the use of signal design and detections for space applications were developed in [21]. In [22], work on ultra-wideband propagation channels—theory, measurement and modelling—was shown. Finally, in [23], the use of wavelet packets in ultra-wideband pulse shape modulation systems was developed.

A number of drawbacks, voids, advantages were there in the works which were carried out by the researchers in our chosen research field. Some of these disadvantages are going to be considered in our research work, and new contributions were produced during the research work, which was well authenticated by simulation results. The research work's problem formulation certified through effective simulation results in the software platform in order to validate the research problem undertaken.

### **3 Limitations of Preceding Methods Developed by Earlier Researchers**

The work done by the various authors was presented in the previous paragraphs [1–23], and there were certain limitations such as consideration of only

- use of conventional methods,
- high compilation time, i.e. computationally very expensive,
- full-fledged automation of algorithms not done,
- less work done on increasing the accuracy and performance,
- real-time implementation (h/w), very few people done, etc.,
- lot of noise was observed, which was affecting the output performance,
- the developed algorithms were not reliable and were prone to noise effects,
- modulating and de-modulating effects were seen leading to high PSNR in sound signal receptions.

## 4 Rectification to the Works Done by Researchers

Some of the above-mentioned limitations which were existing in the works carried out by the past researchers were considered in our research work, new algorithms are proposed in order to overcome some of the insufficiency of the existing algorithms, and also, sincere effort is made to develop some highly efficient algorithms for proposed research work. Once the problem is defined after thorough literature review of the work done by other authors, the problem can be defined, and thus, the various parameters could be analysed while designing the wireless communication system to make it more efficient and reliable.

## 5 Design Challenges

In this section, the design challenges that are faced in designing of the wireless communication system are presented. There is a less research work towards designing hardware-based digital code-shifted reference (DCSR) considering UWB systems. Significant trade-off for UWB architecture considers frequently used reference schemes in wireless communication systems. Few optimization techniques are being implemented towards enhancing the throughput and other communication-related performance in wireless networks. The significant design challenges identified for the proposed study are as follows.

- Transmitting pulses with very short era fix up a critical problem in the transceiver design for the IR-UWB systems.
- At present, there is less investigation towards impact on DCSR-based systems towards the architectural design of UWB transceiver.
- The robustness factor of hardware-based transmitter and receiver designs in UWB transceiver extremely less explored.
- The empirical impact on various security and communication performances towards the existing architectures/models of UWB transceiver is not standardized yet.

## 6 A Objective Formulation

Based on the limitations of the works done by the earlier authors, problem formulation was carried out and the objective of the research was carried out. The prime purpose of the proposed research study was to design and develop a novel hardware-based architecture of ultra-wideband (UWB) transceiver in wireless communication systems. The secondary study objective is to formulate a novel computational model of transmitter and receiver of ultra-wideband system for enhancing the security in physical layer using digital code-shifted reference (DCSR). The complete framework of DCSR in ultra-wideband (UWB) transceiver is finally going to be prototyped using FPGA for effective model validations.

## 7 Time Framework of UWB Signals

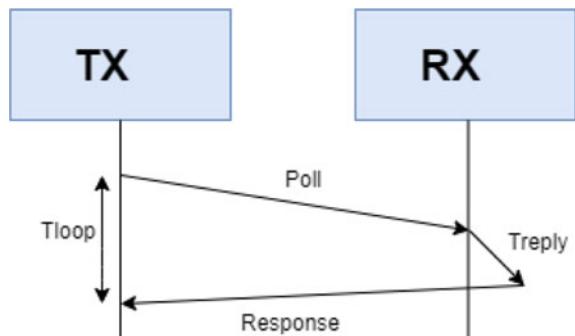
The expected research study's outcome of the proposed system is anticipated to obtain a chip reduction in terms of less area utilization, better throughput, low latency and minimized power consumption. The overall objective of the study will be to obtain an enhanced security in ultra-wideband (UWB) using digital code-shifted reference (DCSR) for wireless communication system.

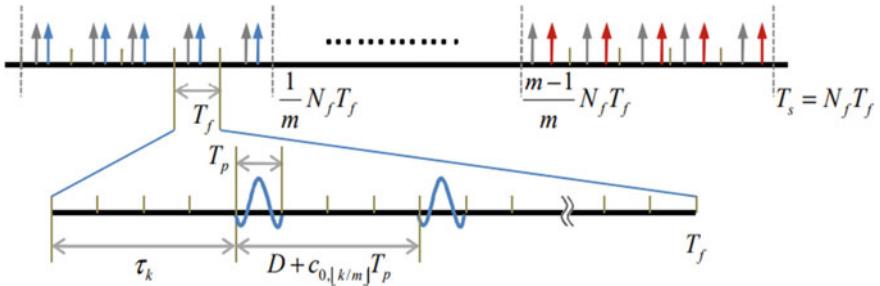
Using the physical properties of ultra-wideband transmissions, the UWB signalling paradigm speeds up transmission. Figure 1 depicts the proposed signalling models, which are based on a time-hopping (TH) technique and binary pulse amplitude modulation. With the help of quantifying waveforms, the techniques for obtaining the UWB Channel. The method of extracting the UWB channel impulse response (CIR) from the quantifying waveforms.  $q(p)$  denotes the template waveform sent by the  $T(p)$  at the  $p$ th location index and the delayed received waveform  $r(p, \tau)$  after propagating through the channel  $h(p, \tau)$  given by

$$r(p, \tau) = q(p) \otimes h(p, \tau) \quad (1)$$

where  $\otimes$  denotes the convolution operator. Concurrent reception of b-bit  $K$  key to set UWB pulses within token time. Contrary to traditional distribution systems, we do not use a registry switch for key-cut connections to produce artificial sound sequence (PN), as this does not improve the cryptographic power of the system [10]. Figure 2 shows the UWB signing system for compatible acquisition. Ideally, each pulse would be obtained independently using key pieces, but the buttons were usually high enough to support that. Therefore, we divide the shared b-bit  $K$  key into parts of  $mK = (K_1, K_2, K_3, \dots, K_m)$  using the specified key fragments, with each  $K_i$  with  $\frac{b}{m}$  bits which  $\in (1, 2, 3, \dots, m)$ ; this is used to select the position indicator at  $0.1\dots, (2^{\frac{b}{m}} - 1)$ , and this is shared pulses on the corresponding frames.

**Fig. 1** UWB signalling





**Fig. 2** UWB signalling scheme intended for TR reception

## 8 Expected Outcome

The research study outcome of the proposed system is anticipated to obtain a chip reduction in terms of less area utilization, better throughput, low latency and minimized power consumption. The overall objective of the study will be to obtain an enhanced security in ultra-wideband (UWB) using digital code-shifted reference (DCSR) for a wireless communication system. The complete DCSR-UWB transceiver is designed, which includes UWB transmitter and receiver designs, and UWB transceiver introduces security in transmission side to enhance the security in wireless communication systems. DCSR-based UWB transceiver architecture is designed to reduce the hardware complexity.

## 9 Methodology of the Proposed Research

The propound research work will be carried out considering analytical research methodology with an aim to design and develop a complete digital code-shifted reference (DCSR) ultra-wideband (UWB) transceiver design. The functional specification can be obtained from the enormous literature available on the UWB transmitter and receiver designs along with reference architectures. Following are the respective stages of operations to be carried out sequentially in order to achieve the proposed research objective.

## 10 Conclusion

The survey on the design and implementation of a compact, secure code-shifted reference ultra-wideband (CSR-UWB) transceiver is presented in this article, and it presents the overview of the work done by various authors till date. Also, a brief

survey is performed on the techniques of ultra-wide band signal and its aftermaths are analysed along with the mathematical models.

Once the problem formulation is carried out and defined, the objectives are being set and outcomes have to be obtained which is just highlighted in this research article. The complete DCSR-UWB transceiver is going to be designed, which includes UWB transmitter and receiver designs; UWB transceiver introduces security on transmission side to enhance the security in wireless communication systems. DCSR-based UWB transceiver architecture is designed to reduce the hardware complexity, comparison of the performance metrics with existing UWB transceiver systems with improvements, improvements in DCSR-UWB architecture which includes chip reduction in terms of less area utilization, better throughput, low latency and minimized power consumption. FPGA prototyping and physical verification of the DCSR-UWB architecture are also thought of in the near future if time permits. The paper conveys the recent developments in the insights of ultra-wideband transceiver for wireless communication system based on code-shifted references.

**Acknowledgements** The authors like to acknowledge the researchers of DSU & DSCE for all sorts of help rendered during the preparation of this article.

## References

1. Ko, M., Goeckel, D.L.: Wireless physical-layer security performance of UWB systems. In: 2010-MILCOM 2010 Military Communications conference, pp. 2143–2148. IEEE (2010)
2. Nie, H., Chen, Z.: Code-shifted reference transceiver for impulse radio ultra-wideband systems. *Phys. Commun.* **2**(4), 274–284 (2009)
3. Nie, H., Chen, Z.: Performance evaluations for differential code-shifted reference ultra-wideband (UWB) radio. In: 2009 IEEE International Conference on Ultra-Wideband, pp. 274–278. IEEE (2009)
4. Nguyen, H.V., Tran, M.H.: Synchronization algorithm and FPGA implementation for transmit-reference UWB receiver. In: 2012 Fourth International Conference on Communications and Electronics (ICCE), pp. 506–511. IEEE (2012)
5. Nie, H., Chen, Z.: Code-shifted reference ultra-wideband (UWB) radio. In: 6th Annual Communication Networks and Services Research Conference (CNSR 2008), pp. 385–389. IEEE (2008)
6. Sedaghat, M.A., Nasiri Kenari, M.: Code-shifted reference for internally coded time hopping UWB communication system. In: 2008 International Symposium on Telecommunications, pp. 214–218. IEEE (2008)
7. Strackx, M., Faes, B., D'Agostino, E., Leroux, P., Reynaert, P.: FPGA based flexible UWB pulse transmitter using EM subtraction. *Electron. Lett.* **49**(19), 1243–1244 (2013)
8. Olonbayar, S., Kreiser, D., Kraemer, R.: FPGA and ASIC implementation and testing of IR-UWB baseband transceiver for IEEE802. 15.4 a. In: 2014 IEEE International Conference on Ultra-WideBand (ICUWB), pp. 456–461. IEEE (2014)
9. Shah, P.M.A., Jan, L., Waqas, M.: Performance comparison of impulse radio ultra wideband receivers. In: 2015 International Conference on Emerging Technologies (ICET), pp. 1–4. IEEE (2015)
10. Hennessey, A., Alimohammad, A.: Design and implementation of a digital secure code-shifted reference UWB transmitter and receiver. *IEEE Trans. Circ. Syst. I: Regular Papers* **64**(7), 1927–1936 (2017)

11. Mayer, P., Magno, M., Schnetzler, C., Benini, L.: EmbedUWB: low power embedded high-precision and low latency UWB localization. In: 2019 IEEE 5th World Forum on Internet of Things (WF-IoT), pp. 519–523. IEEE (2019)
12. Lee, G., Park, J., Jang, J., Jung, T., Kim, T.W.: An IR-UWB CMOS transceiver for high-data-rate, low-power, and short-range communication. *IEEE J. Solid-State Circ.* **54**(8), 2163–2174 (2019)
13. Lowe, J.: RF Transceiver for Code-Shifted Reference Impulse-Radio Ultra-Wideband (CSR IR-UWB) System (2010)
14. Win, M.Z., Scholtz, R.A.: On the energy capture of ultrawide bandwidth signals in dense multipath environments. *IEEE Commun. Lett.* **2**(9), 245–247 (1998)
15. Hoctor, R., Tomlinson, H.: Delay-hopped transmitted-reference RF communications. In 2002 IEEE Conference on Ultra Wideband Systems and Technologies (IEEE Cat. No. 02EX580), pp. 265–269. IEEE (2002)
16. Goeckel, D.L., Zhang, Q.: Slightly frequency-shifted reference ultra-wideband (UWB) radio. *IEEE Trans. Commun.* **55**(3), 508–519 (2007)
17. Zhang, Q., Goeckel, D.L.: Multi-differential slightly frequency-shifted reference ultra-wideband (UWB) radio. In: 2006 40th Annual Conference on Information Sciences and Systems, pp. 615–620. IEEE (2006)
18. Abhishek, M.B., Shet, N.S.V.: Data processing and deploying missing data algorithms to handle missing data in real time data of storage tank: a cyber physical perspective. In: 2019 1st International Conference on Electrical, Control and Instrumentation Engineering (ICECIE), pp. 1–6. IEEE (2019)
19. Nie, H., Chen, Z.: Performance analysis of code-shifted reference UWB radio. In: 2009 IEEE Radio and Wireless Symposium, pp. 396–399. IEEE (2009)
20. Nie, H., Chen, Z.: Differential code-shifted reference ultra-wideband (UWB) radio. In: 2008 IEEE 68th Vehicular Technology Conference, pp. 1–5. IEEE (2008)
21. Simon, M.K., Hinedi, S.M., Lindsey, W.C.: Digital Communication Techniques: Signal Design and Detection, vol. 1. Prentice Hall (1995)
22. Molisch, A.F.: Ultrawideband propagation channels-theory, measurement, and modeling. *IEEE Trans. Vehic. Technol.* **54**(5), 1528–1545 (2005)
23. Ciolino, S., Ghavami, M., Aghvami, H.: On the use of wavelet packets in ultra wideband pulse shape modulation systems. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* **88**(9), 2310–2317 (2005)

# Optimisation of the Execution Time Using Hadoop-Based Parallel Machine Learning on Computing Clusters



B. V. V. Siva Prasad, G. Sucharitha, K. G. S. Venkatesan,  
Tulasi Radhika Patnala, Thejovathi Murari,  
and Santoshachandra Rao Karanam

**Abstract** Due to its growing popularity, the discipline of machine learning has been able to take advantage of the large amount of data that is now available. Whilst data set sizes increase, so does algorithm execution time. Data and processing capacity may be distributed across a large number of computer nodes using cluster computing frameworks, allowing algorithms to run in a reasonable length of time even for very huge data sets. On the basis of MapReduce-K-means (MR-KMeans)-based distributed document clustering, MapReduce PSO-K-means (MR-PKMeans), and a hybrid-distributed document clustering method, three alternative approaches to document clustering are proposed in this work (MR-hybrid). Semantically related document clusters with excellent quality and speed are developed after comprehensive examinations. In order to improve clustering quality and speed up the localised clustering solution, the MapReduce-K-means-distributed (MR-K-means) document clustering approach is implemented in the Hadoop framework using an efficient similarity metric.

---

B. V. V. Siva Prasad  
Malla Reddy University, Hyderabad, India

G. Sucharitha  
Department of ECE, Institute of Aeronautical Engineering, Hyderabad, India

K. G. S. Venkatesan  
Department of CSE, MEGHA Institute of Engineering & Technology for Women, Edulabad, Hyderabad, India

T. R. Patnala (✉)  
Sahanax Technologies, Hyderabad, India  
e-mail: [tulasichandra2010@gmail.com](mailto:tulasichandra2010@gmail.com)

T. Murari  
Computer Science and Engineering, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India

S. R. Karanam  
Department of IT, Anurag University, Hyderabad, Telangana, India

**Keywords** Parallel machine learning · Cluster computing · Hadoop · Performance analysis

## 1 Introduction

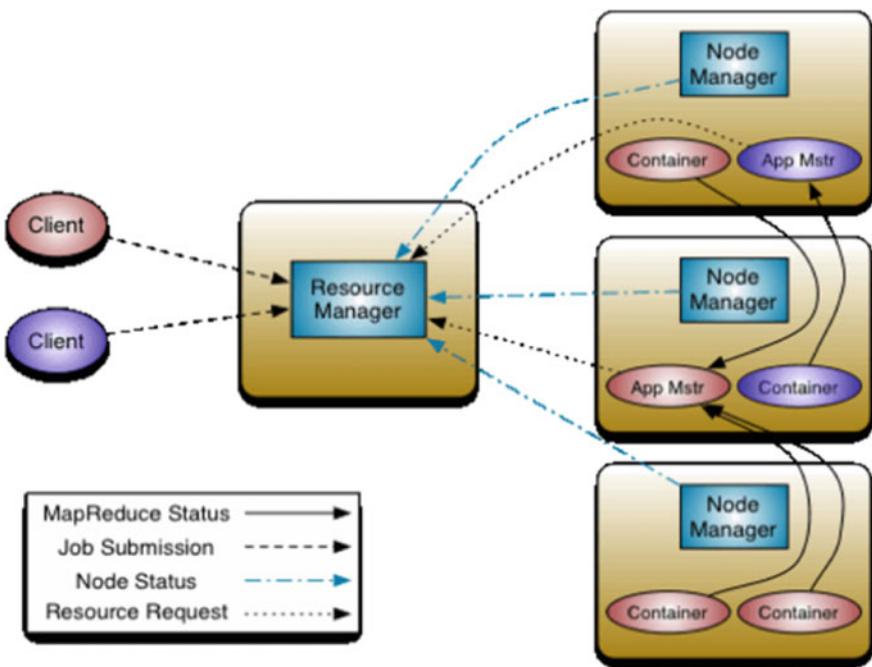
Yarn is the codename for the MapReduce 2.0 component of Apache Hadoop. The existence of yet another resource negotiator in the bargaining process is indicated by the presence of yarn. Furthermore, it includes the Resource Manager and the Scheduler, as well as the Container and the Application Master [1]. The customer assigns the job to the resource management, and the Resource Manager accepts the assignment. To have a job executed, it must be submitted to the Node Manager. A response from the Node Manager will indicate whether the job was successful or unsuccessful, and a response from the Resource Manager will indicate whether the job was successful or unsuccessful to the client. Application Master collaborates with Node Manager to perform and monitor tasks, and with Resource Manager to request resources, as well as communicating back and forth between the two components [2]. The Node Manager communicates with the Resource Manager via heartbeat communication in order to report on its own aliveness. The Resource Manager includes two components: the Scheduler and the Application Master. The Resource Manager is notified by the Application Master that a Resource Request has been received. The scheduler, which serves as a container for resources such as CPU, memory, and so on, is in charge of scheduling jobs and giving them to the Application Master as and when they are required (Fig. 1).

MapReduce is a data processing component in the Apache Hadoop Yarn framework. A computational platform is provided to users. Commodity hardware is used to develop distributed and parallel programming models in order to meet their goals. This package includes utilities such as mapper and reducer. You can think of it in terms of the mapper function, which accepts a key/value pair as an input and then executes the data block based on logic from the programme to create another type of key/value pair. In order to use the reducer function, you must pass it an intermediate key/value pair. The Resource Manager receives jobs from the user and processes them [3]. Job is broken into several block tasks, a map and a reduction task in order to share workload throughout the cluster.

Figure 2 illustrates an illustration of the MapReduce architecture implemented in the Hadoop framework using the framework's building blocks. There are three main processes involved, and they are as follows:

### 1.1 Mapper

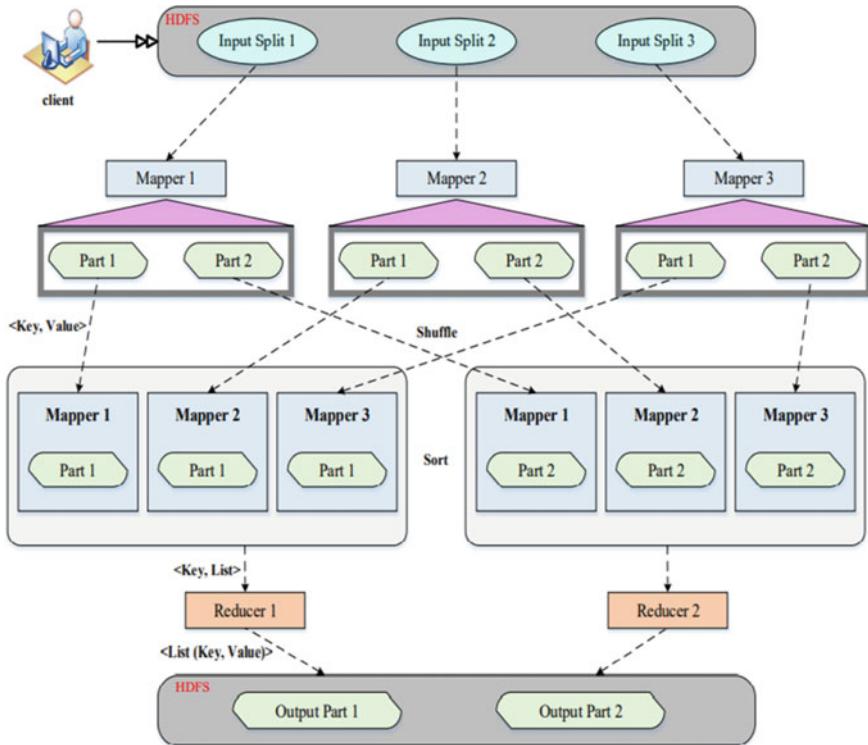
The mapper phase is the initial stage of the MapReduce system, and it is during this stage that the input is broken down into key, and value components are divided. It is



**Fig. 1** Hadoop architecture

possible to write on and swap out the key in this situation because of the way it is handled. A variety of splits are created from the input that has been provided after that. In the environment, there are rational splits, but there are also input splits to be considered [4]. As a result of translating these input splits to this format, key-value pairs are formed in the record reader and stored in the database. It is the Hadoop system's actual input data setup for the mapper input for auxiliary data processing, as opposed to the Hadoop system's actual output data arrangement. Each programme has its own set of requirements for the input format.

As a result, the programmer must be able to comprehend the data that is being received and write code that reflects that comprehension. It is only possible for a mapper to make use of partition and combiner logic to complete a certain data procedure. The words “little reducer” and “combiner” refer to the same device, which is the same size [5]. The high network bandwidth required by Hadoop for managing big amounts of data is essential. This issue is handled by including the combiner stage after the mapper step has been completed, as previously stated. The partition module of the Hadoop framework is crucial in the process of separating data collected from different mappers or combiners.



**Fig. 2** Map reduce architecture

## 1.2 Shuffling and Sorting

The shuffling and sorting process is an intermediary step in the Hadoop system that is required to complete the MapReduce operation. After completing the mapper process, there will be a large quantity of middle data to be moved from all of the Map nodes to the shuffler in order to complete the process. It sorts the key of the given input, and as a result, the entire collection of pairs with the same key value is brought together [6]. In addition, it is necessary to transport the sorted output to the reducer nodes in order to continue the MapReduce operation on those nodes.

## 1.3 Reducer

When using MapReduce, the reducer is the last step in the process. In the reducer operation, the transitional key and the set of values linked to the provided key are obtained. Reducer: A smaller collection of values is created by condensing the input

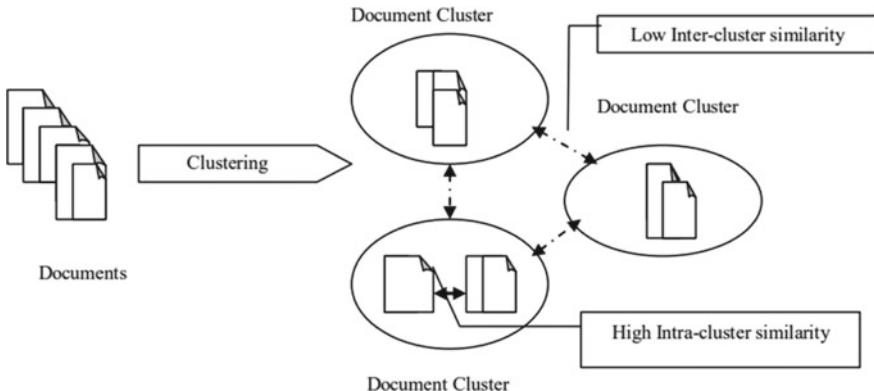
data and then using those smaller sets of values to create the output. Reducers write data to Hadoop using a record writer module that is part of the reducer.

The MapReduce framework has a number of implementations, including Mars, Phoenix, Hadoop, and Google's own. Hadoop is the most extensively utilised software in these situations due to its open-source nature. The most extensively used MapReduce implementation is the Hadoop framework, which allows applications to run on enormous computer clusters [7]. By splitting down a huge task into smaller tasks and a large data set into smaller divisions, the Hadoop framework enables distributed, data-intensive, and related applications. This means that each job processes a distinct partition in parallel. With the Hadoop framework, huge data sets can be processed in a distributed fashion by using a collection of computers that have been programmed according to specified programming paradigms and models. It can quickly grow from a single server to tens of thousands of nodes. It is designed to estimate application failures rather than relying on hardware in order to give high accessibility.

## 2 Background

The clustering or unsupervised learning discipline of machine learning is crucial in extracting or summarising new information by grouping data based on similarities [8], which is critical in extracting or summarising new information. It is used in a variety of applications such as statistics, pattern identification and data mining. The purpose of this study is to investigate how clustering can be employed in data mining. The clustering technique used in data mining is crucial in tackling the particular difficulties faced by every sector of the information technology industry. Clustering very large data sets with multiple dimensions presents a number of obstacles and expenses [9]. Amongst these are the following, data It is the process of categorising a collection of items in such a way that items in one cluster are similar to one another and distinct from those in other clusters [10]. Clustalizing is a data mining approach that makes it feasible to abstract large amounts of data by grouping things that are related together in a single data set. The similarity metric is used to identify which objects should be grouped together.

To make decisions using data clustering, we must specify a measure of similarity or distance over the object feature space. This measure must be specified before we can make decisions. When a piece of data is fed into an automatic categorization system, the system can figure out which category it belongs to based on a set of predefined categories. Clustering methods, on the other hand, let the computer decide how to divide a data set [11]. Categorization is useful when new data need to be classified into an existing category. Clustering is an effective method for uncovering previously unknown structural details. Both classification and clustering can yield visually appealing results from an unknown data set; classification organises these data sets according to a well-known structure, whilst clustering illustrates the



**Fig. 3** Process of document clustering

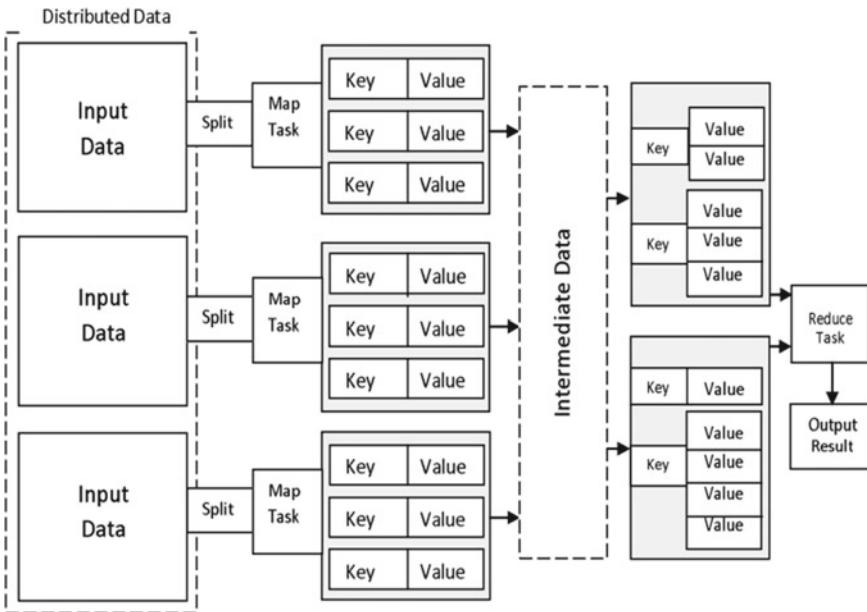
structure of the particular data set in issue. This works aims to boost the efficiency of document data set clustering (Fig. 3).

A new data mining study topic known as document clustering has gained significance as a result of significant developments in information retrieval and text mining. The third graphic depicts document clustering's end purpose [12]. Document clustering is the principal approach of data organisation since it maximises intra-cluster similarity whilst minimising inter-cluster similarity.

### 3 Hadoop and MapReduce for Distributed Data Clustering

MapReduce is a distributed programming approach for handling large amounts of data. Cloud computing is now possible as a result of the development of this enabling technology. When processing data over numerous workstations, the MapReduce functional programming method is a helpful tool to have on hand. This programme does two primary operations: a map and a reduction, which are described below. Each document set is automatically parallelised and distributed [13] so that it can manage large document sets without requiring manual intervention. Nodes in the system communicate with one another through a network of communication links. Map operations are capable of efficiently handling input data that is distributed across multiple nodes. It is possible to integrate variables by lowering the number of variables. In a logical sense, all document data are represented as a key (K) and value (V) pair. You can use many mappers and reducers at the same time if you want to save time. Each map operation takes a K1 and V1 pair as input and outputs an intermediate list of K2 and V2 pairings as output [14]. In order to construct a new key-value pair, the intermediate list of key-value pairs is sorted and grouped together.

A reduction task processes each key-value combination separately to generate the final output pairs. “To sort intermediate key-value pairs, use a reduce task. Reducing

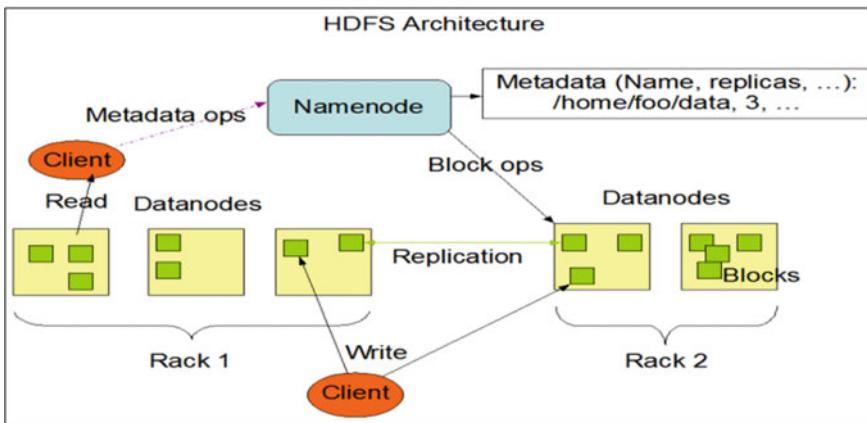


**Fig. 4** Distributed MapReduce function with clustering

activities cannot begin until their corresponding map counterparts have finished, but regardless of when they are started, all map and reduce operations run in parallel and independently of one another.” Each map function is executed in parallel and, upon completion, outputs data from the corresponding input [15]. Similarly, each key reducer works independently and concurrently (Fig. 4).

In a cluster, HDFS is a distributed file system that works well with large files. Rather than saving files one at a time, it saves them in blocks. Fault tolerance is built into the system by duplicating all of the blocks [16]. In Bitcoin, the name node is responsible for replicating block data. Each data node sends this server a heartbeat and a block report on a regular basis.

As you can see in Fig. 5, HDFS is designed in this way [17]. The HDFS interface allows the processing to be shifted to the data rather than the data to the application, which is useful for working with enormous data sets like this. As a result, the computational algorithm runs more smoothly and efficiently. Utilising distributed document clustering techniques built on Hadoop-MapReduce can help solve the problems associated with distributed document storage. Customers don’t have to worry about distributed programming issues because of the MapReduce model’s simplicity. Consumers only have to think about the computing element of the algorithm [18]. By utilising the MapReduce-distributed document clustering technique, existing document clustering algorithms will run much more quickly.

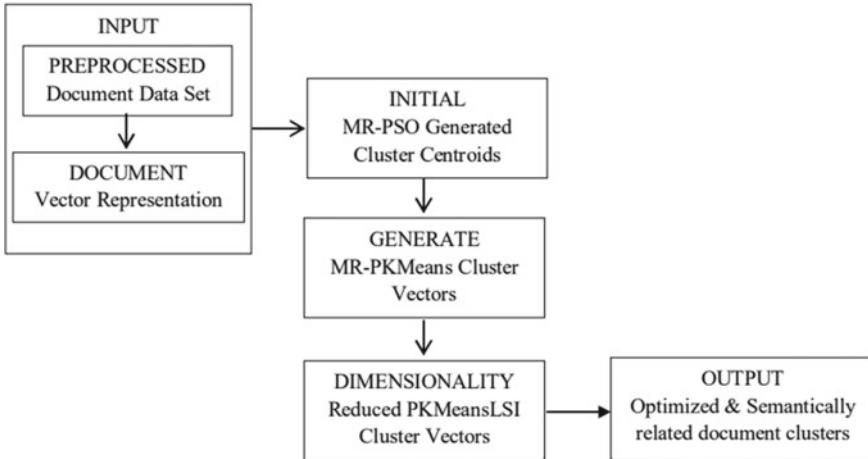


**Fig. 5** HDFS architecture

## 4 Methodology

The algorithm proposed is a hybrid PSO-K-means LSI (PK-meansLSI) algorithm based on MapReduce that combines the ability to cluster huge texts with the global search power of particle swarm optimization (PSO) [19]. It is hoped that by using this methodology, the hybrid algorithm will be able to perform better in clusters and take advantage of distributed frameworks such as the Hadoop-based MapReduce framework. For the purpose of determining the optimal cluster centroids for k-means clustering, this proposed methodology employs a particle swarm optimization (PSO) module that runs in parallel with a latent semantic indexing (LSI) dimensionality reduction module to reduce the number of possible cluster centroids (Fig. 6).

- Make a decision on which document data set you would like to examine. In order to conduct text mining research, a wide range of publically available document databases can be used. Examples of actual document data sets are provided in the following sections.
- Document preprocessing is used to minimise the amount of attributes in a document. There are various processing phases that must be completed before the real processing may begin. It is necessary to convert the input text documents into a set of terms in order to use the vector space model. Documents are represented as document term matrixes (DTMs) in this vector space model, which is achieved by determining the word weight (VSM). When it comes to document term matrixes, there is an entry for term weight (DTM).
- Initial seeds for the MR-PK-means module are derived from these term vectors, with the optimal centroids for the MR-PK-means module derived from the MR-PSO module. Finding the most optimal document clustering configurations is accomplished through the use of PSO-K-means document clustering based on distributed MapReduce (MR-PK-means).



**Fig. 6** Architecture of proposed method

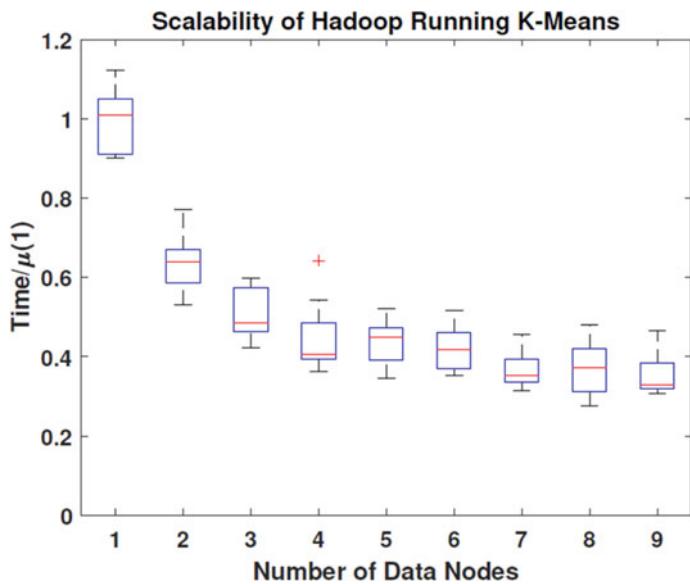
- Input is provided to the MR-hybrid module in the form of document vectors. To reduce the dimensionality of document clusters generated, latent semantic indexing (LSI) is used. This dimension reduction technique results in increased speed and higher quality document clusters that are semantically related.

## 5 Results

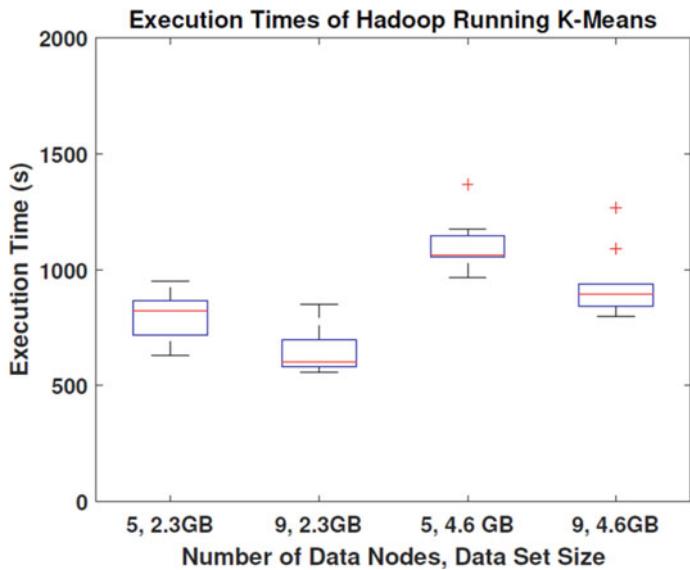
A total of 120 measurements were taken on a total of 20 million rows of synthetic data in the original experiment for k-means, as detailed in the preceding section. Figure 7 shows the outcomes of the experiment. The median is shown by the red line in the box plots once more.

To make a point of comparison, Fig. 8 displays the execution times of the identical configuration when using the original data set. To construct the data set, 40-dimensional data point vectors with each member representing a random integer in the range [1; 100] were randomly generated and then randomly sampled. There are no clusters to be found in this particular data set. When compared to data sets that contain some sort of clustering, it is likely that the algorithm will have a more difficult time achieving convergence before the given maximum number of iterations, which was set to 10. As a result, the data set that has been generated can be viewed as a worst-case scenario of sorts. Using a sample script that was included in Mahout 0.13, the k-means algorithm was run to determine its results. Some execution stages that were not of interest were included in the measured execution time because they were necessary. The data transformation process is the first step.

Whilst text files can be used to store input data in HDFS, Mahout requires sequence files, which are collections of binary key-value pairs, as opposed to standard input



**Fig. 7** Scalabilities of the k-means experiments with the 20 million row data set as the cluster size changes



**Fig. 8** Execution times of the k-means experiments with the 40 million row data set as the cluster size changes. Included also are the execution times for the same node configuration for the original experiment

files. Data transformation can be done ahead of time, but because the example script will do it, it must be included in the time measurement. A text file will most usually be used as the input for comparing MapReduce algorithms to other algorithms, and the time required will only start when the text file is read in. For comparison's sake, the input will almost certainly be text files, and the time taken can only fairly begin at that moment rather than after a transformation; thus this limitation is sensible. Whilst the cluster dump is an unnecessary step in the execution process, it is included in the script because it cannot be avoided. A list of all the data points and their cluster assignments may be found in this dump. Normally, this would be done into a file on the local file system; however instead of a file, the dump is done directly into the command-line window with this script. Because of this, the Java heap becomes overflowing, and as a result, the execution is stopped.

## 6 Conclusion

The MapReduce framework performs better when employing the Map-Optimize-Reduce approach. Include the new optimizer block between the mapper and reducer to decrease data items transferred between the two. The execution times of a sequential k-means algorithm were recorded on one of the worker nodes. Sequential technique outperformed the distributed algorithm's four-node configuration. This demonstrates the usefulness of Hadoop. There are, however, a few important things to keep in mind. This sequential algorithm's execution duration is greatly influenced by the amount of data being loaded from disc. The data could be loaded more quickly if an alternative library is used. When loading the data, only, one CPU core was accessible; nevertheless, all cores were required to run the k-means algorithm.

Instead of utilising Hadoop, the Spark framework will be investigated in future iterations of this project. Due to Spark's improved performance over Hadoop, large data sets' execution time, response time and data load all go smaller.

## References

1. Xian, G.: Parallel machine learning algorithm using fine-grained-mode spark on a mesos big data cloud computing software framework for mobile robotic intelligent fault recognition. *IEEE Access* **8**, 131885–131900 (2020). <https://doi.org/10.1109/ACCESS.2020.3007499>
2. Dhiman, G., Oliva, D., Kaur, A., Krishna, K.K., Vimal, S., Ashutosh, S., Korhan, C.: BEPO: a novel binary emperor penguin optimizer for automatic feature selection. *Knowl. Based Syst.* **211**, 106560 (2020)
3. Srinivasulu, A., Ramanjaneyulu, K., Neelaveni, R., et al.: Advanced lung cancer prediction based on blockchain material using extended CNN. *Appl. Nanosci.* (2021)
4. Rajan, M.S., Dilip, G., Kannan, N., et al.: Diagnosis of fault node in wireless sensor networks using adaptive neuro-fuzzy inference system. *Appl. Nanosci.* (2021)
5. Sangeetha, Y., Majji, S., Srinagesh, A., et al. Authentication of symmetric cryptosystem using anti-aging controller-based true random number generator. *Appl. Nanosci.* (2021)

6. Kothapalli, S., Samson, M., Majji, S., Patnala, T.R., Karanam, S.R., Pasumarthi, C.S.: Comparative experimental analysis of different Op-amps using 180 nm CMOS technology. In: 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), pp. 1–6 (2020). <https://doi.org/10.1109/ic-ETITE47903.2020.440>
7. Jana, B., Thotakura, H., Balyan, A. et al.: Pixel density based trimmed median filter for removal of noise from surface image. *Appl. Nanosci.* (2021)
8. Patnala, T.R., Jayanthi, D., ShyLu, D.S., Kavitha, K., Chowdary, P.: Maximal length test pattern generation for the cryptography applications. In: Materialstoday Proceedings (In press). <https://www.sciencedirect.com/science/article/pii/S2214785320305368>. Available online from 20 Feb 2020
9. Patnala, T.R., Jayanthi, D., Majji, S., Valletti, M., Kothapalli, S., Karanam, S.C.R.: Modernistic way for KEY generation for highly secure data transfer in ASIC design flow. <https://ieeexplore.ieee.org/document/9074200>. Published in IEEE digital Xplore, Electronic ISSN: 2575-7288. Available from 23 Apr 2020
10. Haoxiang, W., Smys, S.: Big data analysis and perturbation using data mining algorithm. *J. Soft Comput. Paradigm (JSCP)* **3**(01), 19–28
11. Sivaganesan, D.: A data driven trust mechanism based on blockchain in IoT sensor networks for detection and mitigation of attacks. *J. Trends Comput. Sci. Smart Technol. (TCSST)* **3**(01), 59–69 (2021)
12. Lessanibahri, S., Gastaldi, L., Fernández, C.G.: A novel pruning algorithm for mining long and maximum length frequent itemsets. *Expert Syst. Appl.*, 1–21, Article number 113004 (2020)
13. Yulong, Z., Weiting, L.: A research on battlefield situation analysis and decision-making modeling based on a Hadoop framework. In: 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), pp. 388–391 (2020). <https://doi.org/10.1109/MLBDBI51377.2020.00083>
14. Quoc, P.H., Küng, J.: FPO tree and DP3 algorithm for distributed parallel frequent itemsets mining. *Expert Syst. Appl.* **140**, 1–13, Article number 112874 (2020)
15. Titarenko, S.S., Titarenko, V., Aivaliotis, G., Palczewski, J.: Fast implementation of pattern mining algorithms with time stamp uncertainties and temporal constraints. *J. Big Data* **6**, 1–34, Article number 37 (2019)
16. Du, S., Li, J.: Parallel processing of improved KNN text classification algorithm based on Hadoop. In: 2019 7th International Conference on Information, Communication and Networks (ICICN), pp. 167–170 (2019). <https://doi.org/10.1109/ICICN.2019.8834973>
17. Huang, S.-y., Zhang, B.: Research on improved k-means clustering algorithm based on Hadoop platform. In: 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), pp. 301–303 (2019). <https://doi.org/10.1109/MLBDBI48998.2019.00067>
18. Yildirim, I., Celik, M.: An efficient tree based algorithm for mining high average-utility itemset. *IEEE Access* **7**, 144245–214263 (2019)
19. Alexandropoulos, S.A.N., Kotsiantis, S.B., Vrahatis, M.N.: Data preprocessing in predictive data mining. *Knowl. Eng. Rev.* **34**, 1–33 (2019)

# Medical Image Analysis Using Deep Learning Algorithm Convolutional Neural Networks



D. Raghu and Hrudaya Kumar Tripathy

**Abstract** Medical image analysis using deep learning algorithm (CNN) area is of foremost significance and maybe it is anything but a high need area. A great deal of concern is as yet needed in this area. To be sure, in larger part of cases, the medical information is deciphered by human master, whilst examination of medical information is very demanding and muddled errand. Regularly, discrete examination is performed by various human specialists which brings about mistaken identification of illness. The astounding exhibition of deep learning (DL) in various domains pulled in the specialists to apply this method inside the domain of medical area as DL gives incredible exactness and precision in conclusive yield. Hence, it has been imagined as a centre strategy for medical image analysis using deep learning algorithm (CNN) and in different floods of the medical services area. Further, division measure is the basic, viable and centre advance of the medical image investigation. Scientists are reliably endeavouring to increase the precision of medical image examination. In ongoing past, machine knowledge-based strategies have generally been utilised for this pursuit. The new pattern in the domain of medical image examination is the ramifications of profound learning-based methodologies. Maybe, the use of profound learning improves the prescient correctness of the individual. Additionally, it likewise mitigates the intercession of human specialists in the analysis marvel. This paper aims to survey on medical image analysis using deep learning algorithm convolutional neural network (CNN). The use of deep learning algorithms for medical image analysis is very significant since it is producing enough and reliable results comparing to human tasks; it also reduces human work and time; basing on all these aspects, a survey is about to done. In this paper, besides medical image analysis, convolutional neural network (CNN) architectural implementation and its features are also discussed.

---

D. Raghu (✉) · H. K. Tripathy

School of Computer Engineering, Kalinga Institute of Industrial Technology, Deemed to be

University, Bhubaneswar, Odisha, India

e-mail: [raghu.dhumpati@gmail.com](mailto:raghu.dhumpati@gmail.com)

H. K. Tripathy

e-mail: [hktripathyfcs@kiit.ac.in](mailto:hktripathyfcs@kiit.ac.in)

**Keywords** Image analysis · Convolutional neural networks · Classification · Segmentation

## 1 Introduction

Deep learning is used in a neural network as an automatic tool for studying features [1]. This is the opposite of the traditional handmade methods [2]. This is the same. It is a difficult task to select and calculate these properties [3]. Deep coevolutionary networks are used for the analysis of medical imagery amongst deep learning techniques [4]. Science in clinical practise, used to analyse clinical problems, is called analytical images [5]. The objective is to extract data for enhanced clinical diagnosis in an affective and effective manner [6]. Recent developments make image analysis one of biomedical engineering's leading research and development sectors [7]. One of the reasons for this progress is the use of techniques for machine learning to analyse health images [8]. Deep learning becomes an important aspect in the area of medical image analysis as a learning machine and a tool for pattern recognition [9]. Medical imagery was a long-standing diagnostic method in clinical practise [10]. The field of health images in hardware design [11], security procedures, computer resources and storage of data has been greatly benefited by recent developments [12]. At the moment, segmenting, classifying and detection of abnormality using images generated in a broad range of clinical imaging modalities are the main applications in analysing medical images [13]. The analysis on medical images seeks to support radiologists and clinicians in diagnostic and treatment processes [14]. Due to the direct effects of the clinical diagnosis and therapy process [15], computer-aided design (CADx and CAD) reliance on an effective medical image analysis is vital for performance [16]. Therefore, accuracy and precision in recall and sensitivity, such as F-measurement, are key aspects, and high values in the analysis of the medical image are highly desirable [17]. Due to the increasing availability of digital images dealing with clinical data [18], the best method is needed for big data analysis [19]. The state-of-the-art in areas such as computer vision in data centres shows that the best candidate can be in-depth learning methods [20]. Deep learning imitates a profound architecture consisting of several layers of human brain transformation [21]. The way information in the human brain is processed is similar [22]. A good understanding of the underlying nature of data collection is necessary for the extraction of the most relevant features [23]. A large collection of data could be tedious and difficult to efficiently manage [24]. Their ability to learn the complicated features of raw data is a major benefit for the use of deep study methods [25]. This enables us instead of producing functions to define a system that is primarily necessary to other machine learning technologies [26]. These properties explored the advantages of deep learning in analysing medical images [27]. The future of medical applications can be helped by recent developments in deep learning techniques [28]. There are many open source DL platforms available to mention just a few, like Caffe, TensorFlow, Theano, Keras and Torch [29]. Due to the limited clinical expertise available

for the DL experts and limited DL expertise, the challenges arise [30]. In a recent tutorial, DL's application to digital pathology images is being overcome by providing step-by-step details [31].

Objectives of this paper using CNN include:

1. Disease classification.
2. Image segmentation.
3. Detection of abnormality.
4. Diagnosis of disease.

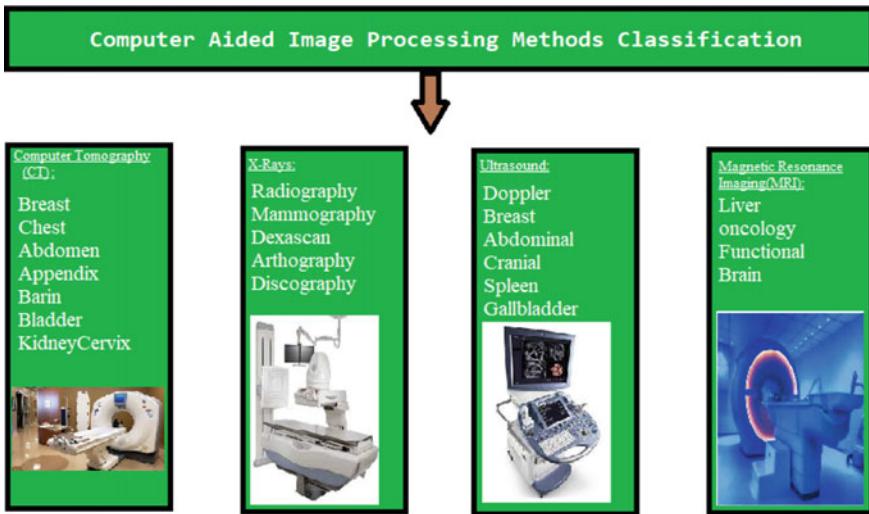
The clinic included mainly experts in the interpretation of medical imaging from people such as radiologists and doctors. Nevertheless, scientists and physicians have recently begun to benefit from computer assistance because of major disease changes and human experts' potential fatigue. Although the analysis of computed images was tardy compared with the progress of image technology, recent improvements were made using machine learning techniques.

In the area of medical imaging, computerised image analysis has been an ongoing problem. Recent progress in machine education, in particular in the field of in-depth learning, has made a major breakthrough in imaging to help identify, categorise and quantify medical image models. The use of hierarchical representations derived solely from data is, in particular, at the forefront of development rather than crafted features based largely on domain understanding. This makes profound learning quickly the cutting edge for improving performance in various medical applications. In this article, we discuss the basic techniques for profound learning, their success in recording image, anatomical and cellular detection structures, tissue segmentation, computer-aided condition diagnosis or prognosis and so on. Finally, we raise research issues and propose future improvements. The rest of the paper is organised as follows.

Medical Image processing Analysis, Convolutional Neural Network (CNN), Data set Generation, Conclusion and Future Scope.

## 2 Medical Image Processing Analysis

As projected in Fig. 1, medical image processing would be classified in different kinds of categories such as ultrasound, MRI and CT scan. Machine learning is an important part of their success in the implementation of target tasks for data analysis. It is Important representation of function. Humans have traditionally designed meaningful or job-related features based on their expertise in specific target areas, making machine teaching techniques difficult for non-experts to use for their respective studies. But, integrating the functional technology in a learning step removes deep learning obstacles. Deep education therefore requires just a few data, if necessary small pre-processing, and does not manually extract features and finds informative representations automatically. Feature engineering moved from human to computer side so that non-experts can efficiently use detailed learning, particularly for the



**Fig. 1** Classification of medical image processing methods

analysis of medical images. For example, pre-screening, diagnosis, treatments of CT, MR, positron emissary tomography (PET), mammography, an ultrasound, X-rays have been shown in recent decades to be of importance for medical image. Some of the medical image processing methods have discussed below.

## 2.1 *Image Segmentation*

The segmentation of images is referred to as the identification of areas in various classes. Automatic image division is now one of the trendiest research fields. A new model is also discovered to segment the picture better for the computer viewing task. The segmentation of images is an important element in the medical field. In the microscopic image of human blood, for example, we can consider the identification of cancer cells. If you want to identify cancer cells, the shape of blood cells should be recognised, and the presence of cancer cells should be diagnosed for the unusual growth of blood cells. This leads to early recognition of blood cancer, so it can be cured in due course. Now, in the field of research picture segmentation, this method is the latest technology. It works on 3D pictures, e.g. height, width and number of channels. First and third dimensions are the channel numbers (RGB) and red, green and blue intensity values. The image resolution in images inserted in the neural network will generally be reduced to actions that reduce the processing time and avoid insertion problems. Even if you take a 1-dimension image  $224 * 224 * 3$ , it becomes an input vector of 150,528. Therefore, this vector is too large to be supplied to the neural network.

## 2.2 *Image Classification*

The CNN is a class of deep and learnable neural networks. In image identification, CNNs are a major step forward. They are often used to classify images behind the scenes in visual images. From Facebook photo to automotive driving, they are at the very heart of anything. In every aspect of health care and safety, they work hard behind the scenes. Image classification is the process of entering (like an image) and exiting a class (like “cat”) or the likelihood that the insert comes from a particular class (“the probability of this insert being a cat is 90%”). You can look at a picture and see your own face in a dreadful shot, but how can a computer learn it? The answer is with CNN. Because of their high accuracy, CNNs are used to classify and recognise images. The CNN is a hierarchical model, which runs in a network such as a funnel and produces a fully linked layer that connects all neurons and processes the output.

### Detection of Abnormality

The detection of medical images abnormalities means that a certain kind of illness such as tumour is identified. Clinical professionals usually identify abnormalities but take a long time and effort. Therefore, it is increasingly important to develop automated systems for the detection of errors. Different approaches for detection of anomalies in medical images are presented in the literature.

### Medical Image Retrieval

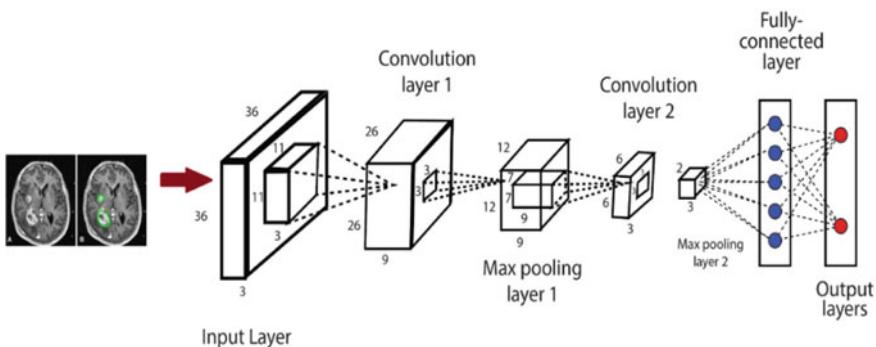
With the widespread use of digital image data in hospitals, the size of medical image repository is rapidly increasing. Over the past thirty years, the manufacture of images by means of more images, more resolution and new types of images have increased exponentially. Medical imaging is one of the biggest data manufacturers in the world. In addition to some images used for publication or teaching, most images are only made for a patient and a single point of time. Data are usually distributed across many institutions and cannot even be combined for the treatment of a single patient. A great deal of expertise is maintained in these medical archives, and from the very beginning, medical, clinical and content-based information are provided through visual and textual information and structured information. When it comes to feature representations that fully characterise high-level information, the effectiveness of these systems is of greater importance. Classification of medical images in convolution neural network is done with specific classes and some modes, an intermodal data set is trained in the network. Medical photographs obtain the learned characteristics and classification results. The best results can be obtained for recovery by using class predictions. Some of the related work is shown in Table 1.

**Table 1** Various medical image retrieval methods and its accuracy with CNN

Application	Method	Framework	Accuracy	Data set
Medical image retrieval	Convolutional neural network and supervised hashing [32]	New content-based medical image retrieval (CBMIR)	98%	Early lung cancer action programme (ELCAP) and the vision and image analysis (VIA) research groups
Medical image retrieval	Using deep convolutional neural network [33]	Content-based medical image retrieval (CBMIR) systems	Classification accuracy of 99.77%, mean average precision of 0.69	Intermodal data set that contains twenty-four classes and five modalities
Medical image retrieval	Medical image retrieval using ResNet-18 [34]	Content-based image retrieval framework (CBIR)	92% and the mean average precision of 0.90 for retrieval	A multi-modality data set that contains twenty-three classes and four modalities
Medical image retrieval	A sequential search-space shrinking using CNN transfer learning and a radon projection pool [18]	Two-step hierarchical shrinking search space when local binary patterns	90.30%	IRMA data set,
Medical image retrieval	Stacked auto-encoder-based tagging with deep features [35]	Content-based image retrieval framework (CBIR)	–	IRMA data set,
Medical image retrieval	Effective diagnosis and treatment through content-based medical image retrieval (CBMIR) by using artificial intelligence [36]	Content-based image retrieval framework (CBIR)	81.51% and 82.42%	12 databases including 50 classes

### 3 Convolutional Neural Network CNN

The weight determines the behaviour of each neuron. The artificial neurons of a CNN choose different visual characteristics when provided with pixel values. Each layer generates a series of activation maps when you enter an image into a ConvNet. The image characteristics are emphasised in the activation maps. Each neuron takes out the input pixels, multiplied by colour values and is resumed and activated by their



**Fig. 2** CNN block diagram

weights. In the first (or lower) CNN layer basic functions, like horizontal, vertical and diagonal borders are usually detected.

In the second layer, the output of the first layer extracts more complicated elements such as corners and edge mixtures. These layers detect higher levels of properties like objects, faces and more when you go deeper into the neural network. “Convolution” is the function of weight multiplication and summation of pixel values (hence the name convolutional neural network). A CNN usually consists of several convolution layers but also contains supplementary components. A classification layer is the final CNN layer that recalls that highly convolution layers detect complex objects as its input, the output of the final convolution layer as shown in Figs. 2 and 3.

### Input Layer

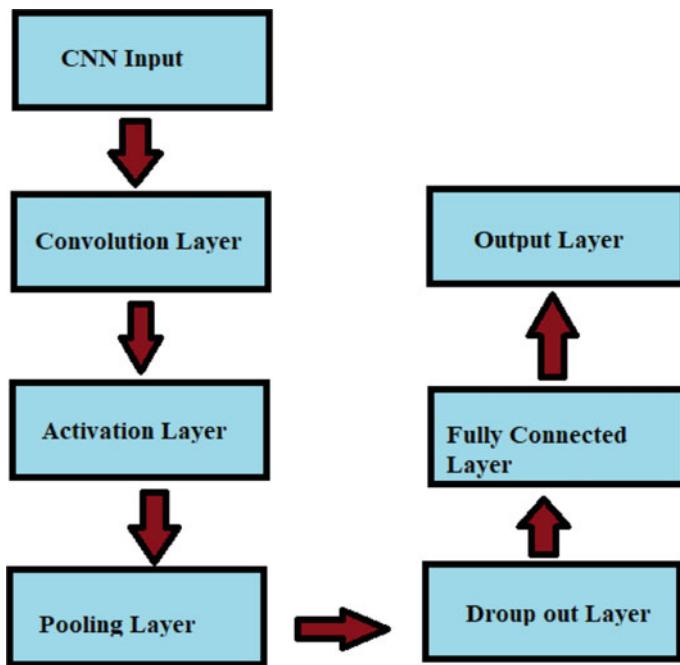
Image data should be included in the input layer of the CNN. The three-dimensional matrix represents picture data as we previously saw. It needs to be transformed into one column. In case of a  $28 \times 28 = 784$  dimensions, the image must be converted to  $784 \times 1$ . When you have “ $m$ ” examples, the input dimension is  $(784, m)$ .

### Convolutional Layer

Convo layer is sometimes called the extractor layer because the layer removed the image functionality. Part of an image is associated with convo layer to perform the convolution as we have seen previously and calculate the point between the receiving field and the filter. This causes the output volume to be integrated. Then, we dive into the following receptive field and perform the same filter operation for the same image. Until the whole image is completed, we repeat the same process. The output is the next layer’s input.

### Max Pooling Layer

The pooling layer is used to decrease volumes after the image input has been converted. It is used in two layers of confusion. It is computer price, and we don’t want it if we use FC after convo layer without bundling or max bundling. The maximum pooling can therefore only reduce the input image volume. In the above example,



**Fig. 3** CNN layers

Stride 2 was used to pool maximally at one depth. The size of the inputs is  $4 \times 4$  to  $2 \times 2$ .

### Fully Connected Layer

The fully connected layer is weighing, damage and neurons. In one layer, the neurons are connected in a different layer to the neurons. It is used to classify images in different categories by training. Softmax or logistic is the last CNN layer. The layer is in the FC at the end of the layer. In binary and multi-classification, softmax logistics are used.

### CNN for Medical Image Analysis

The various modes for medical imaging are used, and images most of the time are identical when it comes to clinical prognosis and diagnosis. The architecture of the network is carefully studied to find complicated information. Where there is expertise and stringent assumptions, handmade features work. These assumptions cannot help with certain tasks, including medical images. It is therefore not easy to distinguish between a healthy and a skilful image in certain applications. For a classifier like SVM, there is no solution. The techniques extracted, regardless of the work or objective function concerned, are features, such as the SIFT (invariant transforming function). Samples are displayed using a word bag vector or any other system. Samples are then taken. This rating, like the use of SVM, differs from the

**Table 2** Various types of CNNs

S. No.	Type of CNN	Year
1	LeNet	1998
2	AlexNet	2012
3	VGG Net	2012
4	GoogleLeNet	2014
5	ResNet	2015
6	ZFNet	2013

other by no loss mechanism to increase the local characteristics of the extraction and classification process.

However, a DCNN can learn the features from the underlying data. The data-driven features are taught at the end of the course. DCNN has the strength of depicting the exhalation part (CNN filters studied from the initial layers) of the function better by using the error signal for the loss function. The other advantage in the initial layers is that neurons concentrate more in high layers on various areas of human organs, and that in the last layers, some neurons take whole organs into account. Figure 3 shows a medical image classification CNN architecture accepting N classes 32/32 patches in 2D medical images. The network has been fully connected with max connection layers. Each super lay produces a characteristic map of different sizes, which reduces the overlay layers to the next level. The fully connected power layers provide the prediction of the required class. The number of parameters is dependent on how many neurons and layers are present on each layer and whether the neurons are connected to the network. The training phase of the network offers the best and most efficient solutions to this problem. The researchers have developed computer technology based on a better understanding of the field of analysis of the medical image. Recent studies have shown that the diagnosis and categorisation of diseases and the medical image success have been based on in-depth algorithms.

Some of the various types of CNNs are listed in Table 2.

## 4 Data set Generation

Data collection is an information collection. The data set will match one or more database tables for tabular data, where each column of a table represents the variable and each row will match the record set. Data gathering is collection of databases. The data set will match one or more database tables for tabular data, where each column of a table represents the variable and each row will match the record set. For each part of the series, the data set shows values such as object height and weight for each of those variables. It is known as a date for every value. It is also possible to collect documents or files as data sets.

Various data set sources for medical image processing would be shown in Table 4.

**Table 4** Various data set sources for medical image processing

Sno	Data set	Link
1	NIH Database of 100,000 Chest X-Rays	<a href="https://nihcc.app.box.com/v/ChestXray-NIHCC">https://nihcc.app.box.com/v/ChestXray-NIHCC</a>
2	The Cancer Imaging Archive (TCIA)	<a href="https://www.cancerimagingarchive.net/">https://www.cancerimagingarchive.net/</a>
3	National Biomedical Imaging Archive (NBIA)	<a href="https://imaging.nci.nih.gov/ncia/login.jsp">https://imaging.nci.nih.gov/ncia/login.jsp</a>
4	Lung Image Database Consortium (LIDC)	Open Source
5	Reference Image Database to Evaluate Response (RIDER)	Open Source
6	Breast MRI	Open Source
7	Lung PET/CT	Open Source
8	Neuro MRI	Open Source
9	CT Colonography	Open Source

## 5 Conclusion Future Scope

In the field of medical image assessment, an extensive review and use of in-depth training techniques are presented. In all subfields of medical image analysis including classification, detection and segmentation, it was found that convolutive network-based deep learning techniques are more acceptable. We can solve the deep learning problem by means of techniques like data improvement and transmission, thanks to the small amount of data and a small number of labels. More computer power and better DL architecture are available for larger data sets and pave the way for better performance. Ultimately, this would result in better diagnostics and computer detection systems. Further research is needed on the imagery methods not used today. Advances in medical image testing recently have demonstrated a great deal of benefits for deep learning technology. The future work of this paper is to develop a framework for medical image analysis.

## References

1. Qiu, T., Wen, C., Xie, K., Wen, F.Q., Sheng, G.Q., Tang, X.G.: Efficient medical image enhancement based on CNN-FBB model. *IET Image Proc.* **13**(10), 1736–1744 (2019)
2. Hariharakrishnan, J., Bhalaji, N.: Adaptability analysis of 6LoWPAN and RPL for healthcare applications of internet-of-things. *J. ISMAC* **3**(02), 69–81
3. Mugunthan, S.R., Vijayakumar, T.: Design of improved version of sigmoidal function with biases for classification task in ELM domain. *J. Soft Comput. Paradigm (JSCP)* **3**(02), 70–82 (2021)
4. Manoharan, J.S.: Study of variants of extreme learning machine (ELM) brands and its performance measure on classification algorithm. *J. Soft Comput. Paradigm (JSCP)* **3**(02), 83–95

5. Sungheetha, A., Sharma, R.: Design an early detection and classification for diabetic retinopathy by deep feature extraction based convolution neural proceedings by previous publications network. *J. Trends Comput. Sci. Smart Technol. (TCSST)* **3**(02), 81–94 (2021)
6. Neelapu, R., Devi, G.L., Rao, K.S.: Deep learning based conventional neural network architecture for medical image classification. *Traitement du Signal* **35**(2), 169–182 (2018)
7. Srinivasulu, A., Ramanjaneyulu, K., Neelaveni, R., et al.: Advanced lung cancer prediction based on blockchain material using extended CNN. *Appl. Nanosci.* (2021)
8. Rajan, M.S., Dilip, G., Kannan, N., et al.: Diagnosis of fault node in wireless sensor networks using adaptive neuro-fuzzy inference system. *Appl. Nanosci.* (2021)
9. Sangeetha, Y., Majji, S., Srinagesh, A., et al.: Authentication of symmetric cryptosystem using anti-aging controller-based true random number generator. *Appl. Nanosci.* (2021)
10. Kothapalli, S., Samson, M., Majji, S., Patnala, T.R., Karanam, S.R., Pasumarthi, C.S.: Comparative experimental analysis of different Op-amps using 180 nm CMOS technology. In: 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), pp. 1–6 (2020). <https://doi.org/10.1109/ic-ETITE47903.2020.440>
11. Jana, B., Thotakura, H., Baliyan, A., et al.: Pixel density based trimmed median filter for removal of noise from surface image. *Appl. Nanosci.* (2021)
12. Patnala, T.R., Jayanthi, D., Shylu, D.S., Kavitha, K., Chowdary, P.: Maximal length test pattern generation for the cryptography applications. In: Materialstoday Proceedings (In press). <https://www.sciencedirect.com/science/article/pii/S2214785320305368>. Available online from 20 Feb 2020
13. Patnala, T.R., Jayanthi, D., Majji, S., Valleti, M., Kothapalli, S., Karanam, S.C.R.: Modernistic way for KEY generation for highly secure data transfer in ASIC design flow. <https://ieeexplore.ieee.org/document/9074200>. Published in IEEE digital Xplore, Electronic ISSN: 2575-7288. Available from 23 Apr 2020
14. Chen, J.-Z.: Design of accurate classification of COVID-19 disease in X-ray images using deep learning approach. *J. ISMAC* **3**(02), 132–148 (2021)
15. Balasubramaniam, V.: Artificial intelligence algorithm with SVM classification using dermoscopic images for melanoma diagnosis. *J. Artif. Intell. Capsule Netw.* **3**(1), 34–42.
16. Selvathi, D., Poornila, A.A.: Deep learning techniques for breast cancer detection using medical image analysis. In: Biologically Rationalized Computing Techniques for Image Processing Applications, pp. 159–186. Springer, Cham (2018)
17. Sudha, S., Jayanthi, K.B., Rajasekaran, C., Sunder, T.: Segmentation of ROI in medical images using CNN-A comparative study. In: TENCON 2019–2019 IEEE Region 10 Conference (TENCON), pp. 767–771. IEEE
18. Cai, Y., Li, Y., Qiu, C., Ma, J., Gao, X.: Medical image retrieval based on convolutional neural network and supervised hashing. *IEEE Access* **7**, 51877–51885 (2019)
19. An, F.P., Liu, Z.W.: Medical image segmentation algorithm based on feedback mechanism CNN. *Contrast Media & Molecular Imaging* (2019)
20. Kumar, E.S., Bindu, C.S.: Medical image analysis using deep learning: a systematic literature review. In: International Conference on Emerging Technologies in Computer Engineering, pp. 81–97. Springer, Singapore
21. Gacsádi, A., Grava, C., Grava, A.: Medical image enhancement by using cellular neural networks. In: Computers in Cardiology, pp. 821–824. IEEE (2005)
22. Lou, A., Guan, S., Loew, M.H.: DC-UNet: rethinking the U-Net architecture with dual channel efficient CNN for medical image segmentation. In: Medical Imaging 2021: Image Processing, vol. 11596, p. 115962T. International Society for Optics and Photonics (2021)
23. Zhao, C., Han, J., Jia, Y., Fan, L., Gou, F.: Versatile framework for medical image processing and analysis with application to automatic bone age assessment. *J. Electr. Comput. Eng.* (2018)
24. Miranda, E., Aryuni, M., Irwansyah, E.: A survey of medical image classification techniques. In: 2016 International Conference on Information Management and Technology (ICIMTech), pp. 56–61. IEEE (2016)
25. Dabeer, S., Khan, M.M., Islam, S.: Cancer diagnosis in histopathological image: CNN based approach. *Inf. Med. Unlocked* **16**, 100231 (2019)

26. Moreno, S., Bonfante, M., Zurek, E., San Juan, H.: Study of medical image processing techniques applied to lung cancer. In: 2019 14th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1–6. IEEE
27. Hajabdollahi, M., Esfandiarpoor, R., Sabeti, E., Karimi, N., Soroushmehr, S.R., Samavi, S.: Multiple abnormality detection for automatic medical image diagnosis using bifurcated convolutional neural network. *Biomed. Signal Process. Control* **57**, 101792 (2020)
28. Niyaz, U., Sambyal, A.S.: Advances in deep learning techniques for medical image analysis. In: 2018 fifth international conference on parallel, distributed and grid computing (PDGC), pp. 271–277. IEEE (2018)
29. Teng, L., Li, H., Karim, S.: DMCNN: a deep multiscale convolutional neural network model for medical image segmentation. *J. Healthcare Eng.* (2019)
30. Feng, N., Geng, X., Qin, L.: Study on MRI medical image segmentation technology based on CNN-CRF model. *IEEE Access* **8**, 60505–60514 (2020)
31. Jung, K.H., Park, H., Hwang, W.: Deep learning for medical image analysis: applications to computed tomography and magnetic resonance imaging. *Hanyang Med. Rev.* **37**(2), 61–70 (2017)
32. Vetova, S.: Comparative analysis on CNN and wavelet features based technology for medical image classification. In: AIP Conference Proceedings, vol. 2333, no. 1, p. 030003. AIP Publishing LLC
33. Madhu, B., Holi, G.: CNN approach for medical image authentication. *Indian J. Sci. Technol.* **14**(4), 351–360 (2021)
34. Tang, Z., Chen, K., Pan, M., Wang, M., Song, Z.: An augmentation strategy for medical image processing based on statistical shape model and 3D thin plate spline for deep learning. *IEEE Access* **7**, 133111–133121 (2019)
35. Qayyum, A., Anwar, S.M., Awais, M., Majid, M.: Medical image retrieval using deep convolutional neural network. *Neurocomputing* **266**, 8–20 (2017)
36. Ayyachamy, S., Alex, V., Khened, M., Krishnamurthi, G.: Medical image retrieval using Resnet-18. In: Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications, vol. 10954, p. 1095410. International Society for Optics and Photonics

# Data-Intensive Physics Analysis in Azure Cloud



Igor Sfiligoi, Frank Würthwein, and Diego Davila

**Abstract** The Compact Muon Solenoid (CMS) experiment at the Large Hadron Collider (LHC) is one of the largest data producers in the scientific world, with standard data products centrally produced, and then used by often competing teams within the collaboration. This work is focused on how a local institution, University of California San Diego (UCSD), partnered with the Open Science Grid (OSG) to use Azure cloud resources to augment its available computing to accelerate time to result for multiple analyses pursued by a small group of collaborators. The OSG is a federated infrastructure allowing many independent resource providers to serve many independent user communities in a transparent manner. Historically, the resources would come from various research institutions, spanning small universities to large HPC centers, based on either community needs or grant allocations, so adding commercial clouds as resource providers is a natural evolution. The OSG technology allows for easy integration of cloud resources, but the data-intensive nature of CMS compute jobs required the deployment of additional data caching infrastructure to ensure high efficiency. The month-long provisioning exercise was successful, both from a user experience and system utilization point of view, and demonstrated the viability of this approach.

**Keywords** Cloud · OSG · CMS · Data caching · Provisioning · Physics

---

I. Sfiligoi (✉) · F. Würthwein · D. Davila  
University of California San Diego, La Jolla, CA, USA  
e-mail: [isfiligoi@sdsc.edu](mailto:isfiligoi@sdsc.edu)

F. Würthwein  
e-mail: [fkw@ucsd.edu](mailto:fkw@ucsd.edu)

D. Davila  
e-mail: [didavila@ucsd.edu](mailto:didavila@ucsd.edu)

## 1 Introduction

Scientific computing needs typically vary with time, and most fields occasionally experience significant spikes in demand, e.g., right before major conferences. Since provisioning-dedicated local resources for peak demand is prohibitively expensive, multi-domain research platforms like the Open Science Grid (OSG) [1] have seen significant success in both aggregating resources located at participating institutions and by provisioning additional resources to scientists in times of need, by either borrowing compute capacity from unrelated science domains or by abstracting access to specialized resources, like XSEDE HPC centers [2]. Adding commercial clouds as resource providers is a natural evolution of the trend.

A major problem when aggregating resources from multiple providers is data handling. While network throughput has improved significantly over time, thanks to projects like the Pacific Research Platform [3], the network latency of wide-area networks is a fundamental physics property and cannot be significantly improved upon. Data-intensive applications that process a variety of non-contiguous remote data can thus incur a significant penalty in such environments, if accessed directly. Many OSG user communities thus rely on caching services, creating an effective content delivery network (CDN) [4] for the most often used data.

This work is focused on how a local institution, University of California San Diego (UCSD), partnered with the OSG to use Azure cloud resources to augment its available computing to accelerate time to results for multiple physics analyses pursued by a small group of collaborators. This paper is structured as follows: Sect. 2 provides a short description of the Compact Muon Solenoid (CMS) [5] science community that was the focus of this work. Section 3 provides an overview of the OSG infrastructure. Section 4 provides the technical setup used to integrate Azure cloud resources into OSG, including the additional CDN services needed by CMS. Finally, Sect. 5 provides an overview of the operational experience running jobs on this setup.

### 1.1 Related Work

Several communities have implemented their own cloud provisioning solutions to extend on-prem compute pools, an example being the Fermilab-focused HEPCloud [6]. Custom cloud-based scheduling solutions have also been proposed [7]. This work differs from those solutions in that it leverages existing, production OSG services, instead of creating a proprietary solution. The desired outcome was achieved by adding only a lightweight cloud provisioning setup that did not require any new code to be written.

## 2 The CMS Collaboration

The Compact Muon Solenoid (CMS) experiment at the Large Hadron Collider (LHC) is operated by a collaboration of more than 200 institutions in more than 40 countries. The experiment produced several exabytes (EBs) of data in the first decade of its operations, with over one EB of data currently stored in archival storage alone. In support for all that data, the CMS collaboration operates about 200 PB of active storage infrastructure, globally.

Most of the computing is contributed in-kind by institutions and national funding agencies the world over. The experiment is very versatile and allows many independent high-energy physics (HEP) studies. The data produced during the first decade of operations allowed the scientists to publish more than 1000 scientific papers. Consequently, at any point in time, hundreds of unique analyses are being worked on by the thousands of members of the collaboration. Standard data products are centrally produced and then used by often competing teams within the collaboration.

This work is focused on how a local institution, University of California San Diego (UCSD), used commercial cloud to augment its available computing to accelerate time to result for multiple analysis pursued by a small group of collaborators. A half dozen collaborators from UCSD, UC Santa Barbara (UCSB), Boston University, and Baylor University accomplished in a few days what would normally take them multiple weeks. This in turn motivated them to pursue additional studies that they would have normally not dared to do, given their compute intensive nature. Cloud integration thus both accelerated their science and allowed them to pursue science that would have otherwise been out of reach.

Some of the analyses that used the cloud resources are summarily described in the subsections below.

### 2.1 *The Scouting Dimuon Analysis*

The CMS scouting dimuon analysis uses scouting tier data (collision data collected at a higher rate and with smaller event content, compared to normal triggers) to look for displaced dimuon resonances at low masses (as low as twice the muon mass).

Based on the observed and predicted event counts at different dimuon masses, the computing resources are used to calculate the upper limits on various models of new physics, through high-dimensional maximum likelihood fits. The fit for one model hypothesis typically takes a few CPU days, and in this analysis, there were tens of thousands of hypotheses to test. The outcome of the analysis [8] was the exclusion of several models of potentially new physics, e.g., where the Standard Model Higgs boson decays into a pair of hypothetical “dark photons,” which can travel some distance in the detector before decaying into a pair of muons.

## 2.2 *The Two Gauge Boson Analysis*

This analysis is looking for a Higgs boson produced in association with two gauge bosons. In the standard model, this is a very rare signature, but deviations from the predicted quartic couplings of two Higgs and two gauge bosons would lead to potentially large excesses at high energies in this final state. In addition, this final state is also sensitive to triple Higgs couplings. As these couplings have never been measured, looking for non-standard values is a high profile endeavor at the LHC. In essence, the Higgs boson that was only discovered in 2012 (and lead to the 2013 Nobel Prize in Physics) is now being used for studies of its properties in order to verify that the Higgs found in nature is indeed the Higgs predicted by theory.

In CMS simulation of event generators, restriction on the phase-space is often required due to the branching fraction of boson decays to reduce the total computation time. The UCSD team performed the full chain of the CMS simulation of event generations for the desired rare processes with an inclusive phase space. The about 50 M events generated will be used to study and design several future physics analyses leading to potentially multiple publications.

## 2.3 *The Top-W Scattering Analysis*

The Top-W scattering analysis searched for a process where a top quark and a W boson scatter on each other, leading to a top quark pair, a W boson plus a high momentum forward quark in the final state. This particular analysis is looked for the process in the final state of two leptons of the same charge, a b-tagged jet, and multiple other jets.

Jets are remnants that the detector sees as a signature for quarks and gluons. A b-tagged jet is indicative of a b-quark. The top quark decays to a b-quark and a W. A collision with two tops and one W will thus have three W's and two b-quarks in the final state. This work has thus a lot in common with looking at a multiple-car traffic accident, trying to figure out what happened in the original collision by studying the debris. What was the original accident, and who piled on afterward.

The targeted process is a sensitive probe of modifications in the couplings between top quarks and the W, Z, and Higgs bosons.

## 3 **The Open Science Grid Computing Model**

The OSG computing model is based on federation principles. There is no central policy entity. Each resource provider is autonomous, both in terms of resource acquisition and user community access policies. Similarly, each user community, also known as a virtual organization (VO), is autonomous, both in terms of setting

priorities among their constituents and in contracting for access with any resource provider. OSG provides the necessary trust and technical mechanisms, a.k.a. the glue that allows seamless integration of the many resource providers and user communities without combinatorial issues.

From a technical point of view, OSG provides 4 + 1 main service categories:

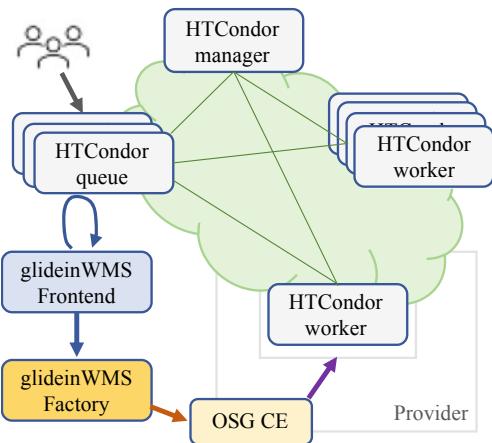
1. An authentication framework, currently based on the Grid Community Toolkit (GCT) X.509 Proxy certificate infrastructure [9], minimizes the security risks of distributed data distribution [10].
2. A portal implementation for accessing compute resources at various resource providers, also known as a Compute Entrypoint (CE), alongside a set of standard packages expected on all compute resources.
3. A set of data handling tools and services, including community-specific software distribution mechanism, data storage portals, and content-distribution services.
4. A central accounting system [11].
5. An overlay pilot workload management system, i.e., glideinWMS [12]. While not a requirement, most OSG user communities do use it, including the CMS collaboration.

Most of the OSG elements are completely generic and will concurrently serve many user communities. Each individual resource provider decides which VOs to serve, advertising the supported VOs through a central OSG registry [13]. With few exceptions, authentication and authorization are based on group membership, i.e., the resource providers are only concerned with which VO a credential belongs to, and any intra-VO policies are dealt outside the OSG security framework.

Like most OSG communities, the CMS collaboration relies on HTCondor for all its workload management needs, with a significant fraction of resources provisioned through OSG services. The provisioning of compute resources from OSG is regulated by a glideinWMS fronted, which monitors the job queues and decides whether more resources are needed. When appropriate, it instructs the glideinWMS factory, operated by OSG as a central service, to provision additional compute resources through one or more Compute Entrypoints using delegated credentials. The factory provisioning request includes the necessary setup details needed for the provisioned compute resources to securely join the CMS HTCondor pool. From that point on, those resources can be scheduled to run any suitable job; this mode of operation is usually referred to as an overlay or pilot setup. A schematic overview is available in Fig. 1. Furthermore, to better understand the system utilization, any job running on OSG resources is recorded in the central OSG accounting system.

Note that each CE converts OSG requests into local ones, leaving full control of the provisioning logic to the resource provider. The serviced provisioning requests are however still recorded in the central OSG accounting system. More details on CE internals are available in Sect. 4.

**Fig. 1** Overview of the OSG workload management system



### 3.1 Data Handling Considerations

CMS jobs rely on CVMFS [14], a POSIX-like read-only distributed file system for software distribution. The CVMFS driver is expected to be installed on all provisioned resources, while all the content is dynamically downloaded by the driver from a central location using the HTTP protocol. Given the relatively static nature of the software, caching servers, based on Apache Squid, should be deployed by resource providers to both compensate for large network latencies and minimize network traffic on the central servers.

Most jobs also require access to the detector calibration parameters, which are relatively big but change very slowly. Many independent jobs will thus access the same data, so CMS relies on an Apache Squid-based content delivery network (CDN), named Frontier [15], to both reduce the access latency and decrease the network bandwidth used. This same infrastructure is typically used for CVMFS caching, too.

Handling of physics data is managed by the jobs themselves, with the bulk of the data being transferred from and to CMS-managed servers. Since this data are contiguous and rarely reused, no OSG-operated support services are needed.

## 4 Azure Cloud Resource Integration

In OSG, all resource provisioning details are abstracted behind a portal, also known as the Compute Entrypoint (CE). The used portal implementation is HTCondor-CE [16] and relies on a batch system paradigm, translating global requests into local ones. After successful authentication, the presented credential is mapped to a local system account, and all further authorization and policy management are handled in the local account domain, typically at the VO level. Several backend batch systems

are supported; we used HTCondor for this work, due to its proven scalability and extreme flexibility.

We used a single HTCondor pool setup per CE to manage resources from many Azure regions. Since HTCondor can easily manage resources distributed over many WAN links and can deal with NAT-ed network environments, there was no compelling reason to create and maintain several independent instances. All connections were secured with mutual authentication using a shared secret, with integrity checks on all transfers. All the worker node resources obviously were provisioned from the Azure cloud. Since OSG computing model allows for preemption, we only used the cheaper pre-emptible instances [17], also known as spot instances.

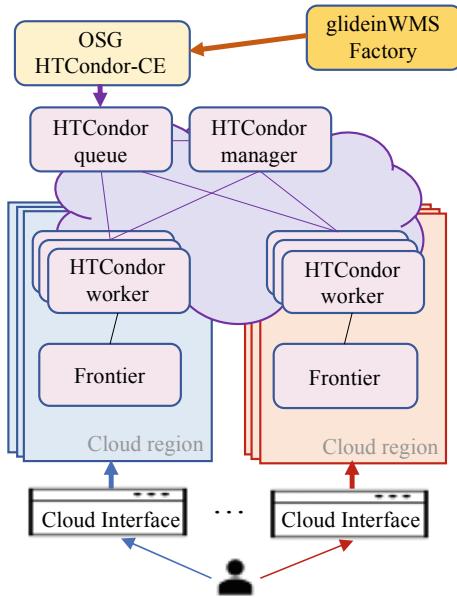
We created a custom virtual machine image, starting with the base operating system, then adding the prescribed OSG worker node software, including the CVMFS drivers, and the fully configured HTCondor daemons pointing to our CE pool. Since the images are private, we embedded the shared secret, too. To transparently use the created image in all the targeted cloud regions, we registered the image in the Azure shared image gallery and used its replication mechanisms.

For ease of use, we used the native Azure interface for resource provisioning, relying on its group provisioning mechanisms, namely the virtual machine scale sets (VMSS). One VMSS was needed per targeted region. Since only the resource manager is exposed to this interface, our previous experiences [18] convinced us that the ease of use and flexibility of native interfaces significantly outweighs the added burden of dealing with multiple independent interfaces. With VMSS, one only has to set the desired number of instances at runtime, and each VMSS will provision as much as it is available at that point in time without any further operator intervention. The number of resources requested in each of the regions was mostly manually determined based on the desired spend rate and observed preemption rates. We would prefer less expensive but stable regions at smaller scales but would expand into more expensive regions when observing non-trivial preemption rates in the more favorable ones.

Once provisioned, the worker node instances would join the HTCondor pool and be available to run any jobs present in the queue. HTCondor would manage the priorities between jobs belonging to any number of CE users, just like it would at any HTCondor-managed resource provider. As already mentioned in the previous section, an OSG accounting probe installed on the CE collects usage statistics and registers them with the central OSG accounting system.

To minimize both network latency experienced by the jobs and network volume flowing into the cloud, we also provisioned, installed, and operated one Frontier cache server per employed virtual private network, which in practice meant one instance per Azure cloud region. Note that a Frontier cache sever is really just a properly configured Apache Squid instance. We ensured that the relative IP address was always the same, e.g., X.Y.0.4, making it easy for the worker instances to programmatically determine its location at startup time. Since many worker instances relied on Frontier cache servers for their operation, we provisioned them using the more expensive but reliable on-demand instances.

**Fig. 2** Overview of a cloud-enabled OSG CE



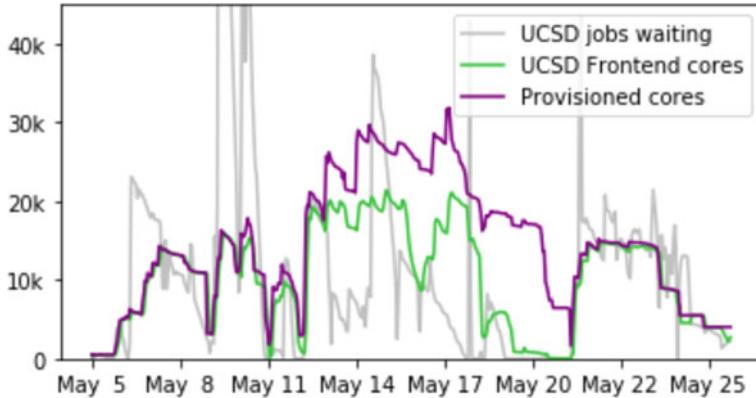
A summary overview of the setup is available in Fig. 2. Note that the depicted HTCondor processes are those managing the internal CE resource pool. The service jobs submitted by the glideinWMS will bring in, configure, and start another set of HTCondor processes, as shown in Fig. 1; HTCondor allows for nested resource management.

## 5 The Operational Experience

For convenience, the OSG CE was deployed as an on-demand instance in one of the Azure cloud regions. The public IP was associated with a UCSD-managed DNS record and then registered with the OSG production services.

CMS jobs are tuned to expect 2 GB of RAM for each CPU core, and prefer utilizing 8 cores concurrently although a subset of them is single-threaded. At the time of the exercise, i.e., May 2021, the most cost-effective Azure instance type that fit those needs was the F16s v2, providing 16 Intel Xeon Platinum 8272CL CPU cores and 32 GB of RAM at an average cost of \$3.3/day per spot instance. HTCondor was configured to dynamically partition the available CPU cores among the users, based on their requests. In order to keep preemption rates reasonably low, we provisioned compute from 5 regions, four in North America and one in Europe. One Frontier cache instance was deployed in each of those regions, too.

As mentioned in Sect. 2, CMS has several user groups who may compete for the same resources. For the purpose of this work, we admitted only pilot jobs serving the



**Fig. 3** Overview of the provisioned cores versus time, and the share going toward UCSD users

UCSD community and pilots serving the Fermilab community, with the UCSD pilots having priority over the Fermilab pilots. We used this setup due to the limited number of users served by the UCSD glideinWMS Frontend, thus resulting in occasionally very spiky usage pattern. Having low-priority jobs in the local HTCondor queue allowed us to make provisioning changes only a couple times a day, while still both providing a high number of CPU cores to the high priority users and fully utilizing the provisioned resources at all times. That said, we limited the UCSD pilots to approximately 20 k CPU cores, so the Fermilab pilots did get a significant fraction of the resources when we decided to provision up to 30 k CPU cores for a few days.

An overview of the provisioned resources and its correlation to the UCSD Frontend is shown in Fig. 3. The figure also shows the impact of Azure preemption at high instance counts; most of the gradual downward slopes are due to Azure preemption alone; we chose not to automatically replace the preempted instances. Note that at lower provisioning levels, e.g., around May 22nd, the line is instead quite flat, indicating that we experienced very little preemption in Azure.

The resource utilization of the compute instances was generally reasonably high. As can be seen from the monitoring snapshot in Fig. 4, the CPU utilization hovered around 80% most of the time, with short-lived dips due to either temporary lack of demand or a change in user jobs' mixture, which introduces inefficiencies due to the Frontier cache not having the necessary data in its cache yet. As mentioned before, the UCSD user job activity was rather spiky. A similar utilization pattern would have been expected from an on-prem system, too.

The cloud run achieved its goal of doubling the CPU core hours delivered by UCSD for the month of May, using about \$70 k worth of Azure credits. CMS UCSD on-prem resources typically deliver around 7 M core hours each month, and the cloud-enabled CE added another 7.3 M. A screenshot of the weekly OSG accounting for the UCSD-associated CMS resources is available in Fig. 5.



**Fig. 4** Screenshot of Azure monitoring, showing CPU utilization versus time of one VMSS



**Fig. 5** Screenshot of the weekly OSG accounting for UCSD-associated CMS resources. Each bar represents one week worth of compute

The additional resources had a very large impact on the produced science, especially because most of them were used to support just a few select analyses of interest to UCSD.

The changes to the user workflow were minimal, too. As an example, the dimuon analysis group reports that they only had to add one HTCondor ClassAd attribute to their job config files, and that alone allowed them to transparently run on the cloud resources. The additional resources allowed them to get an estimated  $5\times$  faster turnaround time, i.e., from months to just over a week, compared to using just regular CMS resources as usual, both because of higher peak resource availability, i.e., 20 k CPU cores versus more typical 9 k, and larger integrated resource availability over the longer period of time. Due to high resource demand and CMS fair share policies, a single CMS user cannot really expect more than  $O(100\text{ k})$  CPU core hours per week using only the regular CMS resources.

## 6 Summary and Conclusion

During the month of May 2021, we more than doubled the compute capacity of the UCSD CMS center by expanding it into the Azure cloud. This was done in such a way that it mostly benefited a set of high profile CMS data analyses the UCSD physics department is involved with. Graduate students, post-docs, and faculty accessed Azure cloud resources with a simple one-line statement in their workload configuration files. The added resources allowed those users to accomplish in days what would have otherwise taken those months to do. Seeing this acceleration in time to science results, several of them added in additional studies they would have normally not dared to do, given their resource intensiveness.

This was possible due to CMS using the OSG provisioning services. By abstracting the cloud resources behind an OSG CE, those resources were essentially identical to the on-prem compute resources those users normally have access to.

Given the data-intensive nature of most CMS analyses, the remote nature of cloud resources required the deployment of content delivery network services in Azure to minimize data-access-related inefficiencies. This was particularly urgent given the use of multiple cloud regions, spanning both the US and European locations. The utilized Frontier caches performed greatly, keeping the CPU utilization on par with on-prem resources.

**Acknowledgements** This work has been partially funded by the US National Science Foundation (NSF) Grants OAC-2030508, MPS-1148698, OAC-1826967, OAC-1836650, OAC-1541349, PHY-1624356, and CNS-1925001. We gratefully acknowledge the credits provided by Microsoft that covered all the Azure cloud expenses.

## References

1. Pordes, R., et al.: The open science grid. *J. Phys. Conf. Series* **78**, 012057 (2007). <https://doi.org/10.1088/1742-6596/78/1/012057>
2. Towns, J. et al.: XSEDE: accelerating scientific discovery. *Comput. Sci. Eng.* **16**(5), 62–74 (2014). <https://doi.org/10.1109/MCSE.2014.80>
3. Smarr, L., et al.: The Pacific research platform: making high-speed networking a reality for the scientist. In: PEARC ‘18: Proceedings of the Practice and Experience on Advanced Research Computing, art. 29, pp. 1–8, July 2018. <https://doi.org/10.1145/3219104.3219108>
4. Fajardo, E., et al.: Creating a content delivery network for general science on the internet backbone using XCaches. *EPJ Web of Conf.* **245**, 04041 (2020). <https://doi.org/10.1051/epjconf/202024504041>
5. Chatrchyan, S., et al.: The CMS experiment at the CERN LHC. *Instrumentation* **3**, S08004 (2008). <https://doi.org/10.1088/1748-0221/3/08/S08004>
6. Mhashilkar, P., et al.: HEPCloud, an elastic hybrid HEP facility using an intelligent decision support system. *EPJ Web Conf.* **214**, 03060 (2019). <https://doi.org/10.1051/epjconf/201921403060>
7. Haixiang, W., Smys, S.: Secure and optimized cloud-based cyber-physical systems with memory-aware scheduling scheme. *J. Trends Comp. Sci. Smart Tech. (TCSST)* **2**(03), 141–147. <https://doi.org/10.36548/jcsts.2020.3.003>

8. CMS Collaboration: Search for long-lived particles decaying into two muons in proton-proton collisions at  $\sqrt{s} = 13$  TeV using data collected with high rate triggers. CERN Report CMS-PAS-EXO-20-014 (2021). <https://cds.cern.ch/record/2767659?ln=en>
9. GridFTP and GSI Migration. <https://opensciencegrid.org/technology/policy/gridftp-gsi-migration/>
10. Shakya, S.: An efficient security framework for data migration in a cloud computing environment. *J. Artif. Intell.* **1**(01), 45–53. <https://doi.org/10.36548/jaicn.2019.1.006>
11. Retzke, K., et al.: GRACC: new generation of the OSG accounting. *J. Phys. Conf. Ser.* **898**, 092044 (2017). <https://doi.org/10.1088/1742-6596/898/9/092044>
12. Sfiligoi, I., Bradley, D.C., Holzman, B., Mhashilkar, P., Padhi, S., Wurthwein, F.: The pilot way to grid resources using glideinWMS. In: 2009 WRI World Congress on Computer Science and Information Engineering, pp. 428–432, July 2009. <https://doi.org/10.1109/CSIE.2009.950>
13. OSG Topology Interface. <https://topology.opensciencegrid.org/>
14. Blomer, J., Buncic, P., Charalampidis, I., Harutyunyan, A., Larsen, D., Meusel, R.: Status and future perspectives of CernVM-FS. *J. Phys. Conf. Ser.* **396**, 052013 (2012). <https://doi.org/10.1088/1742-6596/396/5/052013>
15. Dykstra, D., Lueking, L.: Greatly improved cache update times for conditions data with Frontier/Squid. *J. Phys. Conf. Ser.* **219**, 072034 (2010). <https://doi.org/10.1088/1742-6596/219/7/072034>
16. Bockelman, B., Livny, M., Lin, B., Prelz, F.: Principles, technologies, and time: the translational journey of the HTCondor-CE. *J. Comp. Sci.*, 101213 (2020). <https://doi.org/10.1016/j.jocs.2020.101213>
17. Sfiligoi, I., et al: Demonstrating a pre-Exascale, cost-effective multi-cloud environment for scientific computing. In: PEARC '20: Practice and Experience in Advanced Research Computing, pp. 85–90, July 2020. <https://doi.org/10.1145/3311790.3396625>
18. Sfiligoi, I., Schultz, D., Würthwein, F., Riedel, B.: Pushing the cloud limits in support of IceCube science. *IEEE Internet Comput.* **2**(1) (2021) <https://doi.org/10.1109/MIC.2020.3045209>

# Identification of Assets in Industrial Control Systems Using Passive Scanning



Aju Mathew Thomas, Mounesh Marali, and Lakshmikiran Reddy

**Abstract** Attacks on industrial control system (ICS) and critical infrastructure network (CIN) are becoming increasingly prevalent. Asset inventory is the bedrock and a critical resource for managing cybersecurity risk in ICS and operational technology (OT). Lack of visibility into assets is one of the significant challenges in the ICS environment. Scanning a network can assist in compiling an exhaustive inventory of all connected devices. In ICS and other CIN, passive scanning is preferred over active scanning, as the latter may disrupt operations, resulting in severe consequences such as production, economic, and human losses. This paper proposes and develops a framework for asset inventory creation that utilizes a passive scanning technique. The framework runs a Python program that performs live capture from all active network interfaces in a Linux Docker container. We utilize a MySQL database to store the asset information captured during the scanning process. Additionally, a graphical representation of the asset's network topology is generated. The results demonstrate that our proposed solution can detect and capture all assets associated with a particular interface based on its IP address and accurately identify more than 70 % of all devices.

**Keywords** Industrial control systems · Asset inventory · Passive scanning · Distributed control systems · SCADA · Port mirroring

---

A. M. Thomas (✉)

TIFAC-CORE in Cyber Security, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India

e-mail: [cb.en.p2cys19001@cb.students.amrita.edu](mailto:cb.en.p2cys19001@cb.students.amrita.edu)

M. Marali · L. Reddy

IA Control Technologies, ABB India Development Center, Bangalore, India

e-mail: [mounesh.marali@in.abb.com](mailto:mounesh.marali@in.abb.com)

L. Reddy

e-mail: [lakshmikiran.reddy@in.abb.com](mailto:lakshmikiran.reddy@in.abb.com)

## 1 Introduction

Industry 4.0 is about connectivity and envisions the manufacturing of tomorrow's products through an autonomous production process in intelligent factories with interconnected machines. One of the most significant aspects of Industry 4.0 is the convergence of information technology (IT) and operational technology (OT) environments, which necessitates the use of traditional OT controls and architecture for more effective monitoring and security of critical environments [1]. Traditionally, ICS in a manufacturing platform was isolated from IT networks, but today's increased socioeconomic pressure necessitates an integrated approach. Accurate and up-to-date information about the plant and process must be available at the plant and enterprise levels, as well as to supply chain partners [2]. Integration of IT and OT enables a business to gain a more detailed view of its operations and provides numerous benefits, including increased efficiency, reduced downtime through preventive maintenance solutions, and the ability to predict potential failures before they occur [3]. Additionally, this association may introduce additional complexity, such as misconfigurations and operational errors, thereby increasing the risk of cyber and operational threats. Within their ICS networks, organizations must adhere to strict cyber security best practices [4]. ISA99/IEC-62443 is a set of global standards primarily designed to address technical and process-related aspects of cyber security in industrial automation and control systems (IACS), to improve the systems' and networks' safety, availability, integrity, and confidentiality. Similarly to IEC-62443, NIST 800-82 "Guide to ICS Security" describes how to secure multiple types of ICS against cyber-attacks while also taking into account the performance, reliability, and safety requirements unique to ICS [5].

The number of sophisticated targeted attacks on the ICS is on the rise. Asset inventory has been at the top of the SANS critical cyber security control list for many years. According to a 2017 SANS Institute Survey, [6], 40% of ICS security practitioners "lack visibility or support intelligence into their ICS network." As a result, having an accurate and up-to-date asset inventory is the first step in securing legacy ICS infrastructures, as it is impossible to secure a system without first understanding its content and connectivity [7]. Effective asset management, specifically a comprehensive asset inventory, is critical for effective IT management, security reinforcement, and corporate governance and provides operators with a clear view of their computer infrastructure. Failure to maintain a complete and accurate asset inventory has a cost and affects various use cases. The ICS environment poses significant challenges for automated and secure device discovery. Creating and maintaining a centralized inventory of OT systems is a time-consuming and challenging task. Identifying assets, maintaining them, and monitoring them in real-time are just a few of the requirements necessary to address the challenges above. Due to the non-impacting nature of ICS components, passive scanning techniques are preferred over active scanning for building asset inventories in ICS as the latter may have unintended consequences due to the instability of the ICS networks upon receipt of unsupported network data, and passive scanning is less intrusive than active scanning [8]. It is accomplished by

monitoring a copy of the traffic emanating from a switch's switched port analyzer (SPAN) port that provides comprehensive inventory information about the devices, including their model, make serial number, type, version, and location [7].

This paper has proposed and developed a lightweight framework for building an automated inventory of assets used in ICS through the approach of passive scanning. The proposed framework is built on top of a Linux container and makes extensive use of the PyShark library to execute sniffing operations that aid in the formation of an automated asset inventory. The remaining section of the paper is laid out as follows. Section 2 discusses about the related works. Section 3 highlights our proposed framework developed for building the asset inventory. Section 4 discusses the experimental results obtained, and finally, Sect. 5 concludes the paper with future work.

## 2 Related Work

This section discusses some of the more recent research on asset identification in industrial networks that has gained traction. Matthias et al. [9] proposed and developed a simple passive monitoring tool for identifying industrial devices based on the MAC address contained in the ARP broadcast packets. The tool makes use of a pre-built data pool that contains the MAC address, vendor information, and model names of known devices and employs a MAC address correlation strategy. The unknown device is identified by calculating the distance between its MAC address and the data pool's MAC address. Later, the Common Vulnerability Exposure (CVE) database is used to map the vulnerabilities associated with that particular device. The limitation of this approach is that it makes device identification nearly impossible in static environments due to the absence of ARP broadcast packets. Since it employs a comprehensive MAC address correlation approach for device identification, it may result in device mis-identification if multiple products share the same MAC Organizational Unique Identifier (OUI) address.

IDS tools can also be used to detect the presence of devices on a network. One such implementation that garnered attention was Haas's et al. [10] correlation of network and host data for advanced monitoring, in which the authors implemented a framework using zeek and osquery to provide greater network visibility for all devices on a network. Osquery is an active agent that must be configured on each monitored host separately. Osquery sends each host's information to the centralized zeek instance via broker overlay, correlating the data events. The proposed solution, however, will have limitations in the case of ICS, as it will be unable to detect OT assets such as PLCs, RTUs, and so on.

The authors [11] conducted a case study to identify various assets in ICS by utilizing three different open-source tools including Nmap, GRASSMARLIN, and the Industrial Exploitation Framework (IEF), each of which employs either an active or passive approach for identifying the assets. Additionally, the authors discussed the limitations of each tool and the amount of information that can be gleaned from

each. Similarly, Mavrakis et al. [12] identified the OS using p0f's existing signatures, which included fields such as maximum segment size (MSS), window size (WS), and TTL, as well as additional fields such as option layout and IP address version. Along with TCP fingerprinting, the proposed algorithm will be able to extract exact OS names from both SMB and browser protocols. The authors also proposed a method for passive OS fingerprinting that is superior to the state-of-the-art tools such as p0f and PRADS by utilizing a machine learning algorithm on a decision tree model.

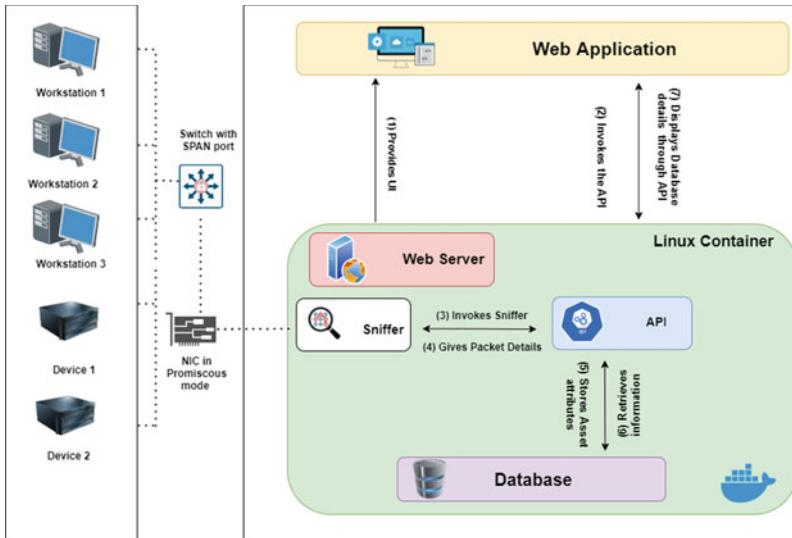
There are numerous different tools available for detecting assets in an ICS environment, both open source and proprietary. Tools such as Forescout's eyeInspect, Nozomi Guardian, Claroty, Verve, Langner's OT Base Asset Discovery, and Microsoft's CyberX are some of the proprietary tools available in the market that can be used for building asset inventory in an ICS environment. These proprietary tools often use a hybrid approach of scanning (passive and active) to build the asset inventory and an active agent to obtain intricate information about an asset. After completing the required installation, the tool's functionality can be accessed via a Web-based GUI. Popular open-source tools such as GRASSMARLIN [13] can also be used to create an inventory of OT assets. However, when compared to proprietary tools, open-source tools have limitations in terms of providing all relevant information about an asset. Wireshark must be installed prior to using GRASSMARLIN, and it currently supports only the older version. Additionally, it does not include cluster-level information or other detailed hardware information. The other open-source tool is the Industrial Exploitation Framework (IEF) [8] that can identify ICS assets such as PLCs based on their communication protocol. The majority of open-source tools are provided as a stand-alone executable file and use either a passive or an active scanning approach.

### 3 Proposed Work

This section details our proposed framework for inventorying assets in ICS networks. The proposed scheme employs complete passive monitoring techniques to extract information from packets, which is then used to construct the inventory. The various components of the proposed framework are described in detail in Fig. 1.

#### 3.1 Architecture

The proposed system has a network of computers and other devices which are connected to a switch with a SPAN port capability. The traffic coming from all the ports of the switch is mirrored to the SPAN port via the idea of port mirroring. Later, this SPAN port is connected to the machine where our proposed framework runs. The network interface of the system where our proposed framework operates is set to a promiscuous mode that can capture all the packets that flow through the SPAN port



**Fig. 1** Proposed architecture

interface. A sample Web application is developed as a test bed for our framework's functionality and is hosted on a Nginx server. The Web application provides a user interface for instructing the Web application to perform live capture or to read packet data from a PCAP file via HTTP requests. The framework invokes an API call to the sniffer module that will initiate the capturing of live packets that comes in the network interface after setting it to promiscuous mode. Our proposed framework necessitates the usage of Wireshark for doing the live capture of packets across all active network interfaces. Wireshark is the fundamental component of our proposed ecosystem.

We utilized Python3 for developing the framework and used the *PyShark* module to accomplish the live packet capture. The PyShark library will operate as the sniffer module in the diagram. The PyShark library relies on the *tshark* and *dumpcap* modules. The packet information obtained by the sniffer module will be taken through the API request and save the required properties pertaining to an asset into a database. We used the MySQL database to store the results of our scanning into respective columns of the asset table. The columns of the asset table hold the information related to various attributes of an asset. Once the information is stored in the database, the information may be retrieved on the UI side by another API call to the database. The complete framework is contained within a Docker container running on Linux. We chose an Ubuntu Docker container to execute our framework because it is easier to initialize the network interface to promiscuous mode in a Linux environment than it is in a Windows one. Additionally, our system ensures that Wireshark is configured and installed programmatically in the Linux Docker container, rather than manually on Windows. This is the critical component of our proposed framework,

**Table 1** List of table attributes

Attributes		
Source IP	Destination IP	Source MAC address
Destination MAC address	Source vendor	Destination vendor
Protocol	Hostname	Device type
Purdue level	Source port number	Destination port number
Model name	Operating system	Last modified date

which is based on the noninvasive passive scanning technology that has no adverse effect on production networks. Furthermore, we used the Python libraries *matplotlib* and *networkx* to visualize the network topology of assets based on the clustering of subnets for each active network interface. The data extracted from the packet is stored in the asset table's respective attributes, as shown in Table 1. Secondly, we maintained a separate table to store the first three or four octets of the MAC vendor address to compare the octets in the packet to the entries in the table to determine the correct model name for the devices such as PLCs, IEDs, HMIs, and switches.

Figure 2 depicts the program flow diagram that our proposed framework is based on. The framework provides the option of performing live packet capture or reading packets from an existing pcap file. Once the live capture mode is selected, the user is prompted to enter a time-out in seconds, which initiates the scanning of all active interfaces for the duration of the time-out specified by the user. The API stores the attributes associated with the assets extracted from the packet in the database and is used to retrieve the details from the database for displaying it on the user interface.

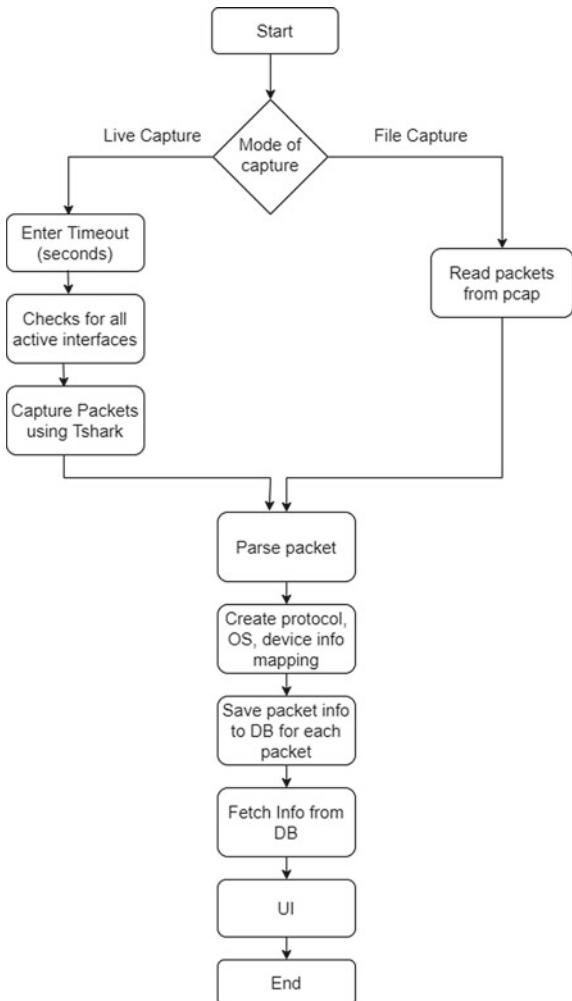
Since industrial networks are a compendium of IT and OT devices distributed across a hierarchical network structure, it is imperative to identify all of the various network assets. The identification of various devices in a network via appropriate scanning techniques assists in the creation of an asset inventory list. As a result, the various approaches we used in our framework to identify assets in a network are discussed in the following subsections.

### 3.2 Identification of Computers

We realized that employing multiple approaches to identifying computers would be more effective than relying on a single approach because it would broaden the scope of detection. Several techniques are used to identify the computer devices on a network, as listed below.

**User-Agent Strings:** The information contained in an HTTP packet's user-agent strings would be critical in determining the device type and Purdue level. This is the first method we used to identify computers or other devices such as mobile phone. The user-agent string contains information about the device's type, operating system

**Fig. 2** Flowchart of the program



version, and browser type. HTTP packets are typically generated in the enterprise network segment's computers. In that case, identifying the Purdue level would be straightforward, as Internet connectivity is unavailable once we reach the control networks. HTTP packets containing user-agent strings, on the other hand, are not generated frequently.

Table 2 shows the operating systems that are typically included in user-agent strings. The algorithm parses the user-agent string information contained in an HTTP packet and then checks for the presence of certain keywords of the operating system's name present in it. The function defined to obtain information about the user-agent string returns the device type and operating system of a given IP address.

**Table 2** Operating system present in user-agent strings

Keywords used from user-agent strings	Operating system	Device type
Linux & Android	Android	Mobile
iPhone	iOS	Mobile
Windows phone	Windows	Mobile
Macintosh	MAC OSX	Desktop
X11 & CrOS	Chrome OS	Desktop
Windows NT 10.0	Windows 10	Desktop
Windows NT 6.3	Windows 8.1	Desktop
Windows NT 6.2	Windows 8	Desktop
Windows NT 6.1	Windows 7	Desktop
Windows NT 6.0	Windows Vista	Desktop
Windows NT 5.2	Windows XP	Desktop
Windows NT 5.1	Windows XP	Desktop
Windows NT	Windows 2000 or lower	Desktop
X11 & Ubuntu	Ubuntu	Desktop

**TTL value and TCP Window Size:** Another straightforward and effective method is to determine the packet's initial TTL value from the IP header and the TCP window size from the TCP header[16]. Typically, the TCP window size is specified in the first packet of a TCP session, i.e., the SYN or SYN+ACK packet. It is included as a fallback method for identifying computers and their underlying operating system in the absence of HTTP packets generated during the sniffing process.

**Protocol Classification** Additionally, the proposed tool considers specific protocols for computer identification.

**Server Message Block (SMB) 2:** SMB2 is a redesigned version of the legacy Windows protocol SMB, and it is used to share files, printers, and serial ports on modern and future Windows hosts. Currently, it has been upgraded to version 3, but the packet signatures for SMB2 and SMB3 are identical. Wireshark uses SMB2 as a display filter for both versions of SMB. It is built on top of the NetBIOS network architecture and communicates via a dedicated TCP port 445. From an SMB2 packet, we can determine the device type as computer, and the packet information also contains the OS version. It is primarily used on Windows-based computers.

**Browser Protocol:** Browser protocol is a Windows-based protocol that runs on top of SMB and enables users to discover computers and network resources. The protocol packet contains information about the operating system version and can be used as an alternative approach to identifying computers in a network. Since this is a Windows-based protocol, the computers detected will be Windows-based machines. Separate logic is included in the program code to parse browser protocol packets and identify the device type as computer.

**NetBIOS Name Service (NBNS):** NBNS is an UDP-based name service protocol. It connects to the network via UDP port 137 and sends broadcast packets once the device is connected to the network following the start of live capturing. The information contained in NBNS packets can be used to determine a machine's hostname. It is the simplest passive technique for obtaining a device's hostname. Windows machines typically generate NBNS packets, and they are an excellent technique for detecting the presence of computers in a network. Additionally, we added this feature in our tool to detect computers and their hostnames in NBNS packets.

### ***3.3 Identification of Printers***

**Internet Printing Protocol (IPP):** We considered the IPP packets generated in the network in order to determine the presence of printers. The IPP protocol communicates via the default TCP port 631. The program logic examines the IPP protocol packet and port number to determine the network printer's identity. Printers are part of the enterprise management network and come in the Purdue level 5.

### ***3.4 Identification of PLC***

PLCs are classified as level 1 in the Purdue Level Hierarchy Reference Model, and they serve as an intermediary between HMIs and field devices. PLCs communicate via industry standard protocols such as Modbus, Profinet, BACnet, and others. Modbus protocol uses the default TCP port 502 for the PLC communication. Through PLCs, operators who monitor and control processes communicate necessary instructions to field devices. We identified PLC models from various vendors using various approaches, including protocol type, port number, MAC vendor OUI, and TTL values. The program logic is written in such a way that it checks the lookup table for information only when it identifies a particular device as a PLC based on the protocol and port number specified in the packet during live capture.

**Siemens PLCs** Siemens PLCs communicate with Siemens HMIs via a proprietary protocol called S7 COMM. It connects via TCP port 102 by default. Siemens PLCs have multiple model names associated with various OUIs, and we used the TTL value and MAC vendor Organizationally Unique Identifier (OUI) to identify the various Siemens PLC model names as mentioned in [15]. Here, the program logic checks the packet for the TTL value, the MAC vendor OUI, and the S7 COMM protocol and then retrieves the model name of the corresponding Siemens PLC from the lookup table. We added four distinct model names in total, as shown in Table 3

**Allen Bradley PLCs** Allen Bradley PLCs from Rockwell Automation typically communicate with other devices via Ethernet/IP or Common Industrial Protocol (CIP). Allen Bradley PLCs in newer versions use the default TCP port 44818, whereas

**Table 3** Model names of Siemens PLC

Protocol name	MAC OUI	TTL value	Model name
S7 Comm	00:1c:06	30	SIMATIC S-1200
S7 Comm	00:1b:1b	30	SIMATIC S-1500
S7 Comm	00:0e:8c	30	SIMATIC S7-300
S7 Comm	00:0e:8c	128	SIMATIC S7-400

older Allen Bradley PLCs used port 2222. Since Ethernet/IP is an open protocol, it can also be used by other leading vendors. If the packet contains the Ethernet/IP or CIP protocol and the port number 44818, the program logic is written to identify the device as a PLC. Once the device type is determined to be a PLC, the logic checks for the Rockwell Automation MAC Vendor OUI and retrieves the appropriate Allen Bradley PLC model name from the lookup table.

**OMRON PLCs:** OMRON PLCs communicate via a proprietary protocol called OMRON and a dedicated TCP port 9600. Wireshark is capable of capturing OMRON packets, and the program algorithm looks up the OMRON Sysmac series PLCs in the lookup table when it detects an OMRON packet during a live capture using the protocol name and port number.

**Additional Models:** Other PLC models from well-known vendors such as Koyo, Yaskawa, Bosch, Keyence, Wago, Fatek, and Korenix communicate via industry standard protocols such as Modbus, Ethernet/IP as shown in Table 4. When the program algorithm detects these protocols and MAC OUIs in the packet, it retrieves the corresponding model name from the lookup table.

### 3.5 Identification of HMIs

Human–machine interface (HMI) or Engineering Workstation is classified as a level 2 component in the Purdue level reference model. Operators communicate with the controller or PLC via HMIs. Typically, HMIs are identified by their protocol, which may include Modbus, S7Comm, Profibus, or Ethernet/IP. Windows-based machines are typically used as Engineering Workstations in large-scale processes, as indicated by the TTL value in the communication packet between the HMI and the PLC. After identifying the device type as HMI, the program logic compares the first three octets of the MAC vendor address to the MAC vendor OUI stored in the table. We have added Beckhoff Embedded Panels and GE Quick Panels that communicate with controllers over dedicated TCP ports 48898 and 57176, respectively.

**Table 4** PLC model names for other leading vendors

Protocol	MAC OUI	Port number	Model name
Modbus/Ethernet IP/CIP/Profinet	00:d0:7c	28784	Koyo Ethernet
Modbus/Ethernet IP/CIP/Profinet	00:20:b5	44818	Yaskawa MP2300 Siec Series
Modbus/Ethernet IP/CIP/Profinet	00:30:de	2455	Wago PFC Series
Modbus/Ethernet IP/CIP/Profinet	00:01:fc	8510	Keyence KV-5000 Series
Modbus/Ethernet IP/CIP/Profinet	70:01:36	500	Fatek FB Series
Modbus/Ethernet IP/CIP/Profinet	00:12:77	502	Korenix 6550 Series
Modbus/Ethernet IP/CIP/Profinet	00:0b:0f	502	Bosch Rexroth ICL
Modbus/Ethernet IP/CIP/Profinet	00:08:53	20547	Schleicher XCX 300

### 3.6 Identification of IEDs or Relays

GOOSE protocol is typically used to communicate between IEDs or relays in a substation and can also be used to determine their presence. To retrieve the appropriate model names, the program logic compares the first three octets of the packet's MAC address to the MAC vendor OUI specified in the lookup table. At the moment, we have added only GE's F650 IEDs with OUI 00:a0:f4.

### 3.7 Identification of Field Devices

Sensors, motors, and actuators are classified as field devices at level 0 of the Purdue reference model. Fieldbus protocols such as HART and PROFINET are used to connect controllers to field devices. The packet containing this protocol information can be used to identify the devices.

### 3.8 Identification of Switches

Typically, switches are identified by the protocols they use for communication. Switches from well-known vendors such as Cisco, HP, and others communicate

via proprietary protocols such as the Cisco Discovery Protocol (CDP), Spanning Tree Protocol (STP), and HP.

**Identification using CDP and STP:** Cisco switches typically use the CDP protocol, and the CDP packet contains the switch's model name and its unique MAC address field. Cisco routers also use the CDP protocol, but the packet information does not include the switch MAC address field in these cases. CDP packets contain information that enables the device to be identified as a switch or a router. The program logic examines the packet's switch MAC address field and classifies the device as a switch. After that, it extracts the first four octets of the packet's MAC address and compares it to the address specified in the lookup table to determine the correct Cisco switch model name. If the packet does not include a switch MAC address field, the device is classified as a router by the program. When STP packets are encountered, the algorithm classifies the device as a switch and compares the first four octets of the packet's MAC address to the address stored in the lookup table to determine the switch's model name.

**Identification using HP:** HP switches communicate via the HP protocol, and the packet pane contains the field "HP switch protocol," enabling the device to be identified as an HP switch. The program logic looks for the MAC vendor OUI and HP protocol in the packet to classify the device as a switch. HP's ProCurve switch family utilizes a dedicated MAC vendor OUI address. Table 5 displays the model names of Cisco and HP switches.

### 3.9 Identification of SCADA/OPC Server

The SCADA or Open Platform Communication (OPC) server is included in level 2 of the Purdue reference model, along with HMI/Engineering Workstations and a

**Table 5** Model names of Cisco and HP switches

Protocol name	MAC vendor OUI	Model name
CDP	0c:68:03	Cisco Catalyst Switches 3850
CDP	00:26:0b	Cisco Nexus 5000 Series
CDP	5c:50:15	Cisco Catalyst 2960 Series
CDP	00:0e:84	Cisco Catalyst 4500 Series
CDP	00:0b:fc	Cisco Catalyst 2960 Series
CDP	00:00:0c	Cisco Nexus 9000 Series
CDP	00:0b:be	Cisco Catalyst 2950 Series
CDP	00:19:2f:a7	Cisco Catalyst 3560 G24PS
CDP	00:18:ba:98	Cisco Catalyst 3560 24TS
HP	00:16:b9	HP ProCurve 2650 Series

Data Historian. It acts as a bridge between the HMI and the controller. The operators obtain data from the controller via the SCADA server. SCADA servers communicate using the OpcUa protocol and the default TCP port 4840. Also, the OpcDa protocol communicates by default via TCP port 11740 or UDP port 1740. Therefore, if the packet contains the OpcUa protocol and the port number as 4840, the program logic determines that the device is an OPC server.

## 4 Experimental Results and Discussion

In this section, we will discuss the experimental results of our proposed solution. Our solution consumes significantly less memory when performing live packet capture. However, memory usage in the read mode increases proportionately with the number of packets parsed. We evaluated the performance of our solution using existing industrial PCAPs from Netresec's 4SICS Geek lounge, which also assisted us in determining the tool's accuracy. These PCAP datasets were collected in a Geek Lounge ICS laboratory equipped with PLCs, RTUs, servers, and other industrial networking hardware such as switches and firewalls. Table 6 summarizes the performance analysis of 4SICS's PCAPs. In the event of live capturing, a timer can be configured to capture packets prior to the start of sniffing. Once the time is set in seconds, the sniffer module lists all active network interfaces and captures packets on each active interface for the duration of the time specified by the user at the start. Additionally, we provide an export option for the scan results for each network interface in CSV format. For file capture mode, the export to CSV feature is also accessible.

Due to the limited availability of data, our solution was unable to parse packets that contain proprietary industry-specific protocols and was less accurate in identifying

**Table 6** PCAP results

File name	Total packets	Time taken to analyze (min)	Total assets	Assets identified	Memory usage range (Mb)
4SICS-GeekLounge-151020.pcap	246137	10	7	5	51.2–51.6
4SICS-GeekLounge-151021.pcap	1253100	56	16	12	107.4–806.8
4SICS-GeekLounge-151022.pcap	2274747	120	24	17	114–1714

**Table 7** Comparison between different open-source passive scanning tools

Parameters	Network miner	GRASSMARLIN	IEF	Zeek-osquery	Our tool
IP address	✓	✓	✓	✓	✓
MAC address	✓		✓	✓	✓
Vendor info	✓				✓
Model names					✓
Hostname	✓				✓
Network topology		✓			✓
Purdue level					✓
Proprietary protocol support					

devices at level 3 (Manufacturing Execution System) and level 4 (Enterprise Resource Planning and Datacenter) in the Purdue reference model.

Additionally, we compared our solution's performance to that of several existing open-source passive scanning tools that provide similar functionality for creating asset inventories by evaluating certain attribute information, as shown in Table 7. Our proposed framework is capable of detecting both OT and IT assets and visualizing their network topology and clustering them according to subnets. We accomplished this goal by utilizing the networkx and matplotlib modules. The framework is also capable of processing approximately 18,000–24,000 packets per minute. Except for GRASSMARLIN, the majority of the open-source technologies presented in Table 7 are primarily focused on detecting IT assets. Additionally, our proposed framework is capable of determining the Purdue level of each asset.

## 5 Conclusion

This research paper describes a tool that can be used entirely passively to create an inventory of ICS assets. We discussed the consequences of active scanning in critical infrastructure networks. Our solution utilizes a comprehensive passive scanning approach to identify both IT and OT assets based on network traffic capture. We were able to store and maintain the inventory by utilizing a centralized database system. We were able to run our Python tool in a lightweight environment with less memory consumption by utilizing a Linux Docker container. Additionally, configuring the network interface to promiscuous mode makes it easier to capture all packets that pass through the interface, which aids in efficiently building the inventory. Based on our findings, we can conclude that an ICS asset inventory can be effectively built using a comprehensive passive scanning approach combined with packet capturing.

## References

1. Marali, M., Sudarsan, S.D., Gogioneni, A.: Cyber security threats in industrial control systems and protection. In: 2019 International Conference on Advances in Computing and Communication Engineering (ICACCE). IEEE (2019)
2. Dzung, D., Naedele, M., Von Hoff, T.P., Crevatin, M.: Security for Industrial Communication Systems. Proc. IEEE **93**(6), 1152–1177 (2005). <https://doi.org/10.1109/JPROC.2005.849714>
3. Yogeshwar B.R., Sethumadhavan M., Srinivasan S., Amritha P.P.: A light-weight cyber security implementation for industrial SCADA systems in the Industries 4.0. In: Senjuu T., Mahalle P.N., Perumal T., Joshi A. (eds.) Information and Communication Technology for Intelligent Systems. ICTIS 2020. Smart Innovation, Systems and Technologies, vol. 196. Springer, Singapore (2021). [https://doi.org/10.1007/978-981-15-7062-9\\_46](https://doi.org/10.1007/978-981-15-7062-9_46)
4. Sivaganesan, D.: A data driven trust mechanism based on blockchain in IoT sensor networks for detection and mitigation of attacks. J. Trends Comput. Sci. Smart Technol. (TCSST) **3**(01), 59–69 (2021)
5. Bhamare, D., Zolanvari, M., Erbad, A., Jain, R., Khan, K., Meskin, N.: Cybersecurity for industrial control systems: a survey, Comput. Secu. **89**, 101677 (2020), ISSN .0167-4048, <https://doi.org/10.1016/j.cose.2019.101677>
6. Brown, B.Gr.: SANS Institute: reading room—Analyst papers. In: SANS, 11 July 2017. <https://www.sans.org/reading-room/whitepapers/analyst/membership/37860>
7. Wedgbury, A., Jones, K.: Automated asset discovery in industrial control systems-exploring the problem. In: 3rd International Symposium for ICS & SCADA Cyber Security Research 2015 (ICS-CSR 2015) 3 (2015)
8. Guide for an Asset Inventory Management in Industrial Control Systems. Spanish National Cybersecurity Institute, Incibe-Cert. [https://www.incibe-cert.es/sites/default/files/contenidos/guias/doc/incibe-cert\\_guide\\_assets\\_inventory\\_2020\\_v1.pdf](https://www.incibe-cert.es/sites/default/files/contenidos/guias/doc/incibe-cert_guide_assets_inventory_2020_v1.pdf). Last accessed 4 June 2021
9. Niedermaier, M.. et al.: Efficient Passive ICS Device Discovery and Identification by MAC Address Correlation. [arXiv:1904.04271](https://arxiv.org/abs/1904.04271) (2019)
10. Haas, S., Sommer, R., Fischer, M.: Zeek-osquery: host-network correlation for advanced monitoring and intrusion detection. In: Hölbl, M., Rannenberg, K., Welzer, T. (eds.) ICT Systems Security and Privacy Protection. SEC 2020. IFIP Advances in Information and Communication Technology, vol. 580. Springer, Cham. [https://doi.org/10.1007/978-3-030-58201-2\\_17](https://doi.org/10.1007/978-3-030-58201-2_17)
11. Abdulrazzaq, M., Wei, Y.: Industrial Control System (ICS) Network Asset Identification and Risk Management (2018)
12. Mavrakis, C.: Passive asset discovery and operating system fingerprinting in industrial control system networks. Wayback archive: <http://web.archive.org/web/20190307110951/> <https://pure.tue.nl/ws/files/46916656/840171-1.pdf> (2015): 840171-1
13. NSA/Cyber Grassmarlin Github. <https://www.github.com/nsacyber/GRASSMARLIN/blob/master/GRASSMARLIN%20User%20Guide.pdf> Last accessed 4 June 2021
14. Hjelmvik, E.: Passive OS Fingerprinting—NETRESEC Blog. Netresec. <https://www.netresec.com/?page=Blog&month=2011-11&post=Passive-OS-Fingerprinting> (2011)
15. Al Ghazo, A.T., Kumar, R.: ICS/SCADA device recognition: a hybrid communication-patterns and passive-fingerprinting approach. In: 2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM), pp. 19–24 (2019)
16. Netresec.: SCADA / ICS PCAP Files from 4SICS. Netresec. <https://www.netresec.com/?page=PCAP4ICS> Last accessed 18 June 2021

# Cybersecurity Governance in Information Technology: A Review of What Has Been Done, and What Is Next



Yang Hoong and Davar Rezania

**Abstract** Cybersecurity is becoming one of the key hallmarks of the world we live in today and is largely present in all organizations in some shape or form. This has refocused the spotlight on the governance of these information technologies (IT) used in cybersecurity. This paper delves to identify the key frameworks used in the governance of information and cybersecurity and to describe the theory underpinnings behind these frameworks. We finish by examining the limitations of the current frameworks and theory, whilst suggesting another framework for future researchers of information security governance to consider.

**Keywords** Governance · Information technology · Frameworks · Theory · Review

## 1 Introduction

Industry 4.0 (The Industrial Revolution of the Internet of Things) has ushered in an age where the increasing use of information technology (IT) has allowed interconnectivity, big data, automation and technology to take the centre stage. Businesses are able to utilize the IT (by making their information and services available round-the-clock online) in order to bring down operating costs and to increase their efficiency. As the adoption of IT grows, so too does the range of threats across a multitude of levels. This has led to a response both at an international level with the United Nations (UN), as well as nationally, with government ministries, provincial and municipal governments all requiring frameworks to protect and govern data. Because of how IT is now embedded in nearly all business systems and processes, any breaches can seriously impact an organization [1, 2]. Cloud computing is a good example of this within the context of Industry 4.0; cloud computing allows for instant access of

---

Y. Hoong (✉) · D. Rezania  
Department of Management, University of Guelph, Guelph, Canada  
e-mail: [yhoong@uoguelph.ca](mailto:yhoong@uoguelph.ca)

D. Rezania  
e-mail: [drezania@uoguelph.ca](mailto:drezania@uoguelph.ca)

services and information, regardless of location or time [3, 4]. Successful cyberattacks can cause harm to not just the financial aspect of an organization; but also to its reputation, ability to attract and retain talent and overall competitive edge [5–7]. These negative impacts have been severe enough to force decision-makers to try to effectively govern security to stop or limit these attacks [8]. The strategic approach towards addressing IT, information security and cybersecurity is often referred to as information security governance (ISG) [9]. In the recent years, due to the changing environment and ever-evolving cyber ecosystem, ISG has become a top priority for many firms at time of present writing [1, 10–12].

The trend towards a more IT-based world has renewed the interest in ISG and placed it firmly back in the spotlight. ISG would be able to provide both academics and practitioners with a clearer understanding of what ISG would mean as we transition from the traditional physical world to a more digitalized world [13, 14]. This paper makes a contribution to the literature by reviewing existing frameworks of ISG, reviewing the theory behind these frameworks and to provide direction for future directions of research by suggesting an alternative framework (multilevel governance) to address the need for change in the current ISG approaches to the new digital world [14].

## 2 Methodology

### 2.1 Searching the Literature

This paper utilized a Boolean search string approach. The literature was primarily sourced from two main databases: Web of Science (WoS) and Google Scholar. Additionally, the search strings can be broken down into three distinct categories listed in Table 1 security, business and governance—with different search expressions for each.

The initial search yielded 120 results found. These results can be categorized into research area: 56 computer science, 33 business economics, 18 information science, 14 government law, 13 engineering, 7 international relations, 6 telecommunications, 5 social sciences and other topics with the remaining being solitary articles scattered amongst other disciplines. By reading abstracts, papers were filtered down (papers

**Table 1** Security, business and governance

Topic	Boolean search string
Security	(‘Cybersecurity’ OR ‘Cyber Security’ OR ‘Cyber-security’ OR ‘Computer Security’ OR ‘Cyber breach’ OR ‘Cyber attack’ OR ‘Data breach’ OR ‘Data leak’)
Business	(Business OR Businesses)
Governance	(Governance)

were excluded with respect to their relevance, as well as the language used in the paper) and the full paper was read (if required). In total, 49 papers were included in the final sample.

### 3 Governance

#### 3.1 *What Is Governance and How It Is Defined in the Literature*

The definitions of governance of informational systems remain fluid, largely due to its interdisciplinary nature. Early conceptualizations of informational systems governance (ISG) largely centred around the technical aspects; for example, it has previously been defined as ‘the overall way in which Information Security (IS) is deployed to mitigate Information Technology (IT) risks’ [8, 15]. However, more and more scholars [16, 17] have started to incorporate other dynamic and flexible factors such as organizational structure towards a more interdisciplinary and enterprise encompassing meaning of ISG, largely due to the nature of the socio-technical environment undergoing frequent and ongoing changes [14].

#### 3.2 *Different Approaches to Governance Found in the Literature*

*Governance in practice.* Existing practices are largely guided by frameworks such as the ones created by the International Organization for Standards (ISO). Specifically, frameworks such as ISO27001 create a set of rules, giving managers concrete steps to follow [18]. These steps help guide their IT and IS to ensure ongoing security compliance. It is important to note that ISOs are not governmentally regulated, rather relying on an opt-in policy—businesses choose to engage in these compliances and frameworks in order to strengthen their security standards.

Due to the increasing shifts to a digital environment, organizations are under pressure to exhibit effective processes and mechanisms in place to manage and respond to the increasing severity cybersecurity threats and frequency of cyberbreaches. In the accounting industry, the world’s largest association representing the accounting profession, the American Institute of Certified Public Accountants (AICPA)—developed a framework that is part of the System and Organization Control (SOC) for informational cybersecurity [19]. This framework largely revolves around the domain of risk management, and functions as a form of internal control—whereby members of the organization utilizes the framework to evaluate the effectiveness of their cybersecurity risk management programmes.

However, researchers note that these frameworks are generic in nature and fail to take into account the nuances of different environments and operations; a lack of theoretically grounded models and empirical evidence [20] on their effectiveness has also been noted—despite these factors, these frameworks continue to be well-established in the industry.

*Corporate governance.* Corporate governance has emerged as one of the first, and dominant theories applied in ISG. In the literature, existing frameworks largely provide the foundational pieces of ISG as they heavily emphasize the technical aspect of IT—i.e. security controls in the form of firewalls [15]. For instance, researchers propose that IT security is paramount and needs to be woven into ISG (and in particular upper-management) in order to create higher gains and productivity [21, 22]. This perspective largely focuses on the management of risks and internal controls of an organization, with upper-level management being seen as the key actors in this approach to governance [12]. In this perspective of ISG, laws and regulations provides the greatest incentive for upper-level management to implement these security controls, as they could potentially suffer heavy losses and penalties [23]. However, researchers are increasingly turning their attention towards the ethics of governance: Matwyshyn [24] argues that companies have a moral and ethical obligation towards ISG, as they should avoid causing unnecessary harm whenever possible. Furthermore, these ethical issues have been found to be wide ranging and context relative. For example, privacy arises in terms of data breaches and keeping information secure from unauthorized access [24–26]. Similarly, autonomy is discussed in terms of data collection, processing, analysis and storage [24, 27]. Despite these findings, limited attention is given to how the business context can affect ISG: largely due to its complex and context-sensitive situation, security is often relegated in importance [14].

*Socio-technical systems.* The second perspective towards ISG is that of a socio-technical perspective. Bostrom and Heinen [28], and Walker et al. [29] state that social-technical system (STS) consists of two dimensions: social and technical. The view of STS is that joint consideration of social and technical elements is necessary [30]. In a modern holistic view, STS is made up of humans applying technology solutions to execute work activities through processes within a social structure (organization) to accomplish set goals. Bostrom and Heinen [28] argue that the social dimension consists of the organizational structure and actors (including people), and the technical dimension consists of the technology and work activities (tasks). Frameworks that are socio-technically oriented include: Dutta and McCrohan's [31] three-pillar framework of critical infrastructure, organization and technology to help upper-level management address IT security from both the social and technical lenses; Veiga and Eloff's [8] people-orientated approach and Maleh et al.'s [32] culture-focused approach towards ISG and IT security. To that end, researchers often underline the importance of this perspective towards ISG, as it combines both the technical perspective and takes into account the human and social elements [8, 20, 33, 34].

*Resilient business perspective.* An emerging perspective towards ISG is one of organizational resilience. Specifically, this perspective takes a business-centric

approach, with von Solms and von Solms [35] coining this ‘business security governance’ (BSG). In BSG, instead of taking a pre-emptive perspective like corporate governance, or a more human-centric and holistic approach like STS, researchers suggest that organizations should deal with IT and IS issues by developing a resiliency-based framework [16, 17]. In this perspective, researchers have proposed aligning security with other key strategic pillars in the company (such as missions, goals and objectives) in order to achieve resiliency [36]. For example, Kauspadiene et al. [6] proposed that due to the increasing threats created by the ever-evolving cyber landscape, businesses must consider multiple stakeholders (partners, collaboration, outsourcing and third parties) in their decision-making and suggest adopting a resilient view of IT security. The researchers propose a holistic, consolidated approach for a high functioning and self-sustaining framework. In other words, researchers propose a more cyber-oriented, topic-specific approach towards becoming resilient.

*ISG as an ongoing process.* One of the main criticisms of ISG frameworks thus far is the lack of flexibility and modification to reflect the changing cyber landscape [37]. To that end, some researchers view ISG as a continuous process—one that constantly requires re-evaluation [38]. Researchers like Knapp et al. [11] examine how external and internal influences can impact organizational IT processes. Furthermore, Haufe and Colomo-Palacios [18] develop a framework to examine how the operation of an IT system can impact ISG, a departure from the focus on measures and controls.

## 4 Discussion and Theoretical Underlying’s

STS. The term ‘socio-technical’ refers to the joint affinity between ‘social’ and ‘technical’. Socio-technical theory is therefore largely established upon two main concepts. Social factors consist of the human aspects; similarly, technical consists of machine aspects [39]. Firstly, the level of synchronization of social and technical factors will determine whether the conditions of the organization or system are successful or unsuccessful. Increasingly, researchers have shown that the mixing of the ‘social’ with the ‘technical’ is a complex issue, as people are not machines, and machines are not human, leading to the ‘social’ and the ‘technical’ exhibiting nonlinear behaviour [29]. Secondly, optimizing one aspect of ‘socio-technical’ over the other does not necessarily lead to better performance; rather, researchers have found that improving one without the other leads to a more detrimental performance [29]. Therefore, socio-technical theory is based around the joint optimization of both the social and the technical.

In the 4th Industrial Age (Industry 4.0), a vast majority of organizations blend a mix of those two factors, thereby becoming an embodiment of socio-technical theory: a socio-technical system [39]. Socio-technical system (STS) is grounded in general systems theory, which illustrates how all disciplines have a part to play in a system, and no one discipline can have a monopoly over any system [40]. Specifically, the concept of open systems [41] describes how joint optimization should be kept in

mind when designing any system, so as to embody the characteristics of both the internal and external environments. This open system concept underlines the idea that organizations that adopt such a policy are better positioned to cope with shifting dynamics, environments, competition and technology [40].

*Corporate governance.* For organizations operating in complex and highly dynamic environments, the significance of effective security governance (how decisions are reached) and management (what types of decisions are made) cannot be understated [42]. Corporate governance draws its roots from agency theory—which is concerned about the relationship between two parties: the principal (the party that assigns tasks) and the agent (the party that executes the task [40]). The main problem, the ‘principal-agent problem’, arises when the interests of the principal and the agent are not aligned—leading to sources of conflicts as the agent does not act in the principal’s best interests [43]. Corporate governance draws on agency theory to design incentives or deterrents for the agent, in order to encourage them to act in accordance with the principal’s interests. Without the appropriate incentives or deterrents in place, the agent is more likely to act in a manner that places their own interests above the principal’s [40].

Take for instance an organizational setting where the board (the principal) assigns tasks and responsibilities to management (the agent)—in this case, perhaps policy implementation. Management would then proceed to execute the policy the best way they could through their expertise (for instance, drawing up contracts and attending courses). If the board had no knowledge of the measures management used, they would not be able to verify that management was acting in their best interests. Monitoring and measuring provide a board with information about what is currently taking place in terms of its IT strategic direction in the organization. This information also enables the board to become comprehensive in terms of its own and management’s responsibilities to keep IT aligned with the business goals.

## 5 A New Perspective: Multilevel Governance in ISG and Future Implications

*ISG as a multilevel issue.* Issues of ISG and cybersecurity are increasingly becoming a multilevel issue. At an international level, the United Nations (UN) has set up a dedicated cyber open-ended working group (OEWG) to address state-level IT concerns, as well as the International Telecommunications Union (ITU) to deal with all matters related to information and communication technologies [44]. At a national level, countries use a variety of different frameworks to guide their policies: for instance, the USA has the Sarbanes–Oxley Act of 2002, the European Union has the General Data Protection Regulation (GDPR), and the UK has the Data Protection Act of 2018. At the industry and organizational level, there is the ISO/IEC 27000 series of standards. At the individual level, operational responsibility and accountability

lie with middle to low levels of management, whilst upper-level management must support and prioritize IT security [45].

*Limitations of current frameworks.* The corporate governance framework largely frames and examines ISG from the internal perspective of an organization. Corporate governance was designed to study the relationship between two parties: broadly classified as the managing bodies and instruments that help safeguard the interests of the organization's stakeholders [22]. This approach largely emphasizes technical controls and aspects, with little attention being paid to the environment and context that the organization is embedded in [14]. In today's rapidly changing environments, this traditional view of corporate governance as a risk management and internal control mechanism, coupled with its roots in agency theory studying just the principal and the agent, is potentially ill-equipped to tackle ISG from a holistic and multilevel perspective.

Whilst the socio-technical perspective strives to adopt a more encompassing approach towards ISG by incorporating the human element, researchers have found that this approach does not take into account ISG at an organizational level [33]. For example, Ruighaver et al. [33] found that organizations that require high levels of security are more risk-averse and prefer stability over change, whilst organizations that require or have lower levels of security are more open and tolerant of change. In the ever-evolving cyber landscape, these organizational characteristics and relationships are important factors when considering ISG approaches and are often excluded from the socio-technical perspectives.

*Multilevel governance.* Multilevel governance (MLG), as the name suggests, comprises of actors located at various levels—for instance, the local (sub-national), the national and the global (supranational). MLG aims to capture and understand processes that span across these three levels of governance. As previously established, cybersecurity and information security are growing concerns for stakeholders across all levels (supranational with the UN, national with governments, and sub-national with organizations). In order to achieve the collective goals of the actors, alignment must be achieved across all levels. Within the MLG literature, two types of MLG have emerged: the first, a type of governance with a vertical hierarchy and a clear structure. In this form of MLG, because power and decision-making are concentrated amongst a limited number of individuals [46], the MLG analysis largely focuses on the interactions between the different levels of governance and the outcomes of the policies. The second type of MLG, often referred to as 'polycentric', contrasts the first hierarchical model with a more evenly spread out spread of power and decision-making. In this model, MLG operates within 'spheres of authority' [47], or 'complex overlapping networks' [48] where collaboration is utilized in order to discuss and negotiate policies and decisions.

A multilevel perspective has been utilized as a framework to analyse socio-technical issues before, specifically the issue of sustainability [49]. In particular, Geels [49] highlights that a multilevel perspective can provide a suitable canvas on which to study transitional phases by observing and identifying dynamic patterns. This is particularly relevant for the issue of ISG and cybersecurity, as the ecosystem shifts from the traditional physical world to a largely cyber and digital environment.

Within the context of ISG and IT security, multilevel governance exhibits many of the characteristics suitable for application in future studies. As previously mentioned, the issue of IT security and ISG is now increasingly becoming a multilevel issue, with supranational (UN) national and sub-national actors all playing a role in the state of ISG and IT security. At present, the dominant theories of corporate governance and STS do not address the full complexities of the multilevel issues ISG, and IT security has become. Because of the way the central actors of governance and IT security are increasingly shaped by and operated at multiple levels, MLG is well-positioned to address these gaps in the literature.

## References

1. Soomro, Z.A., Shah, M.H., Ahmed, J.: Information security management needs more holistic approach: a literature review. *Int. J. Inf. Manag.* **36**(2), 215–225 (2016). <https://doi.org/10.1016/j.ijinfomgt.2015.11.009>
2. Horne, C.A., Maynard, S.B., Ahmad, A.: Organisational information security strategy: review, discussion and future research. *Australas. J. Inf. Syst.* **21** (2017)
3. Mugunthan, S.R.: Soft computing based autonomous low rate DDOS attack detection and security for cloud computing. *J. Soft Comput. Paradigm* **2019**(2), 80–90 (2019). <https://doi.org/10.36548/jscp.2019.2.003>
4. Samuel Manoharan, J.: A novel user layer cloud security model based on chaotic arnold transformation using fingerprint biometric traits. *J. Innov. Image Process.* **3**(1), 36–51. <https://doi.org/10.36548/jiip.2021.1.004>
5. Higgs, J.L., Pinsky, R.E., Smith, T.J., Young, G.R.: The relationship between board-level technology committees and reported security breaches. *J. Inf. Syst.* **30**(3), 79–98 (2016). <https://doi.org/10.2308/isys-51402>
6. Kauspadiene, L., et al.: High-level self-sustaining information security management framework. *Balt. J. Mod. Comput.* **5**(1), 107–123 (2017). <https://doi.org/10.2236/bjmc.2017.5.1.07>
7. Shakya, S.: An efficient security framework for data migration in a cloud computing environment. *J. Artif. Intell. Capsule Netw.* **01**(01), 45–53 (2019). <https://doi.org/10.36548/jaicn.2019.1.006>
8. Veiga, A.D., Eloff, J.H.P.: An information security governance framework. *Inf. Syst. Manag.* **24**(4), 361–372 (2007). <https://doi.org/10.1080/10580530701586136>
9. Nicho, M.: A process model for implementing information systems security governance. *Inf. Comput. Secur.* **26**(1), 10–38 (2018). <https://doi.org/10.1108/ICS-07-2016-0061>
10. Kayworth, T., Whitten, D.: Effective information security requires a balance of social and technology factors. Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2058035, May 2012. Accessed: 8 Oct 2021. [Online]. Available: <https://papers.ssrn.com/abstract=2058035>
11. Knapp, K.J., Franklin Morris, R., Marshall, T.E., Byrd, T.A.: Information security policy: an organizational-level process model. *Comput. Secur.* **28**(7), 493–508 (2009). <https://doi.org/10.1016/j.cose.2009.07.001>
12. McFadzean, E., Ezingeard, J., Birchall, D.: Perception of risk and the strategic impact of existing IT on information security strategy at board level. *Online Inf. Rev.* **31**(5), 622–660 (2007). <https://doi.org/10.1108/14684520710832333>
13. Holgate, J., Williams, S., Hardy, C.: Information security governance: Investigating diversity in critical infrastructure organizations. In: BLED 2012 Proceedings, June 2012. <https://aisnet.org/bled2012/13>

14. Williams, S.P., Hardy, C.A., Holgate, J.A.: Information security governance practices in critical infrastructure organizations: a socio-technical and institutional logic perspective. *Electron. Mark.* **23**(4), 341–354 (2013). <https://doi.org/10.1007/s12525-013-0137-3>
15. von Solms, B.: Information security—the fourth wave. *Comput. Secur.* **25**(3), 165–168 (2006). <https://doi.org/10.1016/j.cose.2006.03.004>
16. Maynard, S., Tan, T., Ahmad, A., Ruighaver, T.: Towards a framework for strategic security context in information security governance. *Pac. Asia J. Assoc. Inf. Syst.* **10**(4) (2018). <https://doi.org/10.17705/1pais.10403>
17. Tan, T., Maynard, S., Ahmad, A., Ruighaver, T.: information security governance: a case study of the strategic context of information security. *PACIS 2017 Proceedings*, July 2017, [Online]. Available: <https://aisel.aisnet.org/pacis2017/43>
18. Haufe, K., Colomo-Palacios, R.: A process framework for information security management. *IJISPM Int. J. Inf. Syst. Proj. Manag.* **4**, 27–47 (2016). <https://doi.org/10.12821/ijispdm040402>
19. AICPA: SOC for cybersecurity. In: AICPA (2017). <https://www.aicpa.org/interestareas/frc/assuranceadvisoryservices/aicpacybersecurityinitiative.html>. Accessed 9 Oct 2021
20. Rocha Flores, W., Antonsen, E., Ekstedt, M.: Information security knowledge sharing in organizations: investigating the effect of behavioral information security governance and national culture. *Comput. Secur.* **43**, 90–110 (2014). <https://doi.org/10.1016/j.cose.2014.03.004>
21. Park, H., Kim, S., Lee, H.J.: General drawing of the integrated framework for security governance, pp. 1234–1241 (2006)
22. Posthumus, S., von Solms, R.: A framework for the governance of information security. *Comput. Secur.* **23**(8), 638–646 (2004). <https://doi.org/10.1016/j.cose.2004.10.006>
23. Gillon, K., Branz, L., Culnan, M., Dhillon, G., Hodgkinson, R., MacWillson, A.: Information security and privacy—rethinking governance models. *Commun. Assoc. Inf. Syst.* **28**(1) (2011). <https://doi.org/10.17705/1CAIS.02833>
24. Matwyshyn, A.M.: CSR and the corporate cyborg: ethical corporate information security practices. *J. Bus. Ethics* **88**(4), 579–594 (2009)
25. D'Arcy, J., Hovav, A.: Does one size fit all? Examining the differential effects of IS security countermeasures. *J. Bus. Ethics* **89**(1), 59–71 (2009)
26. Leiwo, J., Heikkuri, S.: An analysis of ethics as foundation of information security in distributed systems, vol. 6, pp. 213–222 (1998)
27. Brey, P.: Ethical aspects of information security and privacy. *Secur. Priv. Trust Mod. Data Manag.*, 21–36 (2007)
28. Bostrom, R.P., Heinen, J.S.: MIS problems and failures: a socio-technical perspective, part II: the application of socio-technical theory. *MIS Q.*, 11–28 (1977)
29. Walker, G.H., Stanton, N.A., Salmon, P.M., Jenkins, D.P.: A review of sociotechnical systems theory: a classic concept for new command and control paradigms. *Theor. Issues Ergon. Sci.* **9**(6), 479–499 (2008). <https://doi.org/10.1080/14639220701635470>
30. Davis, M.C., Challenger, R., Jayewardene, D.N., Clegg, C.W.: Advancing socio-technical systems thinking: a call for bravery. *Appl. Ergon.* **45**(2), 171–180 (2014)
31. Dutta, A., McCrohan, K.: Management's role in information security in a cyber economy. *Calif. Manage. Rev.* **45**(1), 67–87 (2002). <https://doi.org/10.2307/41166154>
32. Maleh, Y., Ezzati, A., Sahid, A., Belaissaoui, M.: CAFISGO: a capability assessment framework for information security governance in organizations. *J. Inf. Assur. Secur.* **12**(6) (2017)
33. Ruighaver, A.B., Maynard, S.B., Chang, S.: Organisational security culture: extending the end-user perspective. *Comput. Secur.* **26**(1), 56–62 (2007). <https://doi.org/10.1016/j.cose.2006.10.008>
34. Thomson, K.-L., von Solms, R.: Information security obedience: a definition. *Comput. Secur.* **24**(1), 69–75 (2005). <https://doi.org/10.1016/j.cose.2004.10.005>
35. von Solms, B., von Solms, R.: From information security to...business security? *Comput. Secur.* **24**(4), 271–273 (2005). <https://doi.org/10.1016/j.cose.2005.04.004>
36. Caralli, R.A., Allen, J.H., Stevens, J.F., Willke, B.J., Wilson, W.R.: Managing for enterprise security. Carnegie-Mellon Univ Pittsburgh PA Software Engineering Inst (2004)

37. Mishra, S.: Organizational objectives for information security governance: a value focused assessment. *Inf. Comput. Secur.* **23**(2), 122–144 (2015). <https://doi.org/10.1108/ICS-02-2014-0016>
38. Carcary, M., Renaud, K., McLaughlin, S., O'Brien, C.: A framework for information security governance and management. *It Prof.* **18**(2), 22–30 (2016)
39. Appelbaum, S.: Socio-technical systems theory: an intervention strategy for organizational development. *Manag. Decis.* **35**, 452–463 (1997). <https://doi.org/10.1108/00251749710173823>
40. Posthumus, S., von Solms, R.: Agency theory: can it be used to strengthen IT governance? In: Jajodia, S., Samarati, P., Cimato, S. (eds.) *Proceedings of the IFIP TC 11 23rd International Information Security Conference*, vol. 278, pp. 687–691. Springer US, Boston, MA (2008). [https://doi.org/10.1007/978-0-387-09699-5\\_46](https://doi.org/10.1007/978-0-387-09699-5_46)
41. Von Bertalanffy, L.: An outline of general system theory. *Br. J. Philos. Sci.* (1950)
42. Peppard, J.: The conundrum of IT management. *Eur. J. Inf. Syst.* **16**(4), 336–345 (2007)
43. Panda, B., Leepsa, N.: Agency theory: review of theory and evidence on problems and perspectives. *Indian J. Corp. Gov.* **10**(1), 74–95 (2017)
44. International Trade Union: Cybersecurity (2010). Accessed: 12 Sept 2021. [Online]. Available: [https://www.itu.int/net/itunews/issues/2010/09/pdf/201009\\_20.pdf](https://www.itu.int/net/itunews/issues/2010/09/pdf/201009_20.pdf)
45. Knapp, K.J., Marshall, T.E., Rainer, R.K., Ford, F.N.: Information security: management's effect on culture and policy. *Inf. Manag. Comput. Secur.* (2006)
46. Fairbrass, J., Jordan, A.: European Union environmental policy and the UK government: a passive observer or a strategic manager? *Environ. Polit.* **10**(2), 1–21 (2001)
47. Rosenau, J.N.: Governing the ungovernable: The challenge of a global disaggregation of authority. *Regul. Gov.* **1**(1), 88–97 (2007). <https://doi.org/10.1111/j.1748-5991.2007.00001.x>
48. Bache, I., Flinders, M.: Multi-level governance and the study of the British state. *Public Policy Adm.* **19**(1), 31–51 (2004)
49. Geels, F.W.: The multi-level perspective on sustainability transitions: responses to seven criticisms. *Environ. Innov. Soc. Transit.* **1**(1), 24–40 (2011). <https://doi.org/10.1016/j.eist.2011.02.002>

# ML-Wasm Entropy and Plot: Dataframes and Plotting Powered by WebAssembly and Rust



Dion Pinto, Arpit Bhat, Immanuel Gnanadurai, and Trupti Lotlikar

**Abstract** This project aims to create a library for data analysis and visualization of datasets. It will be similar in structure to popular data analysis and manipulation libraries in Python so that the library is easily adaptable for new users. This project builds the basic blocks for creation of series and data frames like structures in rust which can be accessed in JavaScript so as to run it on the browser or any other JavaScript environment like NodeJs. This approach would be faster than its pure JavaScript alternative.

**Keywords** Wasm · Series · Dataframe · Line chart · Scatter plot · ML-wasm · Rust

## 1 Introduction

Machine learning is an integral part of society. A fundamental cog in the machine learning wheel is data analysis, feature detection [20], and manipulation along with visualization. However, the current machine learning landscape is dominated by Python and its ML ecosystem consisting of NumPy, Pandas, Matplotlib, and Sklearn. The problem with using Python for this is that they need a Python runtime on the system which is not available on browsers. The aim of this project is to build a data analysis and manipulation tool along with a plotting library that runs in the browser powered by WebAssembly and Rust.

---

D. Pinto (✉) · A. Bhat · I. Gnanadurai · T. Lotlikar

Faculty of Information Technology, Fr. Conceccao Rodrigues Institute of Technology, Navi Mumbai, India

e-mail: [dion.pinto@fcrit.ac.in](mailto:dion.pinto@fcrit.ac.in)

A. Bhat

e-mail: [bhat.arpit@fcrit.ac.in](mailto:bhat.arpit@fcrit.ac.in)

I. Gnanadurai

e-mail: [immanuel.gnanadurai@fcrit.ac.in](mailto:immanuel.gnanadurai@fcrit.ac.in)

T. Lotlikar

e-mail: [trupti.lotlikar@fcrit.ac.in](mailto:trupti.lotlikar@fcrit.ac.in)

## 1.1 What Is ML-Wasm/Entropy

Many organizations now run machine learning models on the browser. The most accessible way to interact with machine learning is through the browser [1] as such using the browser for data preprocessing, and wrangling is the logical next step.

Entropy stands for the degree of disorder, and the ml-wasm/entropy library lets its users reduce the entropy of your data by structuring your data into series and dataframes from where they can manipulate the data, preprocess it, and make it their own.

## 1.2 What Is ML-Wasm/Plot

Visualizing data is another essential step in machine learning. The ml-wasm/plot library lets users import data in JavaScript and utilize Rust methods to render a svg file utilizing the tera [2] templating engine.

## 1.3 Why Ndarray

The ndarray crate provides an n-dimensional container for general elements and for numerics. ml-wasm/entropy utilizes ndarray as its internal data representation. When we considered the creation of a series and dataframe like structures, the options to consider for its internal data representation were a standard Rust vector or a ndarray.

The problem with utilizing a standard vector is we would have no access to fast and efficient operations which we would have to perform on the series and dataframes. The ndarray crate provides various methods which facilitate quick operations. It also provides various additional functionality such as its support with serde [3] which allows us to quickly serialize and deserialize data from JavaScript to Rust and vice versa. Other than that, ndarray also supports rayon [4] which gives us access to parallel iterators and parallel methods. ml-wasm/entropy utilizes ndarrays with the use of a ml-wasm/linalg utility wrapper.

## 1.4 What is Wasm-Bindgen

It is a Rust library that allows wasm and JavaScript to interact with each other. This is the package that lets us write code in Rust and then talk to JavaScript and interact with it. It allows us to import JavaScript functionality into Rust such as DOM manipulation and console logging.

Wasm-bindgen [5] allows the end-user to access methods written in Rust with the advantages of its speed and added memory safety all in JavaScript. As it is compiled into wasm, it also has near-native speed, which is comparable to stock Rust.

This project aims to create a fast library on the web which implements features of data structuring and visualization libraries and runs them at near native speed on the browser. It will build on modules such as data analysis and manipulation, in which we can ingest data and manipulate it, preprocess the datasets, and visualize them all utilizing Rust and WebAssembly for greater performance on the browser.

## 2 Background

Programs which are compiled to machine code generally work faster than those which are interpreted. This is considered true because the step of interpreting a piece of a program and then converting to machine code, would be slower than just converting it to machine code.

### 2.1 Why Rust

For the entire lifetime of programming languages, there has always been the focus on either the speed of the language or the memory safety of the language. A classic example would be programming languages with or without a garbage collector. Languages with inbuilt garbage collector generally tend to be slower example ruby [6] than the likes of those without one like C++ [7]. However, many memory safety issues arise due to the lack of a garbage collector.

Rust however comes up with a solution that is both fast since it does not require a garbage collector but also ensures memory safety with concepts such as ownership and borrowing [8]. However, no solution is perfect and the cost of utilizing Rust is slower compilation time and many compilation errors along with a higher learning curve just to begin using Rust.

For the purposes of this library, however Rust is excellent as it has great support with libraries such as wasm-bindgen and excellent documentation with Rust used together with WebAssembly and tools such as wasm-pack [9].

Senior software engineer Mohit Agarwal outlined his involvement with Rust in his blog [10]. After evaluation of runtime performances of Rust against other popular compiled and interpreted languages like Java and Python, some surprising results were found. Rust is twice as fast when compared to Java but only uses 1% of its memory. When compared with Python, Rust is 150 times faster but it uses the same amount of memory. In a study by IBM [11], it was found that Rust and WebAssembly are nearly 15 times faster than Scala which is conventionally considered a high-performance language. Luca Palmieri published an article [12] claiming that for simple machine tasks Rust is up to 25 times faster than Python.

## 2.2 Why WebAssembly

WebAssembly [13] is like a low-level assembly language which can be run on the browser. It allows code compiled on various languages to be run on the browser at nearly their native speeds.

Thus, for our library we will write the code in Rust and then compile it to WebAssembly, often abbreviated as wasm. Also, since the program will be compiled, the compiler can perform various compile time optimizations. Since we are just writing Rust programs, we can also utilize various features that it provides most notably memory safety without negatively impacting performance.

Every machine learning library requires a tool for data analysis and manipulation, the series and dataframe implementation in ml-wasm/entropy makes it simple to do many time-consuming and repetitive tasks on data with ml-wasm/plot allowing easy visualization of the data.

## 3 Existing System

### 3.1 Pandas

Pandas [14] is one of the foundational libraries for data analysis and tools for working with structured data in Python. It is commonly used in statistics, finance, social science, and other fields. It offers various data structures and operations for manipulating data and time series. It strongly complements the existing Python stack while improving upon the tools provided by other statistical languages like R. Pandas is fast, and it has high-performance and productivity for users.

### 3.2 Pandas-Js

Pandas-js [15] is an open-source library mimicking the pandas library from Python in JavaScript. Similar to how the pandas library is built on top of the NumPy library, the Pandas-js library is built on immutable-js. The main data objects in pandas-js are the Series and the Dataframe. It implements many of the numerical analysis methods and reshaping operations from Python pandas.

### 3.3 Danfo-Js

Danfo.js [16] is a fast and flexible JavaScript package that provides expressive data structures to work with “relational” or “labeled” data both easy and intuitive. It is

built on TensorFlow.js. It aims to bring ML to the web. Since it is similar to the Python pandas library, it is easy to pick up.

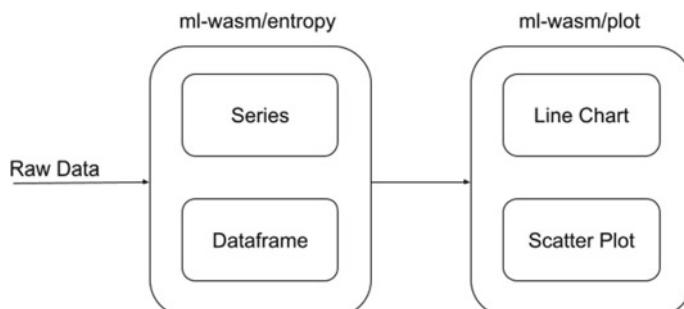
### 3.4 Dataframe-Js

Dataframe-js [17] is an immutable data structure in JavaScript which allows it to work on rows and columns with a SQL and functional programming-inspired API. This library simplifies complex tasks like join, group by, and machine learning. It is designed to work at the server side but can also be used in the browser without the file system-related features.

### 3.5 D3

D3 [18] is a dynamic and interactive JavaScript library for visualizing data using web standards. It uses SVG, Canvas, HTML5, and CSS standards to bring data visualization to the browser. It follows a data-driven approach to DOM manipulation, providing freedom to design the right visual interface for the data.

Almost all of the above systems utilize JavaScript with pandas utilizing Python for its implementation which are interpreted languages that tend to be slower than the approach explained for our library which is to utilize Rust and WebAssembly for greater performance.



**Fig. 1** Flow diagram of entropy and plot

## 4 Implementation

Figure 1 describes the way in which data is processed by the libraries; initially, the user can create the raw data or provide csv data which is then passed to ml-wasm/entropy library which would facilitate in structuring the data in a series or a dataframe and help manipulating it by utilizing various methods listed in Tables 1, 2, 3, and 4. After preprocessing, the data can be visualized via the ml-wasm/plot library as shown in Figs. 2 and 3.

### 4.1 Series

A series consists of a column name and the data array, and the array is represented as an ndarray utilizing the implementation of ml-wasm/linalg wrapper and can have three possible types: Integer 32bit, Float 64bit, or String. Multiple methods can be utilized on the series depending on the use case of the developer.

**Table 1** List of methods for integer and float

Basic methods	Description
Get	Get at index
Set	Set at index
Swap	Swap values between index
Reverse	Reverse the series
Reversed	Reverse the series (with return)
Append	Append to series
Appended	Append to series (with return)
Extend	Extend data in series
Extended	Extend data in series (with return)
Insert	Index put value push rest by one
Inserted	Index put value push rest by one (with return)
Splice	Splice remove the value at the specified index
Spliced	Splice remove the value at the specified index (with return)
Len	Length of series
Dtype	Data type of series
Shape	Shape of series
ToString	Metadata of series
NewWithElement	Construct with specified length
NewWithSimpleFunc	Construct with specified function

**Table 2** Math methods for integer and float

Math methods	Description
Sum	Sum of series elements
Product	Product of series elements
Mean	Mean of series elements
Median	Median of series elements
Max	Max of series elements
Min	Min of series elements

**Table 3** List of methods for string

Methods	Description
Get	Get at index
Set	Set at index
Swap	Swap values between index
Reverse	Reverse the series
Reversed	Reverse the series (with return)
Append	Append to series
Appended	Appended to series (with return)
Extend	Extend data in series
Extended	Extend data in series (with return)
Insert	Index put value push rest by one
Inserted	Index put value push rest by one (with return)
Splice	Splice remove the value at the specified index
Spliced	Splice remove the value at the specified index (with return)
Len	Length of series
Dtype	Data type of series

## 4.2 Dataframe

A Dataframe is a combination of multiple series of different types. It forms a table structure which can be used for many methods to sanitize/manipulate data.

**Table 4** List of methods for dataframe

Methods	Description
Readcsv	Construct dataframe with CSV
Size	Size of dataframe
Columns	Display column names
Dtypes	Data types of columns
Display	String output of dataframe
Loc	Get series with columns name
Iloc	Get series of dataframe
Ilocr	Get dataframe row
Ilocc	Get dataframe column
Head	Get top five rows
Tail	Get bottom five rows
Min	Min of all values
Max	Max of all values
Median	Median of all values

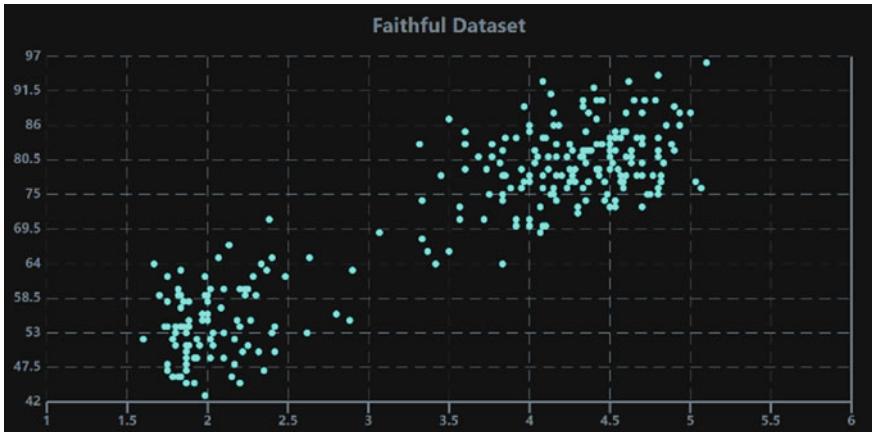
### 4.3 Plot

A library to plot charts with rust and WebAssembly using tera [2]. It is a way to draw svg graphs with rust. It uses an svg templating engine along with rust and WebAssembly to plot data as a line chart or scatter plot.

Figure 2 displays a line chart created by ml-wasm/plot line method. The dataset used is a series of random x-y-coordinates.



**Fig. 2** Line chart using ml-wasm/plot



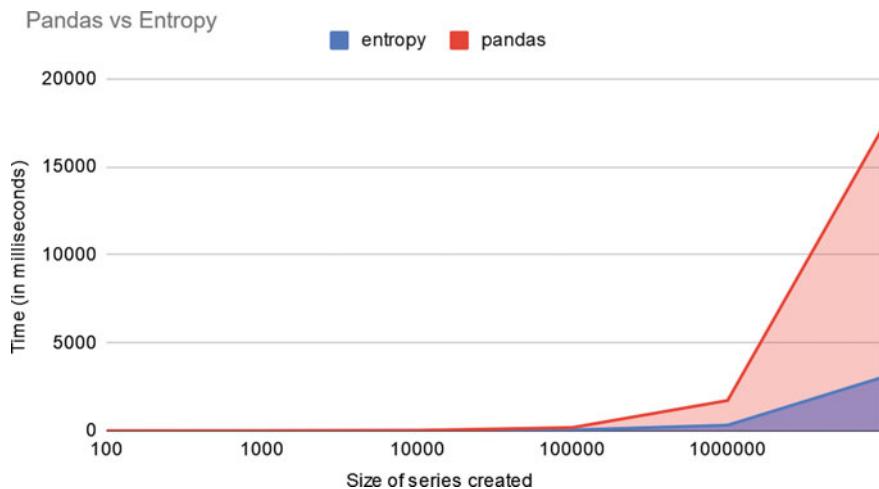
**Fig. 3** Scatter plot using ml-wasm/plot

Figure 3 displays a scatter plot created by ml-wasm/plot scatter method. The dataset used for the following demonstration is the Old faithful geyser dataset [19] which contains the eruption duration and the waiting time in between eruptions.

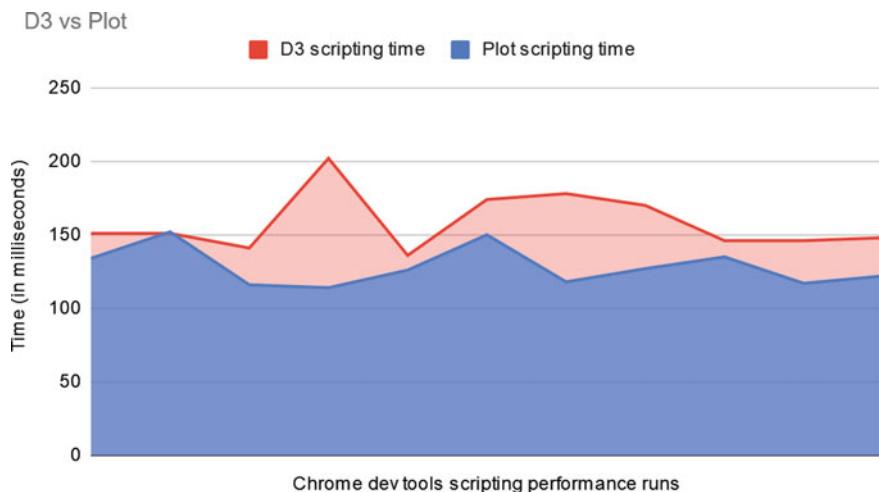
## 5 Results

To test the performance for creation of series, we compare ml-wasm/entropy SeriesI32 to Pandas Series. A random number between 1 and 9 is inserted into the data structure of each series whose size ranges from 100 to 10,000,000. After calculating the time, it took for the creation of the series, it could be seen that our libraries implementation of series is considerably quicker than that of Pandas library. Figure 4 plots the size of the series to the time taken for creation of the series. This result can be attributed to the fact that in our library the underlying data structure of the series is a ndarray which is statically typed array.

To measure the performance of our plotting library, we compared it to D3 JavaScript graphing library. We particularly focused on the scripting time parameter in the chrome developer tools which specifies the JavaScript execution time. Using it, the average scripting time for our ml-wasm/plot library was found to be 19% faster than that of D3 library which is shown in Fig. 5.



**Fig. 4** Time taken to create series



**Fig. 5** Comparison of scripting times

## 6 Conclusion and Future Scope

The trend of moving machine learning to JavaScript [1] can be witnessed in many of the applications developed today. Our library will help manipulate and visualize the data needed for machine learning which can be used to train various models [21] while providing excellent speed with the added memory safety as a by-product of utilizing Rust and WebAssembly.

In this paper, we have shown the creation of structures such as series and dataframes to read, manipulate, process data with ml-wasm/entropy, the benefits of using such an approach, and the speed advantage that can be achieved. Along with that, the paper also shows the visualization of data with ml-wasm/plot.

In the future, ml-wasm will continue to build out the various libraries on aspects of machine learning with the ultimate goal of creating a machine learning framework which would run on JavaScript with speeds comparable to compiled languages.

## References

1. Smilkov, D., Thorat, N., Assogba, Y., Yuan, A., Kreeger, N., Yu, P., Zhang, K., Cai, S., Nielsen, E., Soergel, D., Bileschi, S., Terry, M., Nicholson, C., Gupta, S.N., Sirajuddin, S., Sculley, D., Monga, R., Corrado, G., Viégas, F.B., Wattenberg, M.: TensorFlow.js: machine learning for the web and beyond. arXiv:1901.05350[cs], Feb 2019
2. Tera, <https://crates.io/crates/tera>. Last accessed 2021/3/2
3. Serde, <https://serde.rs/>. Last accessed 2021/3/7
4. Rayon, <https://docs.rs/rayon/1.5.1/rayon/>. Last accessed 21 May 2021
5. Wasm-bindgen, <https://github.com/rustwasm/wasm-bindgen>. Last accessed 12 Feb 2021
6. Ruby, <https://www.ruby-lang.org/en/>. Last accessed 6 Apr 2021
7. Stroustrup, B.: An Overview of the C++ Programming Language (1998). <https://doi.org/10.1201/9780849331350.sec3>
8. Xu, H., Chen, Z., Sun, M., Zhou, M., Lyu, M.: Memory-Safety Challenge Considered Solved? An In-Depth Study with All Rust CVEs. arXiv:2003.03296[cs], Feb 2021
9. Rustwasm, <https://github.com/rustwasm/wasm-pack>. Last accessed 12 Feb 2021
10. Bitbucket, <https://bitbucket.org/blog/why-rust>. Last accessed 12 March 2021
11. IBM study, <https://developer.ibm.com/articles/why-webassembly-and-rust-together-improve-nodejs-performance/>. Last accessed 30 March 2021
12. Taking ML production with Rust, <https://www.secondstate.io/articles/performance-rust-wasm/>. Last accessed 11 Feb 2021
13. WebAssembly, <https://webassembly.org>. Last accessed 10 Feb 2021
14. Pandas, <https://pandas.pydata.org/>. Last accessed 19 Jan 2021
15. Pandas-js, <https://github.com/StratoDem/pandas-js>. Last accessed 2 May 2021
16. Danfo-js, <https://github.com/opensource9ja/danfojs>. Last accessed 17 April 2021
17. Dataframe-js, <https://github.com/Gmousse/dataframe-js>. Last accessed 1 May 2021
18. D3 JavaScript Library, <https://github.com/d3/d3>. Last accessed 7 June 2021
19. Kaggle Old faithful geyser dataset, <https://www.kaggle.com/niteshhalai/old-faithful-data-visualisation-and-modelling/data>. Last accessed 31 Aug 2021
20. Manoharan, S.: Study on Hermitian graph wavelets in feature detection. J. Soft Comput. Paradigm (JSCP) **1**(01), 24–32 (2019)
21. Vijesh, J.C., Raj, J.S.: Location-based orientation context dependent recommender system for users. J. Trends Comput. Sci. Smart Technol. (TCSST) **3**(01), 14–23 (2021)

# Multiclass Hierarchical Fuzzy Classification on Multi-labeled Data



R. Kanagaraj, N. Rajkumar, K. Srinivasan, and E. Elakiya

**Abstract** Multiclass classification is dissimilar from binary classification based on the number of output classes. It can be implemented by using subsequent methods such as extended from the binary case, translate the multiclass into binary-class classification problems and hierarchical classification methods. In this paper, multiclass hierarchical fuzzy rule-based classification tree for using quantitative transactions for correlated pattern mining is proposed. It uses different fuzzy membership linguistic operator values for each fuzzy data item to generate fuzzy classification patterns from multi-labeled data set. The result of experiments shows that the MHFRB mining algorithm is used to generate fuzzy classification rules and efficiently classify infrequent patterns from multi-labeled data.

**Keywords** Fuzzy classification · MHFRB tree · Hierarchical classification · Fuzzy membership

## 1 Introduction

Machine learning techniques are flexible in multi-label classification and clustering. Fuzzy logic, artificial neural networks, and genetic algorithms play a significant role in multiclass classification, and clustering are helpful in analyzing the data. Construction of the HFRBCS has been implemented in hierarchical manner with

---

R. Kanagaraj (✉) · K. Srinivasan  
Sri Ramakrishna Engineering College, Coimbatore 641022, India  
e-mail: [kanagaraj.r@srec.ac.in](mailto:kanagaraj.r@srec.ac.in)

K. Srinivasan  
e-mail: [hod-eie@srec.ac.in](mailto:hod-eie@srec.ac.in)

N. Rajkumar  
KGISL Institute of Technology, Coimbatore 641035, India

E. Elakiya  
National Institute of Technology, Tiruchirappalli, Tiruchirappalli 620015, India  
e-mail: [elakiya@nitt.edu](mailto:elakiya@nitt.edu)

generic rule-based selection process to obtain efficient model [1]. The classification rules are generated from the input training data by fuzzy classification. Hence, a lot of tree-based algorithms are developed for mining classification rules and correlated patterns in the domain of data mining. Most of the developed algorithms are dependent upon certain interesting measures for generating multiple classes which paved the way for multiclass problem. Infrequent class problem in supervised learning classification is solved from the given different output classes. A single-level classification problem is not significant to solve multi-label classification problem in a training data efficiently. To overcome single-label classification problem, alternative interesting measures like linguistic variable, membership function, fuzzy operator, etc., are used. Each measure has a collection bias that validates the importance of the classification rule or pattern mined. In this paper, multiclass hierarchical fuzzy rule-based classification tree (MHFRB) is constructed to mine fuzzy correlated patterns and classification rules. The proposed approach first constructs a MHFRB tree using fuzzy linguistic operator value and generates correlated patterns and classification rules. Multi-level hierarchical fuzzy classification technique is significant classification rules at various levels of taxonomy using interesting measures like fuzzy membership value.

## 2 Related Work

An evolutionary improving algorithm for fuzzy classifier in a hierarchical manner has been proposed to increase the accuracy of classifiers which make use of ordered structure for obtaining fuzzy rules and classifications approaches on a set of well-known classification model has been developed [2]. Fuzzy k-nearest neighbor classification model has been implemented for multi-label learning, and it is progressively more necessary for multi-label classification using the eristic variable model [3]. Fuzzy rule-based classifiers with more than one label is allocated simultaneously to a known samples. A multi-label classification problem is converted into a different single-label classification problems by using different approaches in fuzzy classification model [4].

The output fuzzy rules of conventional rules are arranged in an unordered rule set by RIPPER is enhancement of most popular state-of-the-art rule learner by [5]. Despite of different measures and work done, there is no formal definition for interestingness in mining association rules. Multiple minimum support threshold is allocated to every fuzzy region. In the same way, fuzzy correlated frequent and infrequent mining patterns are extracted by allocating various confidence threshold values to every incoherent region. Recent surveys of interesting measures are given based on their applicability and novelty [6–10].

### 3 MHFRB Tree Construction Algorithm

A single-level classification problem is not significant to solve multi-label classification problem and also not handled the frequent item set in transaction database successfully. To overcome single-label classification problem, alternative interesting measures like linguistic variable, membership function, fuzzy operator, etc., are used. Each measure has a collection bias that validates the importance of the classification rule or pattern mined. An additional linguistic modifiers operators have been used in Fuzzy set is illustrated in Eq. 3.1 which is used to categories the output classes into “high,” “very high,” “low,” “little,” “less,” etc. (Table 1).

$$\mu_{A_i, \text{comp}}(x) = \prod_{i=1}^m \mu_i(x)^{(1-\gamma)} \left( 1 - \prod_{i=1}^m \mu_i(x)^{(1-\gamma)} \right)^\gamma \quad (3.1)$$

INPUT:  $n$  training data set and membership operator  $\gamma$ .

OUTPUT: a constructed MFHRB tree.

1. Initialize  $\gamma = 0.5$
2. To compute the fuzzy membership degree RM ( $A, C_i$ )
3. For each  $V_j^{(i)}$  of  $I_j$  in  $D^i$  do

$$\mu_{A_i, \text{comp}}(x) = \prod_{i=1}^m \mu_i(x)^{(1-\gamma)} \left( 1 - \prod_{i=1}^m \mu_i(x)^{(1-\gamma)} \right)^\gamma$$

4. End for
5.  $\text{Card}(A|C_j) = \sum_{i=1}^n \mu_{A_i, \text{comp}}(x)$
6. To compute final fuzzy membership FM
7. For each  $C_i$  in  $D^{(i)}$  do
8.  $\text{FM}(A, C_i) = \text{RM}(A, C_i)/\text{Card}(A|C_j)$
9. End for

**Table 1** Notations in MHFRB algorithm

Notations	Description
$N$	Number of transactions in D
$\Gamma$	Membership operator $\gamma = 0.5$
$D^i$	Data object, $i = 1$ to $n$
$I_j$	The $j$ th item, $j = 1$ to $m$
$\text{RM}(A_i, C_i)$	Raw fuzzy membership value
$\text{FM}(A_i, C_i)$	Final fuzzy membership value
$C_i$	Multiclass labels
$\mu_{A_i, \text{comp}}$	Linguistic variable of attribute $A_i$
$\text{Card}(A C_j)$	Cardinality of class

10. Multiclass hierarchical fuzzy rule-based classification tree
11. Begin
12. For each  $V$  in  $k$  data times
13. Construct a output tree
14. Compute  $S_i$  of each iteration
15. Update weight value
16. End for
17. Classification model is constructed based on fuzzy final degree membership
18. End

### 3.1 An Example

Sample database for MHFRB tree mining is given to demonstrate the construction of MHFRB tree from input training. The training input database given in Table 2 is used for demonstrating the example.

From the input database in Table 2, turnover of the values ranges between [0, 499] and [500–1000]. Outcome of the multi-labeled data is classified into {Bad, Satisfactory} and {Good, Outstanding}.

The input database consists of eight transactions and three items (Item1, Item2, and Item3). Initialize the fuzzy membership operator  $\gamma$  which is equal to 0.5. The quantity corresponds to in the input data for each item is labelled using two classifications high, low and linguistic variable behavior with its terms of *ok* and *not ok* with four individual class  $C_1$ ,  $C_2$ ,  $C_3$ , and  $C_4$ . Table 3 illustrates the fuzzy sets of the individual

**Table 2** Sample database for MHFRB tree mining

Customer	List of items	Turnover	Outcome
A	Item3	1000	Outstanding
A	Item2	650	Good
B	Item1	400	Satisfactory
B	Item3	450	Satisfactory
C	Item2	100	Bad
C	Item3	300	Satisfactory
D	Item1	60	Bad
D	Item2	800	Outstanding

**Table 3** Individual target class for MHFRB tree mining

Customer	Turnover	Outcome	Class
A	{590}	{Good}	$C_2$
B	{1000}	{Outstanding}	$C_1$
C	{60}	{Bad}	$C_4$
D	{410}	{Satisfactory}	$C_3$

**Table 4** Contingency table for two items

	$y$	$\neg y$
$x$	$C_1$	$C_2$
$\neg x$	$C_3$	$C_4$

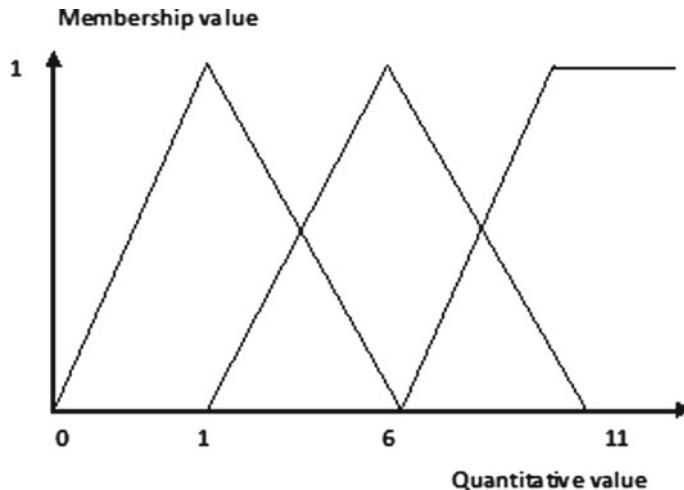
**Table 5** Raw fuzzy membership of multiclassess of customer A

Customer	Fuzzy membership value	Class
A	$\begin{aligned} \text{RM}(A, C_1) &= (\mu_{\text{high}}(590) * (\mu_{\text{Ok}}(\text{good}))^{0.5}) * (1 - (1 - \mu_{\text{high}}(590) * (1 - \mu_{\text{Ok}}(\text{good})))^{0.5}) \\ &= ((0.56) * (0.70))^{0.5} * (1 - (0.44) * (0.30)^{0.5}) \\ &= 0.626 * 0.931 \approx 0.582 \end{aligned}$	$C_1$
	$\begin{aligned} \text{RM}(A, C_2) &= (\mu_{\text{high}}(590) * (\mu_{\text{NotOk}}(\text{good}))^{0.5}) * (1 - (1 - \mu_{\text{high}}(590) * (1 - \mu_{\text{NotOk}}(\text{good})))^{0.5}) \\ &= ((0.56) * (0.30))^{0.5} * (1 - (0.44) * (0.70)^{0.5}) \\ &= 0.409 * 0.831 \approx 0.340 \end{aligned}$	$C_2$
	$\begin{aligned} \text{RM}(A, C_3) &= (\mu_{\text{low}}(590) * (\mu_{\text{Ok}}(\text{good}))^{0.5}) * (1 - (1 - \mu_{\text{low}}(590) * (1 - \mu_{\text{Ok}}(\text{good})))^{0.5}) \\ &= ((0.44) * (0.70))^{0.5} * (1 - (0.56) * (0.30)^{0.5}) \\ &= 0.554 * 0.912 \approx 0.505 \end{aligned}$	$C_3$
	$\begin{aligned} \text{RM}(A, C_4) &= (\mu_{\text{low}}(590) * (\mu_{\text{Notok}}(\text{good}))^{0.5}) * (1 - (1 - \mu_{\text{low}}(590) * (1 - \mu_{\text{Ok}}(\text{good})))^{0.5}) \\ &= ((0.44) * (0.30))^{0.5} * (1 - (0.56) * (0.70)^{0.5}) \\ &= 0.363 * 0.779 \approx 0.283 \end{aligned}$	$C_4$

target class for MHFRB tree mining in the given input database. Fuzzy membership value of customer A for class  $C_1$  is calculated using Eq. (3.1). Table 4 shows the contingency table for fuzzy membership function which is calculated within the fuzzy region. The value of fuzzy raw membership for four classes of customer A is  $C_1 \approx 0.582$ ,  $C_2 \approx 0.340$ ,  $C_3 \approx 0.505$ , and  $C_4 \approx 0.283$  which are illustrated in Table 5. Similarly, fuzzy membership value of customer B, C, and D has been calculated based on Eq. (3.1). The sum of membership value for each fuzzy region, i.e., customer A, is calculated as  $(0.582 + 0.304 + 0.505 + 0.283) (=1.71)$ .

In single-class classification problem, the elements are evaluated by using the fuzzy membership degrees. For multi classes, pre-processing the input samples in which the values of fuzzy membership is normalized, such that sum of all class value is equal to 1 based on the probabilistic fuzzy model. For customer A, the total value fitting to the equivalence class of  $A$  from  $C_1$ ,  $C_2$ ,  $C_3$ , and  $C_4$  is calculated using Eq. (3.2) and final membership value is calculated using Eq. (3.3). A membership function used in fuzzy classification is shown in Fig. 1 which applies to all the data items in the database.

$$\text{Card}(A|C_i) = \sum_{j=1}^n \text{RM}(A, C_j) \quad (3.2)$$



**Fig. 1** Membership functions used for MHFRB tree mining

where  $A$  is attribute,  $C_j$  is Class, and  $n$  is the number of classes.

$$FM(A, C_j) = \frac{RM(A, C_j)}{\text{Card}(A|C_i)}, \quad 0 \leq FM(A, C_j) \leq 1 \quad (3.3)$$

$$\begin{aligned} \text{Card}(A|C_i) &= RM(A, C_1) + RM(A, C_2) + RM(A, C_3) + RM(A, C_4) \\ &= 0.582 + 0.304 + 0.505 + 0.283 \\ &\approx 1.71 \end{aligned}$$

A final membership degree of fuzzy classification is calculated for the following multiple classes. Then, the final membership value is extracted from standardized input data from the four different classes if the cardinality value of particular class  $A$  is greater than one. The cardinality value of class  $A$  is equal to 1.71 using Eq. (3.2), so we need to normalize the value using Eq. (3.3). The sum of final membership value for each fuzzy region, i.e., customer A, is calculated as  $(0.340 + 0.198 + 0.295 + 0.165) = (1.00)$ . The following sum of fuzzy values equal to 1.00.

$$\begin{aligned} FM(A, C1) &= RM(A, C_1)/\text{Card}(A|C_i) \approx 0.340 \\ FM(A, C2) &= RM(A, C_2)/\text{Card}(A|C_i) \approx 0.198 \\ FM(A, C3) &= RM(A, C_3)/\text{Card}(A|C_i) \approx 0.295 \\ FM(A, C4) &= RM(A, C_4)/\text{Card}(A|C_i) \approx 0.165 \end{aligned}$$

Fuzzy classification technique has been used membership operator value to classify the given input database into different multiple classes with in the fuzzy region. The total value of the classes is within the range from 0 to 1. Based on the turnover

**Table 6** Multiclasses results table

Input attributes			Multiclass			
Customer	Turnover	Outcome	$C_1$	$C_2$	$C_3$	$C_4$
A	{590}	{Good}	0.340	0.198	0.295	0.165
B	{1000}	{Outstanding}	1	0	0	0
C	{60}	{Bad}	0	0	0	1
D	{410}	{Satisfactory}	0.165	0.295	0.198	0.340

**Table 7** Fuzzy membership value of multiclasses

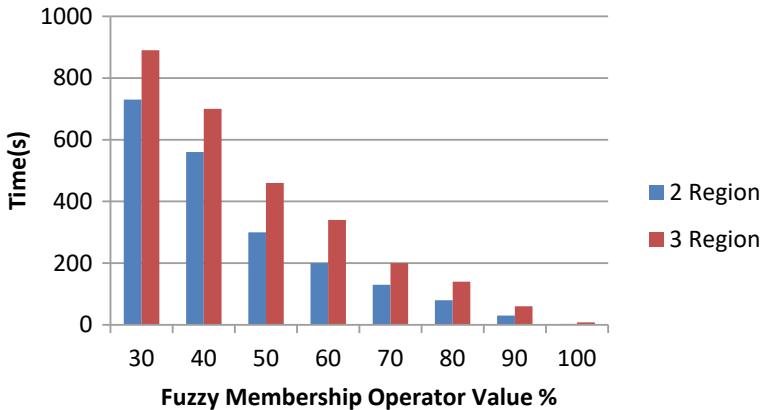
Customer	Outcome	Behavior	Class
{A, 0.340}	High	OK	$C_1$
{A, 0.398}	High	Not Ok	$C_2$
{A, 0.295}	Low	Ok	$C_3$
{A, 0.165}	Low	Not Ok	$C_4$
{B, 1.0}	High	OK	$C_1$
{C, 1.0}	Low	Not Ok	$C_4$
{D, 0.19}	High	OK	$C_1$
{D, 0.30}	High	Not Ok	$C_2$
{D, 0.20}	Low	Ok	$C_3$
{D, 0.31}	Low	Not Ok	$C_4$

attribute, the customers are classified into different classes from class  $C_1$  to  $C_4$  which are shown in Table 6. The frequent fuzzy regions are ordered with respect to multiple classes of attribute customer and their sums to form table, and the fuzzy region in each customer is demonstrated in Table 7. Using the above Table 7, the resultant tree is given in Fig. 1.

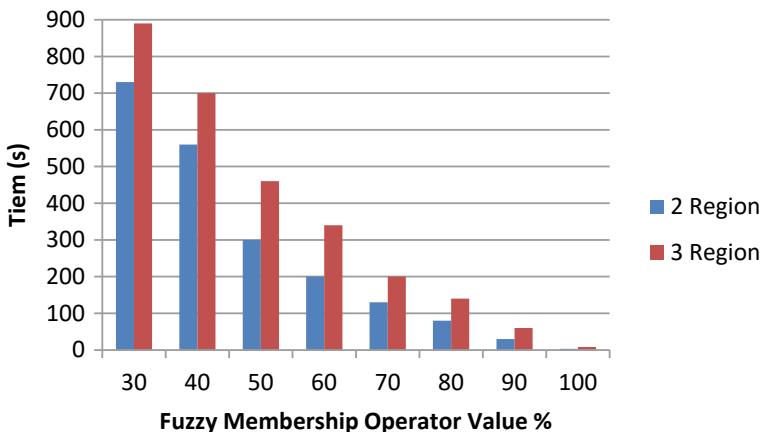
Based on the observation from above table, the customer B belongs to class  $C_1$  and behavior of the class is Ok and customer C belongs to class  $C_4$  and behavior of the class is not Ok.

## 4 Experimental Results

The experiments for the proposed approach are implemented in R Tool. Blood bank dataset and electricity consumption dataset are publicly available at the open data Indian Government Repository. Open Government Data is (OGD) a repository. The value of the fuzzy membership operator  $\gamma$  is set as 0.01 in order to enable the occurrence of all infrequent patterns in the database. Minimum linguistic variable values are experimented at a range from 30 to 100%. Figures 2 and 3 display the execution time taken at different values minimum values of membership operator for electricity



**Fig. 2** Execution time at various membership values for electricity consumption dataset



**Fig. 3** Execution time at various min all-confidence values for breast cancer dataset

consumption and breast cancer datasets. Figures 2 and 3 show that the implementation time is long position for 3 regions of Fuzzy is compared to the 2 regions of the fuzzy. Three regions of fuzzy set would produce more number of tree nodes for multiple classes, for example {high, medium, low}, than two fuzzy regions.

## 5 Conclusions and Future Work

The MHFRB classification mining algorithm has been used to construct multi-level hierarchical in which subsequent mining is processed for storing various linguistic

terms. In the MHFRB tree, each data item has been used to generate the classification rules. The proposed MHFRB algorithm in the Experimental results is used to generates the number correlated patterns is more compared to the decision tree induction algorithm based on classification rules, the number of irregular rules generated is not considered and further neural network classification framework has been used to decrease the number of irregular classification rules. In this proposed technique, the input database for classification algorithm is assumed to be static, and in the future work, database is used dynamically for implementing the process. In the future, the proposed approach is extended to implement the neural network classification technique for reducing the fuzzy irregular rules and also examining the addition of proposed work for developing compact and accurate classification model.

**Acknowledgements** The authors like to thank all the anonymous reviewers for their valuable suggestions and Sri Ramakrishna Engineering College for offering resources for the implementation.

## References

1. Fernández, A., del Jesus, M.J., Herrera, F.: Hierarchical fuzzy rule-based classification systems with genetic rule selection for imbalanced data-sets. *Int. J. Approx. Reason.* **50**, 561–577 (2009)
2. Anuradha, R., Rajkumar, N., Sowmyaa, V.: Multiple fuzzy correlated pattern tree mining with minimum item all-confidence thresholds. In: Proceedings of the Fifth International Conference on Fuzzy and Neuro Computing (FANCCO-2015), pp. 15–29
3. Kanagaraj, Anjana, P., Bavatarani, S. Kumar, D.: A study on human behavior based color psychology using K-means clustering. In: 2020 international conference on inventive computation technologies (ICICT), pp. 608–612 (2020). <https://doi.org/10.1109/ICICT48043.2020.9112442>
4. Amouzadi, A., Mirzaei, A.: Hierarchical fuzzy rule-based classification system by evolutionary boosting algorithm. In: 5th International Symposium on Telecommunications (IST'2010)
5. Smys, S., Raj, J.S.: Analysis of deep learning techniques for early detection of depression on social media network-a comparative study. *J. Trends Comput. Sci. Smart Technol. (TCSST)* **3**(01), 24–39 (2021)
6. Younes, Z., Abdallah, F., Deneux, T.: Fuzzy multi-label learning under veristic variables. In: WCCI 2010 IEEE World Congress on Computational Intelligence, 18–23 July 2010. CCIB, Barcelona, Spain
7. Prati, R.C.: Fuzzy rule classifiers for multi-label classification (2015). <https://doi.org/10.1109/FUZZ-IEEE.2015.7337815>
8. Kanagaraj, R., Rajkumar, N., Srinivasan, K.: Multiclass normalized clustering and classification model for electricity consumption data analysis in machine learning techniques. *J. Ambient Intell. Human Comput.* **12**, 5093–5103 (2021)
9. Huhn, J.C., Hüllermeier, E.: FURIA: an algorithm for unordered fuzzy rule induction. *Data Min. Knowl. Discov.* **19**(3), 293–319 (2009)
10. McGarry, K.: A survey of interestingness measures for knowledge discovery. *Knowl. Eng. Rev.* **20**(1), 39–61 (2005)
11. de Campos, L.M., Moral, S.: Learning rules for a fuzzy inference model. *Fuzzy Sets Syst.* **59**, 247–257 (1993)
12. Hong, T.P., Chen, J.B.: Finding relevant attributes and membership functions. *Fuzzy Sets Syst.* **103**, 389–404 (1999)

13. Hong, T.P., Lee, C.Y.: Induction of fuzzy rules and membership functions from training examples. *Fuzzy Sets Syst.*, 33–47 (1996)
14. Elakiya, E., Kanagaraj, R., Rajkumar, N.: Topic detection using multiple semantic spider hunting algorithm. *Smart Intell. Comput. Commun. Technol.* (2021). <https://doi.org/10.3233/APC210072>
15. Hong, T.P., Lin, C.W., Lin, T.C.: The MFFP-tree fuzzy mining algorithm to discover complete linguistic frequent itemsets. *Comput. Intell.* **30**(1) (2014)
16. Lin, C.W., Hong, T.P., Lu, W.H.: Mining fuzzy association rules based on fuzzy fp-trees. In: The 16th National Conference on Fuzzy Theory and Its Applications, pp. 11–16 (2008)
17. Elakiya, E., Rajkumar, N.: A comprehensive survey on topic and subtopic detection. *Int. J. Appl. Eng. Res.* **10**(51) (2015)
18. Geng, L., Hamilton, H.J.: Interestingness measures for data mining: a survey. *ACM Comput. Surv.* **38**(3), 1–32 (2006)
19. Lallich, S., Teytaud, O., Prudhomme, E.: Association rule interestingness: measure and statistical validation. In: *Quality Measures in Data Mining*, vol. 43, pp. 251–276 (2006)
20. Lenca, P., Meyer, P., Vaillant, B., Lallich, S.: On selecting interestingness measures for association rules: user oriented description and multiple criteria decision aid. *Eur. J. Oper. Res.* **184**(2), 610–626 (2008)
21. Elakiya, E., Rajkumar, N.: In text mining: detection of topic and sub-topic using multiple spider hunting model. *J. Ambient Intell. Humanized Comput.* **12**(3), 3571–3580
22. Kanagaraj, R., Rajkumar, N., Srinivasan, K., Anuradha, R.: *Regional Blood Bank Count Analysis Using Unsupervised Learning Techniques*. Lecture Notes on Data Engineering and Communications Technologies, vol. 35. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-32150-5\\_100](https://doi.org/10.1007/978-3-030-32150-5_100)

# An Efficient Approach for Identification of Multi-plant Disease Using Image Processing Technique



**K. Kranthi Kumar, Jyothi Goddu, P. Siva Prasad, A. Senthilrajan, and Likki Venkata Krishna Rao**

**Abstract** Plant diseases must be identified to prevent loss of yields and quantities of agricultural products. Plant disease studies mean plant patterns are observable visually. Health monitoring and detection of plant diseases are very important for sustainable agriculture. Manually monitoring plant diseases is very difficult. It requires huge work, expertise in plant diseases and too long to deal with them. Image processing is therefore used to detect diseases of plants. Disease detection includes steps like image capture, image preparation, fragmentation of images, extraction of functions, segmentation and characteristic extraction. Farmers need automatic disease monitoring to improve crop growth and productivity. Manual Identification of plant diseases. Manual disease monitoring is not effective, because old naked eyes require a more time-limited process for expertise in disease recognition, and is therefore ineffective. This paper addresses plant disease detection using computer-aided method like image processing with MATLAB. Results show that 87% diseases identification has generated by proposed system.

**Keywords** Plant leaf features · Image disease detection · Image processing · Classification · Convolutional network

---

K. Kranthi Kumar (✉)

CSE Department, Dhanekula Institute of Engineering and Technology Ganguru, Vijayawada, India  
e-mail: [kk97976@gmail.com](mailto:kk97976@gmail.com)

J. Goddu

IT Department, Vignan Institute of Information Technology, Visakhapatnam, India

P. Siva Prasad

Department of Computer Science, Alagappa University, Karikudi, Tamil Nadu 630003, India

A. Senthilrajan

Department of Computational Logistics, Alagappa University, Karikudi, Tamil Nadu 630003, India

L. V. K. Rao

CSE Department, Lakireddy Bali Reddy College of Engineering, Mylavaram, AP, India

## 1 Introduction

Plants are now susceptible to a variety of diseases as a result of the extensive use of pesticides and sprays; however, recognizing rotting sections of plants early on can help rescue them [1]. Examining plants for disease necessitates looking for a variety of patterns on them [2]. It can take a long time to manually identify disease in plants, thus image processing could be helpful [3]. Plant disease can affect the stem, root, shoot and even the fruit of the plant [4]. Automatic detection of plant sickness saves time and allows the plant to be protected from disease in its early stages [5]. The ancient and conventional method of identifying and distinguishing plant diseases relies on naked eye inspection, which is less thorough than slow methods [6]. Expert consultation to detect plant diseases is costly and time consuming in some nations since experts are available [7]. Plant diseases proliferate due to haphazard plant management, forcing the use of more pesticides to treat them, as well as chemicals that are hazardous to other farm animals, insects and birds. Automatic detection of plant conditions is required to detect disease symptoms in the early stages of plant and fruit growth [8]. Agriculture, or the art and science of raising livestock and plants, has grown increasingly important in the growth of human civilization. Many farmers are unable to identify plant diseases, leading to the loss of agricultural products [9]. Agro scientists can provide a better solution by using images and videos of crops that provide a clearer view. Plants can be infected with a variety of diseases that have no obvious signs at first, resulting in societal and economic losses [10]. To make things easier, image processing is used, which aids in the resolution of such situations by extracting leaf features that can be used to diagnosis ailments. Image acquisition, pre-processing, segmentation, feature extraction and classification are all processes in the image processing process [11]. A new photo identification system is developed based on multiple linear regressions [12]. A number of advancements have gone into the photo segmentation and recognition technology [13]. The recognition system is built using multiple linear regression and feature extraction [14]. The system has great precision, dependability and photo recognition capabilities, according to the findings [15]. The process is useful, according to the findings, because it aids in the detection of disorders with less effort.

## 2 Related Work

Plant diseases and pests have a substantial impact on productivity and quality. Digital image processing can be used to identify plant diseases and pests. In recent years, deep learning has achieved major advances in the field of digital image processing, far outperforming traditional methods. Deep learning has grabbed the curiosity of scientists interested in learning how to utilize it to identify plant diseases and pests [16].

Agriculture's economic importance cannot be overstated. This is one of the reasons why, because disease in plants is a natural occurrence, disease detection in plants is critical in the agricultural sector [17]. The quality, quantity and productivity of the plants may deteriorate if this area is not well kept. Small leaf disease, for example, is a devastating disease that affects pine trees in the USA [18]. In large agricultural farms, utilizing an automated method to identify plant disease decreases the amount of monitoring required and allows for early diagnosis of symptoms [19].

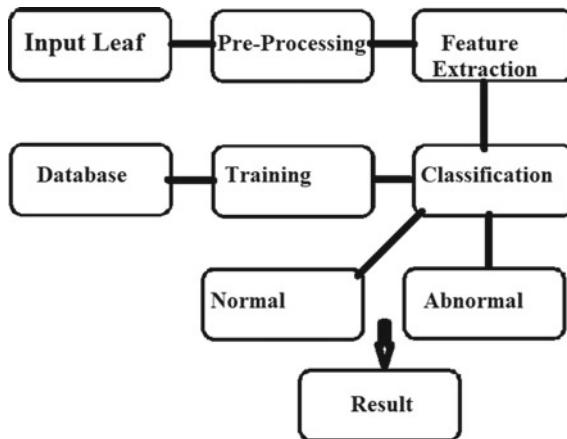
Crop diseases are a major threat to food security, yet early detection is difficult in many parts of the world due to a lack of infrastructure [20]. Thanks to a combination of increased global smartphone usage and recent breakthroughs in computer vision enabled by deep learning, smartphone-assisted disease detection is now a possibility [21].

### 3 Methodology

#### 3.1 *Image Processing*

Researchers in current technology have been working to improve the growth of plants on a regular basis. Growing higher-breed seeds and plants has been a huge success for them. Crop diseases and pesticides, for example, are still major challenges in crop growing. As a result of these issues, crop yields are dropping, and plant cultivation is scarce in the country. The majority of diseases must be prevented early on, but if this is not done, significant harm can result. To reduce such losses, diseases must be identified early. Corn is farmed for a lengthy period of time, between ten and eighteen months, which could lead to disease outbreaks. Fungus, which shows as tiny patches on the leaves, is the most common disease in corn. Serious infections destroy the leaves to the point where they are entirely ruined and covered in patches. Regular pesticide use increases the amount of poisons in the products, posing a number of health risks and contributing to groundwater contamination. Pesticide prices have been steadily rising in recent months. As a result, contemporary technology allows crop yields to be increased while requiring less labour and time. Plant disease diagnosis and categorization are important aspects of plant production and agricultural yield decline. This study uses image processing to develop a system for detecting and classifying plant leaf diseases. Image pre-processing and analysis, feature extraction and plant disease recognition are the three main processes of the method. Because plant diseases are so little, a person's visual capacity limits their diagnosis. Because of the optical nature of the task, computer visualization technologies are applied in plant disease recognition (Fig. 1).

The goal is to accurately recognize the manifestations of an illness that has hurt leaves. After the caught picture has been pre-handled, the different properties of the plant leaf, like power, shading and size, are separated and shipped off a neural organization for characterization. Regardless of the way that numerous frameworks have

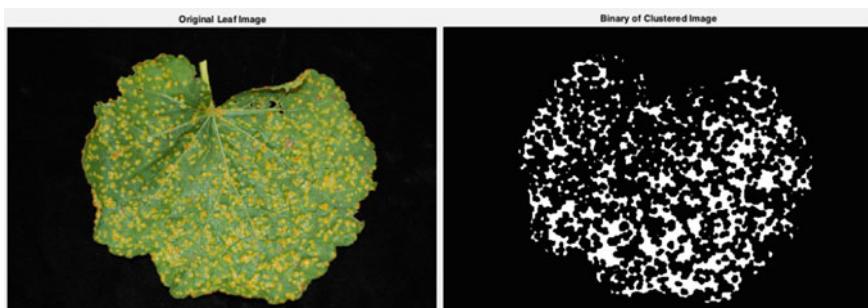


**Fig. 1** Leaf image processing architecture

been created to date utilizing different AI calculations like arbitrary woods, innocent Bayes, and fake neural organizations, their exactness is low, and the work utilizing those arrangement methods is finished and determined to distinguish sickness in just one plant animal groups. These endeavours have just been utilized by a couple of ranchers in Karnataka. Ranchers keep on recognizing ailments with their unaided eyes, which is a significant issue on the grounds that the rancher has no agreement what sickness the plant is tainted with. Ranchers are as yet managing issues, and sickness location strategies are tedious (Figs. 2 and 3).

The proposed method includes the following:

1. Image acquisition
2. Image pre-processing
3. Segmentation
4. Feature extraction
5. Classification.



**Fig. 2** Original and binary clustered image



**Fig. 3** Multi-plant input leaf images

### 3.1.1 Image Acquisition

This stage entails uploading photos of plant leaves into our programme for disease analysis. Because categorization is easier on black and white photographs, which are 2D images, the photos in Fig. 4 are converted to grayscale images at this stage. The system will retrieve the plant snapshot and load it into the system at this point. The following are the steps that follow the acquisition of an image. As a data source (JPG format) for picture analysis, higher standard resolutions will be used, and JPEG is the most commonly used image format.



**Fig. 4** Multi-plant input leaf images processing

### 3.1.2 Image Pre-processing

These strategies have only been employed by a few farmers in Karnataka. Farmers are still using their naked eyes to diagnose ailments, which is a serious worry, because the farmer has no knowledge what disease the plant is suffering from. Farmers are still facing difficulties, and disease detection methods are time consuming.

### 3.1.3 Feature Extraction

The shape highlights utilized in this paper to extricate shape highlights incorporate strength, degree, minor pivot length and unconventionality. The evil part of the leaf in issue is separated utilizing these models. Difference, connection and energy are a portion of the textural highlights utilized in the paper. The ill section of the leaf in issue is extracted using these criteria. The variation of pixels and their neighbours will be determined at the end. Colour feature extraction provides a unique way of exhibiting picture representation when it comes to translation, scaling and rotation. The mean, skewness and kurtosis all have a role in determining colour.

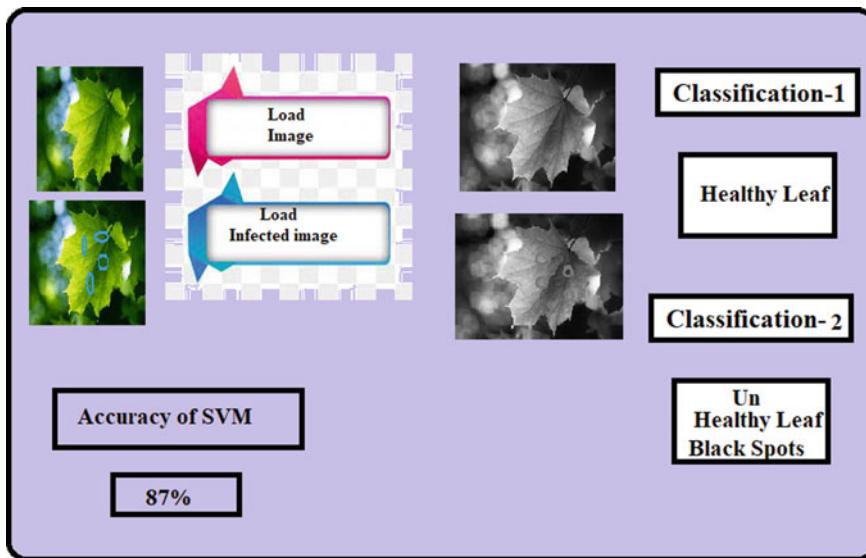
### 3.1.4 Classification

Arrangement incorporates isolating information into two sorts preparing and testing sets. The preparation put out has one objective worth and many provisions for each example or information. The most vital advance is to find the partition hyperplane, which will isolate these focuses into two gatherings: positive (“+1”) and negative (“−”) classes (“−1”). The final outcome is shown in Fig. 5.

## 4 Results and Experimental Evaluation

In plant leaf illness recognition and stage forecast framework, we have been carried out exceptionally prepared model that can precisely perceive infections. In this framework, we utilized Gaussian haze for dim scale change and Otsu’s strategy for paired transformation of pictures after which we utilized raised body for edge recognition. Dark scale change in dim scale transformation shading picture is changed over into a dim structure utilizing Gaussian haze. Shading picture containing commotion and undesirable foundation is eliminated or obscured by utilizing this strategy.

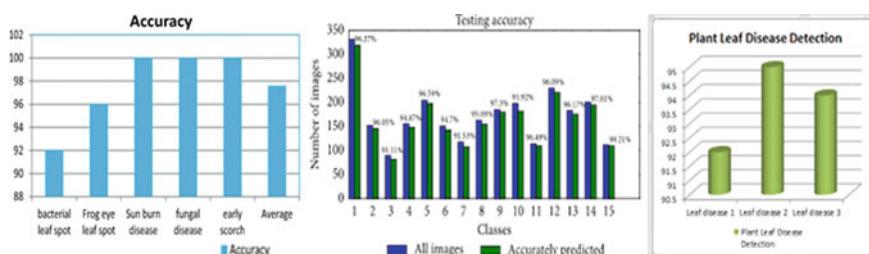
Double change dim scale picture is given to include for Otsu’s technique for parallel transformation. Twofold type of pictures changed over in 0 and 1 structure implies high contrast. In edge discovery, parallel picture gets measurements by counters utilizing curved frame calculation. In which, unconventionality discovers drawing edges around white part of paired picture. In our framework, we are utilizing tensor stream for removing components of preparing data set. In which, 12,000



**Fig. 5** System results generation

picture tests are prepared by utilizing preparing model. At last, plot records were created as a yield of our prepared model. Testing model in definite period of information testing in which leaf illness and typical leaf pictures were coordinated by our preparation model with higher per cent of exactness. In the wake of coordinating with all kind of sickness, pictures separate after-effects of stage and identification that are shown on console and put away in text record too.

The results show that the selected support vector machine classifier outperforms the extreme learning machine. The sensitivity of the support vector machine with a polynomial kernel is likewise higher than that of the other classifiers. Because the developed real-time hardware is capable of detecting a variety of plant illnesses, this work appears to be of considerable social importance (Fig. 6).



**Fig. 6** Accuracy and graphs for results

## 5 Conclusion and Future Scope

The system designed the plant sickness recognition and expectation dependent on profound neural organization and machine acquiring information on techniques. It very well may be exceptionally fundamental for the hit development of yield, and this might be accomplished utilizing picture handling. This paper goes to concoct the various techniques to fragment the disease part of the plant. This paper moreover referenced a couple of element extraction and type techniques to extricate the provisions of excited leaf and the grouping of plant infections. Our future examination will be reached out for additional improvement in crop illness acknowledgement exactness and work for constant sickness acknowledgement.

## References

- Vishnoi, V.K., Kumar, K., Kumar, B.: Plant disease detection using computational intelligence and image processing. *J. Plant Dis. Prot.* **128**(1), 19–53 (2021)
- Singh, V., Sharma, N., Singh, S.: A review of imaging techniques for plant disease detection. *Artif. Intell. Agric.* (2020)
- Singh, D., Jain, N., Jain, P., Kayal, P., Kumawat, S., Batra, N.: PlantDoc: a dataset for visual plant disease detection. In: Proceedings of the 7th ACM IKDD CoDS and 25th COMAD, pp. 249–253 (2020)
- Tete, T.N., Kamlu, S.: Plant disease detection using different algorithms. In: RICE, pp. 103–106 (2017)
- Nagaraju, M., Chawla, P.: Systematic review of deep learning techniques in plant disease detection. *Int. J. Syst. Assur. Eng. Manage.* **11**(3), 547–560 (2020)
- Nandhini, S.A., Hemalatha, R., Radha, S., Indumathi, K.: Web enabled plant disease detection system for agricultural applications using WMSN. *Wireless Pers. Commun.* **102**(2), 725–740 (2018)
- Sawarkar, V., Kawathekar, S.: A review: rose plant disease detection using image processing. *IOSR J. Comput. Eng. (IOSR-JCE)*. e-ISSN 2278-0661
- Bankar, S., Dube, A., Kadam, P., Deokule, S.: Plant disease detection techniques using canny edge detection & color histogram in image processing. *Int. J. Comput. Sci. Inf. Technol* **5**(2), 1165–1168 (2014)
- Sharath, D.M., Kumar, S.A., Rohan, M.G., Prathap, C.: Image based plant disease detection in pomegranate plant for bacterial blight. In: 2019 International Conference on Communication and Signal Processing (ICCSP), pp. 0645–0649. IEEE
- Halder, M., Sarkar, A., Bahar, H.: Plant disease detection by image processing: a literature review. *Image* **1**, 3 (2019)
- Sethy, P.K., Barpanda, N.K., Rath, A.K., Behera, S.K.: Image processing techniques for diagnosing rice plant disease: a survey. *Procedia Comput. Sci.* **167**, 516–530 (2020)
- Cristin, R., Kumar, B.S., Priya, C., Karthick, K.: Deep neural network based Rider-Cuckoo Search Algorithm for plant disease detection. *Artif. Intell. Rev.* **53**(7) (2020)
- Kshirsagar, G., Thakre, A.N.: Plant disease detection in image processing using MATLAB. *Int. J. Recent Innov. Trends Comput. Commun.* **6**(4), 113–116 (2018)
- Venkataraman, A., Honakeri, D.K.P., Agarwal, P.: Plant disease detection and classification using deep neural networks. *Int. J. Comput. Sci. Eng* **11**(9), 40–46 (2019)
- Tlhobogang, B., Wannous, M.: Design of plant disease detection system: a transfer learning approach work in progress. In: 2018 IEEE International conference on applied system invention (ICASI), pp 158–161. IEEE (2018)

16. Liu, J., Wang, X.: Plant diseases and pests detection based on deep learning: a review. *Plant Methods* **17**, 22 (2021). <https://doi.org/10.1186/s13007-021-00722-9>
17. Singh, V., Misra, A.K.: Detection of plant leaf diseases using image segmentation and soft computing techniques. *Inf. Process. Agric.* **4**(1), 41–49 (2017)
18. Manoharan, J.S.: Study of variants of extreme learning machine (ELM) brands and its performance measure on classification algorithm. *J. Soft Comput. Paradigm (JSCP)* **3**(02), 83–95 (2021)
19. Sungheetha, A., Sharma, R.: Design an early detection and classification for diabetic retinopathy by deep feature extraction based convolution neural network. *J. Trends Comput. Sci. Smart Technol. (TCSST)* **3**(02), 81–94 (2021)
20. Sungheetha, A., Sharma, R.: A novel CapsNet based image reconstruction and regression analysis. *J. Innov. Image Process. (JIIP)* **03**, 156–164 (2020)
21. Chen, J.I.-Z., Chang, J.-T.: Applying a 6-axis mechanical arm combine with computer vision to the research of object recognition in plane inspection. *J. Artif. Intell.* **2**(02), 77–99 (2020)

# Polling Cycle Analysis Using Different Modulation Types for IoT-Based Health Control in a Smart City



Imane Benchaib, Salma Rattal, Kamal Ghoumid,  
and El Miloud Ar-Reyouchi

**Abstract** This paper aims to analyze the performance of narrow-band Internet of Things (NB-IoT) technology in the state of release patients tracking using long-term evolution (LTE) cellular networks in the smart city. The study investigates to analyze the time required for the hospital as a control center to poll all patients as remote terminal units (RTUs) one by one in smart city and to receive their responses. The study analyzes the polling cycle concerning the number of hops to reach the patients and the number of patients in the communication range of the base station (BS). The results provide a quick overview of performance based on several fundamental factors in a mobile healthcare application. In particular, the paper focuses on analyzing the impact of modulation types and rates on the depreciation of the polling cycle.

**Keywords** NB-IoT · Polling cycle · Modulation types · Modulation rates · Wireless communication

## 1 Introduction

Nowadays, NB-IoT [1] is considered a major commercial development opportunity in various important activity sectors [2]. NB-IoT is the most advanced technical standard [3] for short messaging services, such as medical sensor data [4], developed by 3GPP Release 13 and beyond NB-IoT. The NB-IoT is also based on the LTE Advanced Pro architecture and can range up to several kilometers from the base station. Health is a new NB-IoT platform [5] in which a patient's health may be diagnosed using a combination of medical sensors and health detection abilities. Using BS-RTU applications, the current approach analyzes wireless bidirectional point-to-multipoint communication in medical NB-IoT networks. The paper [6] describes

---

I. Benchaib · S. Rattal · K. Ghoumid  
ENSAO, University Mohammed First, Oujda, Morocco

E. M. Ar-Reyouchi (✉)  
Faculty of Sciences, Abdelmalek Essaâdi University, Tetouan, Morocco  
e-mail: [e.arreyouchi@m.ieice.org](mailto:e.arreyouchi@m.ieice.org)

how the authors employ an Internet of Things protocol to identify patients' access points and how they may get messages from the hospital advising them to take safety measures.

This paper analyzes the complete polling cycle [7, 8]; it represents the total time required for a control unit in the hospital to perform fast total polling [9] of all necessary connections of patients RTU. The switchover time is an important parameter in the polling cycle [10]. Once a service has been performed, the server takes a brief time to transit a new queue. The factors considered in work for determining the switchover time are the number of patients, the received, transmitted, the transferred packets, the buffer and packet size, and modulation rates.

The proposed system model uses the time division multiple access (TDMA) technique [11] to avoid serious collisions in the wireless network and improve channel utilization. The work in [12] propose wireless collision detection system has been developed for monitoring physical approach.

Medical sensors and base parameters must be specified for each user, allowing for various regulations. The operation of a medical mesh network [13] in polling mode is undoubtedly the future of communications for patients in freedom. For example, it allows for various modulation rates [14] to arise from different limitations on transmitted signal characteristics. When a medical network in a smart city has many patients, each of whom generates traffic on a low-duty cycle, the polling procedure [15] may become inefficient, using a significant amount of bandwidth and occupying the channel with useless polling messages for extended periods.

True complements to traditional wireless medical networks allow flexibility and patient freedom to roam, which has become too rare. However, they are a tool available to patients in the city to communicate, organize and exchange benefits. Polling methods for data transfer on wireless communication networks are evaluated for performance in several research studies [16, 17]. However, the polling process may become extremely inefficient when a network has many terminals, each of which generates traffic with a low-duty cycle. This limitation is because numerous users must be polled while only a small number have data to provide. Consequently, the polling process consumes significant bandwidth and takes up the channel with useless polling message transmission.

The findings in both frequency shift keying (FSK) and quadrature amplitude modulation (QAM) are investigated in a different context [18]. FSK has a lower symbol rate, lower bandwidth, and better sensitivity, while QAM has a wider bandwidth, better spectral efficiency, power efficiency, and superior system gain.

Data collection and transmission methodology for IoT RTUs are described in the article [19]. This paper considers the RTU's like electronic devices controlled by a microprocessor that serves as a physical interface between patients and the BS or the neighbor patients. It benefits from a distributed control system, or SCADA, in that it transmits telemetry data to the hospital control center. It changes the state of linked devices depending on the control messages received.

The paper investigates the performance evaluation of the entire cycle in a medical network architecture with and without a BS, in which patients are linked wirelessly. This mesh technology allows for the creation of scalable self-formed networks. As

a result, even if a patient on the medical network is no longer accessible, network information may still be accessed by using an alternative path. The primary benefits of this technology are its cheap implementation costs and its ability to withstand faults.

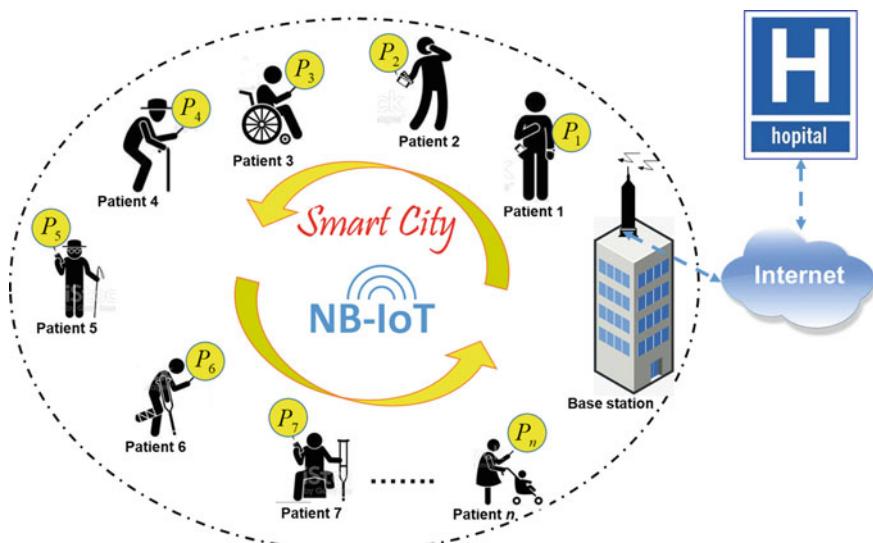
The present research aims to decrease polling cycles significantly, improve healthcare efficiency, and keep patients out of potentially hazardous circumstances, improving productivity, safety, and profitability of the medical network.

The structure of the paper is as follows: Sect. 2 gives a detailed description of the system model to produce a health monitoring report. Section 3 describes the description of the polling system functionality. Section 4 evaluates the analysis of the results. Finally, the article concludes with Sect. 5.

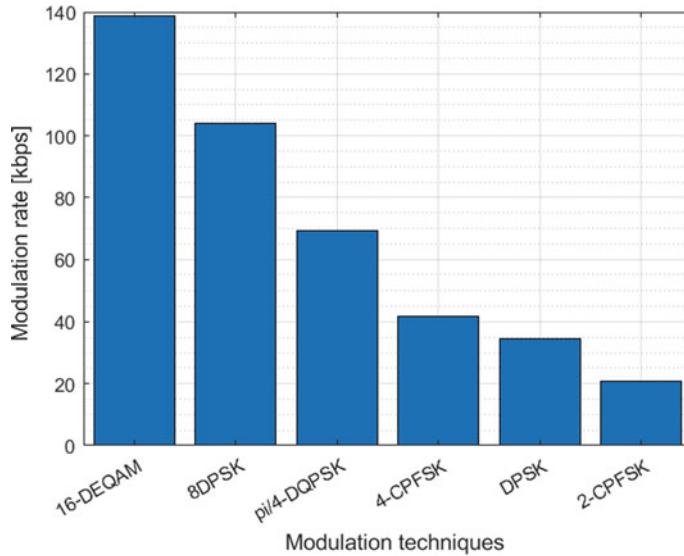
## 2 System Model

Consider the system model given in Fig. 1 of several patients  $P_1, P_2, \dots, P_n$  equipped by RTU, which connects their medical sensors and BS. Figure 1 also presents the meta-model of this medical context implementing NB-IoT into smart cities.

The central control unit hospital polls patients' nodes via the internet and BS, sending polling packets to RTUs. The hospital can be equipped with a supervisory control and data acquisition (SCADA) [20] system, generating the request and receiving medical data.



**Fig. 1** System model for the analysis of complete polling cycle for several patients using medical mesh network in a smart city



**Fig. 2** Modulation rate versus modulation techniques

In Fig. 2, the  $n$  patients share the same channel, and only one user may be serviced at any one moment. The BS polls patient nodes by sending polling packets to its RTUs.

Any patient's RTU in the topology network presented in Fig. 1 (mesh network) can randomly and spontaneously access any other RTU. Polling cycle and report-by-exception applications also can be run over this medical network. It can be done several times. Each patient's packets are broadcast to all remote patients, including the base station. This dissemination allows connecting more RTUs to every remote patient.

In poll-response networks, a medical hospital center communicates with several remote patients one at a time and the BS exchanges medical data with the currently connected remote patient. This process ends when the distant patient in the polling ordering shown in Fig. 1 establishes a new connection with the system. In case of indirect line-of-sight (LOS) visibility, each packet is delayed for a period equal to the time required to transmit the number of bytes specified before being sent over the radio channel.

The selected modulation rate determines the polling duration. The server wants to delay all packets when it takes 150 user data bytes to transmit a UDP packet. Therefore, it has to set the following 178 bytes adding to 150 bytes 8 bytes for the UDP leader and 20 bytes for the IP header. For a successful communication and correct exchange of medical data, IoT patients  $P_1, P_2, \dots$ , and  $P_n$  must be within the BS node's communication range and capable of transmitting packets.

Repeater and bridge modes are appropriate for point-to-multipoint networks with BS-user applications that communicate using a polling-type protocol. Using a complex protocol in the radio channel, medical devices in repeater and bridge modes

are as simple to operate as a basic transparent device while offering communication dependability and spectrum efficiency. It is emphasized that monitoring medical data from a free patient of NB-IoT models in cities is a new approach.

### 3 Description of the Functionality

Consider the control system depicted in Fig. 1; every patient  $P_i \in \{1, 2, \dots, n\}$  has a waiting queue.

#### 3.1 Description

The polling cycle is an essential metric in the present application. The transmitted data in bytes for each polling cycle is

$$(p_l - p_h) \times (n - 1) \times 2 \text{ (bytes)} \quad (1)$$

where  $p_l$  is the packet length,  $p_h$  is the packet header length, and  $n$  is the number of patients in the suggested medical network.

Consider a cellular medical network in a smart city consisting of  $n$  patients numbered as  $P_1, P_2, \dots, P_n$  which are served cyclic order. In this paper, the polling cycle is defined as the time between the instant where  $P_1$  is polled and the instant where the patient  $P_n$  finishes its service period (including switchover time), and the server returns to  $P_1$ .

Consider that the server visits patients by respecting cyclic order. This order is supposed to be  $P_1, P_2, \dots, P_n, P_1, P_2$ , and so on. Then, the operation is continued cyclically without any loss of generality.

Note that, in this paper, the switchover time is the time to switch from  $P_i$  to  $P_{i+1}$  by  $D_i, i = 1, \dots, n$  and the total switchover time in a cycle by  $D_T = \sum_{i=1}^n D_i$ . Further, it is to be emphasized that  $S_T = \sum_{i=1}^n S_i$  ( $S_T$  is the total number of packets served) and  $B_T = \sum_{i=1}^n B_i$  ( $B_T$  is the total service time) for the given polling cycle  $n$ . The service time is the duration of the polling cycle spent transmitting packets.

Assuming that service and switchover durations are statistically independent and equally distributed across various polling events, so  $E[D_T] = E[D]$ ,  $E[S_T] = E[S]$ , and  $E[B_T] = E[B]$ . Moreover, it is also assumed that the statistics controlling service and switchover times are the same for all patients, implying that the system is symmetric.

$$E[D] = \sum_{i=1}^n E[D_i] = P_i \cdot D \quad (2)$$

Overall, assuming that  $P_C$  is the entire overhead time needed by system software for one polling cycle  $T_i$  is the amount of period it takes the patient to transmit every one of its packets when polled and after that the polling cycle period is computed.

$$T_C = P_C + \sum_{i=1}^n T_i \quad (3)$$

### 3.2 Functionality

A typical network with a single BS and several remotes (remote patient monitoring) has a tree-like structure that the addressing scheme can follow. The queue status of distant patients for which the BS does not have queuing information is checked regularly (Table 1).

The medical devices in the system model shown in Fig. 2 are based on direct line-of-sight (LOS) visibility. The medical devices can work in repeater mode with a fully transparent radio protocol. The repeaters are suitable for all polling (request-response) applications, particularly control and monitoring of health care. For example, if  $P_i, i \in \{1, 2, \dots, n\}$  do not get his packet because it is not within the coverage range of BS. The packet will be received by  $P_{i\pm 1}$ , and it is addressed to  $P_i$ .

Table 2 shows an overview of the different simulation settings as well as the values of important parameters.

OFDMA with a carrier spacing of 15 kHz is used for both the downlink and uplink transmissions. Only 12 carriers are used in NB-IoT, resulting in a 180 kHz occupied bandwidth. In order to improve performance in this work, choosing various modulation schemes is requested; each modulation type can distinguish at different

**Table 1** Functionality example based on system model of Fig. 1

	Description
Step 1	The cycle of polling begins
Step 2	BS broadcast requests packets (received by the hospital control center via the Internet) to all patients in its average range. The patients $P_1, P_2, \dots, P_n$ can receive it, but firstly only $P_1$ responds. The other patients within BS coverage listen and store the correctly received packet and wait for their turn to answer
Step 3	$P_1$ sends its response packet to the BS as well as to its neighbors. BS receives this packet, and it will automatically be transferred to the hospital control center for medical intervention
Step 4	Then, $P_2$ sends the response packet to BS as well as to its neighbors as in Step 2. It is suggested that the $P_1, P_2, \dots, P_n$ and BS interact directly with each other in turn
Step 5	The same procedure continues for $P_3, P_4, \dots$ until $P_n$ . Then, the server visits the patients again $P_1, P_2, \dots, P_n$ in a cyclic order. The medical polling cycle operates continuously

**Table 2** Designation, notation, and the values of critical medical network parameters

Notation, definition, and values of key network parameters	Value
Total number of patients (TNPs)	10, 20, 30, 40, 50, and 100
Average message size	100 bytes
Channel spacing [kHz]	50
Acknowledgment (ACK)	Off
Forward error correction (FEC)	Off
Modulation type	QAM (16-DEQAM, 8DPSK, pi/4-DQPSK, and DPSK), FSK (2-CPFSK and 4-CPFSK)
Occupied bandwidth	180 kHz
Carriers	12
Number of hospitals	1, 2, 3, 4, and 5

durations for the polling cycle. In order to analyze the evolution of the polling cycle time in the proposed medical network, the simulation tool employed is MATLAB.

## 4 Evaluation Results with Interpretation

This section provides numerical findings illustrating polling cycle analysis using different modulation techniques for IoT-based health control. The results consider the number of patients, average message size, channel spacing [kHz], FEC, ACK, modulation type, and modulation rates.

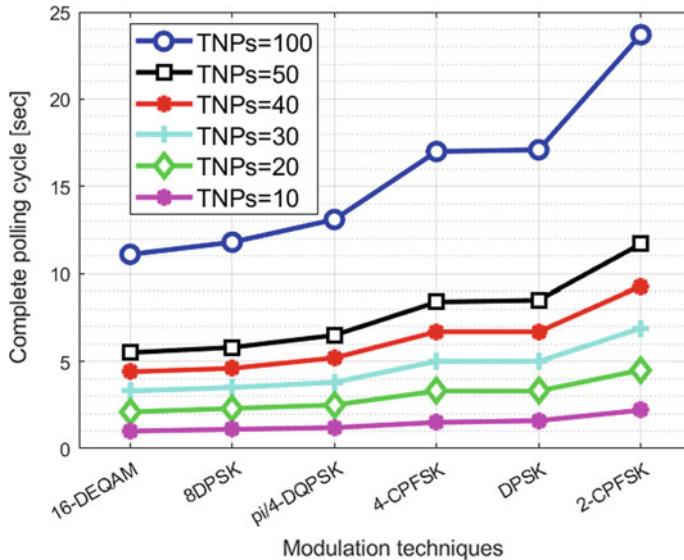
The findings confirm the analysis of the whole polling cycle's performance, highlight the significance of modulation rates, and offer valuable technical insights.

Figure 2 shows the modulation rates (kbps) as a function of the used modulation types such as QAM (16-DEQAM, 8DPSK, pi/4-DQPSK, and DPSK) and FSK (2-CPFSK and 4-CPFSK).

The modulation rates used in a suggested application vary from approximately 30 to 168 kbps. The results revealed a clear specific correspondence between modulation technique and modulation rate.

Consequently, different constraints on transmitted signal characteristics produce varying modulation rates. The findings obtained from Fig. 2, in the present study, allow to achieve better results; the choice of the various modulation schemes is requested; each modulation type can distinguish at different durations for the polling cycle. Therefore, the modulation-type selection determines the value of the modulation rate.

Higher modulation rates in the present medical application result in faster data transfer rates, resulting in lower receiver sensitivity and a smaller coverage range, whereas with lower modulation rates, transmission reliability across a radio channel is always better.



**Fig. 3** Complete polling cycle versus modulation techniques

Figure 3 provides complete polling cycle (s) as a function of the used modulation techniques for TNP<sub>s</sub> = 10, 20, 30, 40, 50, and 1000 patients.

Figure 3 compares the rise in TNPs versus modulation method types to the growth in the whole polling cycle. Furthermore, the findings show that the 16-DEQAM positively reduces the complete polling cycle in environmental monitoring and healthcare applications.

Furthermore, in the NB-IoT system in a wireless medical network, the polling cycle lowers as TNPs drop but increases when modulation rates rise.

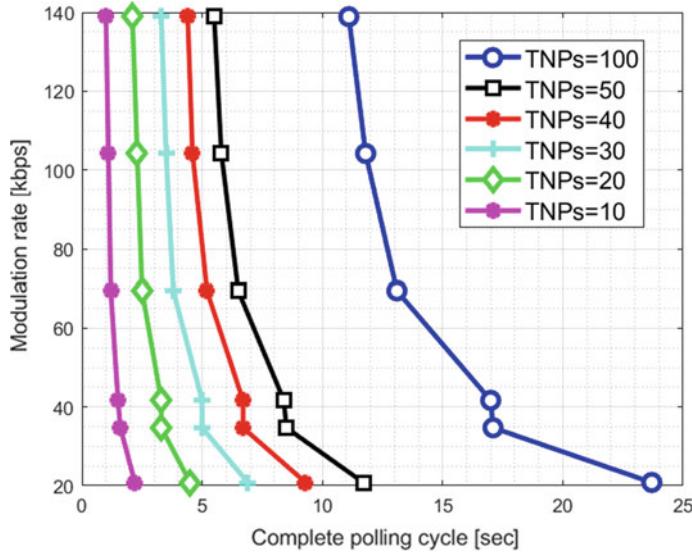
Figure 4 shows the modulation rates (kbps) as the complete polling cycle function. It is observed that the modulation techniques strongly impact the proposed IoT scheme regarding the complete polling cycle.

The comparison results of the present study, illustrated in Fig. 4, show that the complete polling cycle decreases with the increased proportion of the modulation rates.

Table 3 provides a summary of the main results derived from Fig. 4

According to the values shown in Table 3, it can be observed that the complete polling cycle rises significantly for the TNPs values increase. Furthermore, users can easily observe that modulation rates may significantly impact the polling cycle. Table 3 also shows that doubling the modulation rates favors a decrease in the complete polling cycle.

Let us, therefore, denote 50 patients, and with a change of modulation rates from 34.72 to 69.44 kbps, the evolution rate from 8.5 to 6.5 s promotes a polling cycle improvement 23.53%, while for the evolution rate from 6.5 to 5.5 s, a polling cycle improvement becomes 15.38%.



**Fig. 4** Modulation rate (kbps) versus complete polling cycle (s)

**Table 3** Comparison of the main results derived from Fig. 4

Modulation rates (kbps)	138.89			69.44			34.72		
TNPs	20	50	100	20	50	100	20	50	100
Complete polling cycle (s)	2.1	5.5	11.1	2.5	6.5	13.1	3.3	8.5	17.1

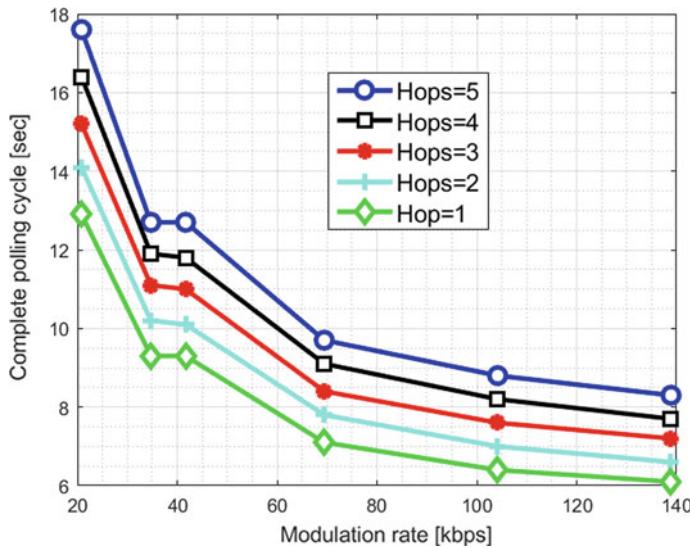
Figure 5 shows the results of complete polling cycles versus modulation rates for hops ranging from 1 to 5 with modulation type QAM and FSK.

The results illustrated in Fig. 5 show that the complete polling cycle is increased when the modulation rate increases. We notice that the polling cycle time decreases significantly compared to FSK modulation for the QAM modulation.

Table 4 gives a summary of a comparative overview of the results providing the main results derived from Fig. 5.

According to the values listed in Table 4, it can be observed that the complete polling cycle decreases considerably for the modulation rate increase. The number of hops in wireless medical networks has a significant effect on the polling cycle.

Therefore, note that the coverage area in wireless communication over a radio channel is usually greater with significant modulation rates. Still, the reliability can be obtained with lower modulation rates. Higher modulation rates provide quicker data rates at the expense of lower receiver sensitivity, resulting in a narrower coverage range. When the modulation rate is lower, communication across a radio channel is always more trustworthy.



**Fig. 5** Complete polling cycle versus modulation rates

**Table 4** Comparison of the main results derived from Fig. 5

Modulation rates (kbps)	138.89			69.44			34.72		
Hops	3	4	5	3	4	5	3	4	5
Complete polling cycle (s)	7.2	7.7	8.7	8.4	9.1	9.7	11.1	11.9	12.5

## 5 Conclusion

This research provides a performance analysis based on modulation choices with various rates. The paper has focused on optimizing the polling cycles time. The simulation results were compared with various hops and patient numbers walking in the smart city. This study can also be considered a reference for a simplistic overview of medical network performance in a multi-hop wireless connection with varied modulations and users. The cyclical sequence of cellular network communication is also helpful in an emergency. Fast algorithms or protocols to improve polling cycle time significantly are strongly required, as well as the need for a priority approach is also important for future research.

## References

1. Chen, M., Miao, Y., Hao, Y., K. Hwang, K.: Narrow band internet of things. *IEEE Access* **5**, 20557–20577 (2017)
2. Nair, V., Litjens, R., Zhang, H.: Optimisation of NB-IoT deployment for smart energy distribution networks. *J. Wirel. Commun. Network* **2019**, 186 (2019)
3. Cengiz, K., Ozyurt, B., Singh, K.K., Sharma, R., Topac, T., Chatterjee, J.M.: The role of IoT and narrow band (NB)-IoT for several use cases (2021)
4. Vitabile, S., et al.: Medical data processing and analysis for remote health and activities monitoring. In: Kołodziej, J., González-Vélez, H. (eds.) *High-Performance Modelling and Simulation for Big Data Applications*. Lecture Notes in Computer Science, vol. 11400. Springer, Cham (2019)
5. Dangana, M., Ansari, S., Abbasi, Q.H., Hussain, S., Ali Imran, M.: Suitability of NB-IoT for indoor industrial environment: a survey and insights. *Sensors* **21**, 5284 (2021)
6. Patil, P.J., Zalke, R.V., Tumasare, K.R., Shiwankar, B.A., Singh, S.R., Sakhare, S.: IoT protocol for accident spotting with medical facility. *J. Artif. Intell. Capsule Networks* **3**(2), 140–150 (2021)
7. Ar-Reyouchi, E.M., Ghoumid, K., Ar-Reyouchi, D., Rattal, S., Yahiaoui, R., Elmazria, O.: An accelerated end-to-end probing protocol for narrowband IoT medical devices. *IEEE Access* **9**, 34131–34141 (2021)
8. Siddiqui, S., Ghani, S., Khan, A.A.: Effect of polling interval distributions on the performance of MAC protocols in wireless sensor networks. In: Barolli, L., Enokido, T. (eds.) *Innovative Mobile and Internet Services in Ubiquitous Computing. IMIS 2017. Advances in Intelligent Systems and Computing*, vol. 612. Springer, Cham (2018)
9. Ar-Reyouchi, E.M., Maslouhi, I., Ghoumid, K.: A new fast polling algorithm in wireless mesh network for narrowband Internet of Things. *Telecommun. Syst.* **74**, 405–410 (2020)
10. Khurram, A.: Effect of switchover time in cyclically switched systems, switched systems. Janusz Kleban, Intech Open (2009)
11. Chatei, Y., Maslouhi, I., Ghoumid, K., Ar-reyouchi, E.M.: Performance enhancement of wireless sensor networks using an efficient coding approach. In: 2018 6th International Conference on Multimedia Computing and Systems (ICMCS), pp. 1–5 (2018). <https://doi.org/10.1109/ICMCS.2018.8525943>
12. Dhaya, R., Kanthavel, R.: A wireless collision detection on transmission poles through IoT technology. *J. Trends Comput. Sci. Smart Technol. (TCSST)* **2**(03), 165–172 (2020)
13. Rattal, S., Ar-Reyouchi, E.M.: An effective practical method for narrowband wireless mesh networks performance. *SN Appl. Sci.* **1**, 1532 (2019)
14. Yousra, L., Benchaib, I., Rattal, S., Ar-Reyouchi, E.M., Ghoumid, K.: Performance analysis on modulation techniques for medical devices sensitivity in wireless NB-IoT network. In: Proceedings of the International Conference on Smart Data Intelligence (ICSDI) (2021)
15. Siddiqui, S., Ghani, S., Khan, A.A.: A study on channel polling mechanisms for the MAC protocols in wireless sensor networks. *Int. J. Distrib. Sensor Networks* (2015)
16. Nannicini, S., Pecorella, T.: Performance evaluation of polling protocols for data transmission on wireless communication networks. In: ICUPC '98. IEEE 1998 International Conference on Universal Personal Communications. Conference Proceedings (Cat. No.98TH8384), vol. 2, pp. 1241–1245 (1998)
17. Borges, L., Mvelez, F.J., Lebres, A.S.: Performance evaluation of the schedule channel polling MAC protocol applied to health monitoring in the context. IEEE 802.15.4. In: 17th European Wireless 2011— Sustainable Wireless Technologies, pp. 1–8 (2011)
18. Ar-Reyouchi, E.M., Lamrani, Y., Benchaib, I., Rattal, S., Ghoumid, K.: NCBP: Network coding based protocol for recovering lost packets in the internet of things. In: Belkasmi, M., Ben-Othman, J., Li, C., Essaaidi, M. (eds.) *Advanced Communication Systems and Information Security. ACOSIS 2019. Communications in Computer and Information Science*, vol. 1264. Springer, Cham (2020)

19. Lekbich, A., Belfqih, A., Zedak, C., Boukherouaa, J., El Mariami, F.: A secure wireless control of remote terminal unit using the internet of things in smart grids. In: 2018 6th International Conference on Wireless Networks and Mobile Communications (WINCOM), pp. 1–6
20. Ar-Reyouchi, E.M.: Analysis of the New method for assessing the total network capacity of wireless mesh networks for IoT devices. *J. Internet Technol.* **21**(4), 1025–1035 (2020)

# Wireless Sensor Network Lifetime Improving Based on Routing Protocols



Bilal Saoud, Mounia Boucif, and Mourad Daas

**Abstract** Wireless networks can be used for many purposes. Among them we find wireless sensor network (WSN). A WSN can be seen as a set of sensors deployed in a capture zone in order to track physical values such as temperature, humidity, and pressure. In general, sensors are powered by irreplaceable batteries with limited capacity. The entire WSN is relied to the energy level in sensors. This feature makes the energy a critical resource to be preserved in order to extend the lifetime of the WSN. Many routing protocols have been proposed to ensure data transmission in WSN and extend the WSN lifetime. The aim of this paper is to evaluate, analyze, and compare some very well-known routing protocols. We have compared LEACH, EAMMH, SEP, and E-SEP. Simulation results show that a routing protocol can improve the lifetime of WSN.

**Keywords** Wireless sensor network · Routing protocol · Network lifetime · Simulation

## 1 Introduction

Networks are used in different areas for different purposes. We can find wired and wireless networks. Among wired networks we find Ethernet, which is very used to transmit data for local area networks. And we have IEEE 802.11 to share data based on electronic waves, which is an example of wireless networks. Wireless sensor network (WSN) is also an example of wireless networks without an infrastructures, which means that there is not an existing devices in the topology (like routers) to ensure the communication and data transmission between sensor nodes. Today, WSNs are a hot topic of research and many scientists work about them for almost every domain [1].

---

B. Saoud (✉) · M. Boucif · M. Daas

Electrical Engineering Department, Sciences and Applied Sciences Faculty, Bouira University,  
10000 Bouira, Algeria

e-mail: [bilal340@gmail.com](mailto:bilal340@gmail.com)

A wireless sensor network (WSN) can be defined as a collection of sensor nodes [2, 3]. The mission of these nodes is data collection from a predefined area. The collected data could be temperature, wind speed, humidity, pressure, etc. Usually, sensor nodes have attached batteries, which are the energy source for them. Typically, the sensor nodes are small and disposable and have a limited amount of energy (power). In WSN, sensor nodes work together in order to observe the target region and to send data to the base station (BS) or sink [3]. Many systems have been designed and developed based on WSNs. WSNs have been used to design healthcare system [4], target tracking [5], smart home [6], agriculture [7], smart cities [8], etc.

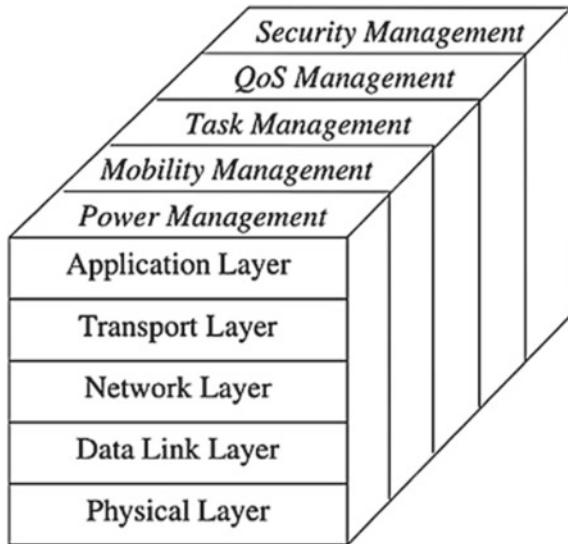
WSN cannot be existed if all its sensor nodes are out of energy. The sensor nodes' energy is the most important fracture in WSN lifetime. This characteristic of lifetime can be defined based on many criteria. According to [9], network lifetime can be defined as a time period that begins from the deployment of sensor nodes until the instant when the network is considered nonfunctional. At [10], WSN lifetime has been defined as an interval of time that starts from the first data transmission in WSN during the setup phase until a percentage of sensor nodes fall below a specific threshold. This threshold is set according to the type of the application. It is not easy to answer the question: When a network should be considered nonfunctional? For instance, the WSN can be considered nonfunctional when the first sensor has died, a percentage of sensors have died, a partition of network was nonfunctional, or coverage loss has occurred.

WSN has a layered architecture or protocol stack. The protocol stack and the associated planes used by the sink, cluster head, and sensor nodes are shown in Fig. 1, where its architecture has five layers which are physical, data link, network, transport, and application layer [11]. WSN protocol stack has received an important attention in order to improve the lifetime of network. Basically, solutions to improve WSN lifetime can be classified according to these layers, some of them improve the physical layer, others try to enhance the mechanism of sharing the medium, and routing protocols have been proposed also to increase the network lifetime. Our study focuses on routing protocols in order to improve the lifetime of WSN. Some of them are based on finding the optimal route between sensor nodes and BS by taking into consideration the residual amount of energy in sensor nodes [12].

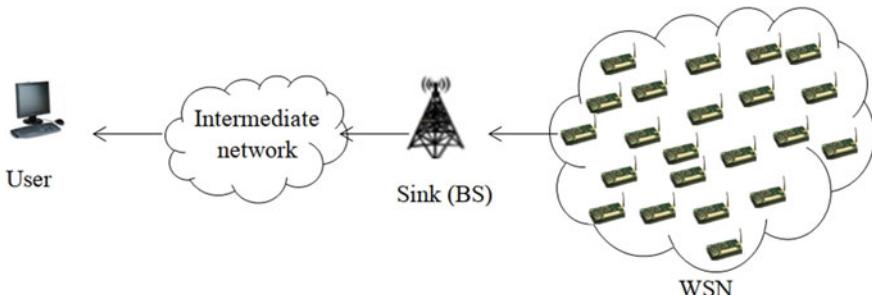
Our paper will be organized as follows; some routing protocols will be presented in the second section. Parameters of our simulation and results will be explained in the third section. We will end our paper by a conclusion.

## 2 Routing Protocols

Sensor nodes are deployed in capture area in order to capture a physical measurement. So, sensor nodes gather data from the capture area and send it to the sink. Each sensor node in the WSN will contribute to reach the sink (BS). After that, the user (or scientist) can get information from the sink through intermediate network like Internet. Figure 2 presents the WSN architecture.



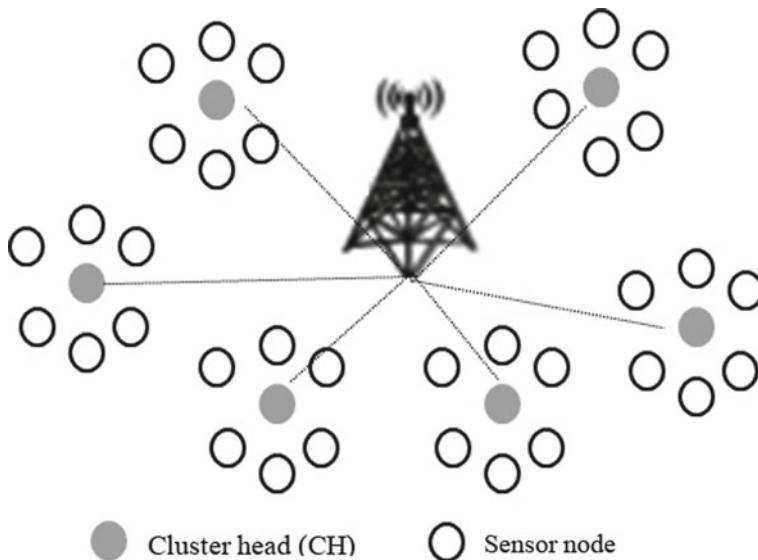
**Fig. 1** Protocol stack of WSNs



**Fig. 2** Architecture of WSNs

Network layer (Fig. 1) ensures the data transmission between sensor nodes and the sink. The main goal of the network layer is to discover ways for energy-efficient clustering and routing of data from sensor nodes to the sink node so that the lifetime of the network can be maximized. Many routing protocols have been proposed over years. In addition, these protocols can be seen as procedures of selecting the right route to reach the destination. They can be classified based on several criteria. We can find load balancing, metaheuristic, and distributed approaches. In the next paragraphs, some WSN routing protocols will be given.

The most famous routing protocol in WSN based on clustering is LEACH [13]. LEACH is a MAC protocol based on the TDMA method. LEACH is based on clustering where sensor nodes are collected into groups (clusters). At each cluster or group, we find two types of nodes, which are cluster head and cluster members.



**Fig. 3** Architecture of WSN with LEACH protocol

When sensor nodes of each cluster want to send data, which is the collected data from the environment, to the sink they must pass by their cluster head (CH). After compression and aggregation of the received data by each CH, it will be sent to the sink. The tasks in LEACH protocol are specified by turns. In addition, two main steps are required for each turn. The first step is the setup, and the second is the data transfer. The selection of CHs and the member sensor nodes for each cluster will be done at the first step. Then, each member sensor node transfers its collected data to its CH based on time slot in the TDMA schedule in the second step. The architecture of WSN with LEACH protocol is illustrated in Fig. 3. CHs send data to sink based on CDMA in order to avoid collision.

CHs are selected at the setup stage. For each round, every sensor node could be head of cluster. The decision will be made based on random numbers, where each sensor node chooses a uniform random number (between 0 and 1). Each sensor node compares its random number with a threshold (see Eq. 1). If it is lower than the threshold, this sensor node will be selected a CH. Otherwise, the node is still a normal node. Then, every sensor node can join a cluster by joining a CH based on the strength of signal received from CHs.

$$T_{(n)} = \begin{cases} \frac{P_L}{1 - P_L \times (r \bmod (\frac{1}{P_L}))}, & n \in G \\ 0 & \text{Otherwise} \end{cases} \quad (1)$$

where  $P_L$  introduces the percentages of CHs in each round,  $r$  is the present round, and  $G$  is a set of sensor nodes that have not yet been CH in the period  $\frac{1}{P_L}$  rounds.

Another interesting protocol is LEACH Centralized (LEACH-C) [14]. It can be seen as an improvement of LEACH protocol. In LEACH-C, information location and amount of energy at each sensor node are sent to the sink at the starting of each round. After the reception of this information, the sink (BS) will make a decision in order to select CHs based on simulated annealing to minimize an objective function. The data transmission is as LEACH protocol.

SEP protocol [15] has two levels of sensor node energy. It is heterogeneous protocol. It improves the clustering process by introducing two parameters: advanced sensor nodes ( $m$ ) and additional energy factor ( $\alpha$ ) to distinguish advanced and normal sensor nodes. The energy level at advanced sensor nodes is more important than normal sensor nodes. In SEP, both sensor nodes (normal and advanced) can be CHs. However, advanced sensor nodes have more opportunities to become CH than normal sensor nodes. E-SEP protocol [16] is proposed for three-level heterogeneous sensor nodes, which are advanced and normal sensor nodes like SEP, and the third type of sensor nodes is intermediate sensor nodes whose level of energy is between normal and advanced sensor nodes. The selection of CHs is done based on the level of energy at each sensor node.

EAMMH routing protocol [17] was proposed based on sensor nodes' energy level and multi-hop intra-cluster routing. EAMMH protocol has many rounds like LEACH protocol, and each round has two stages: the setup and the data transmission stage. Many routes are established from every sensor node and its CH. This strategy is more energy-aware in order to transmit data. The optimal route is chosen. EAMMH chooses CHs based on their residual energy.

### 3 Simulation and Results

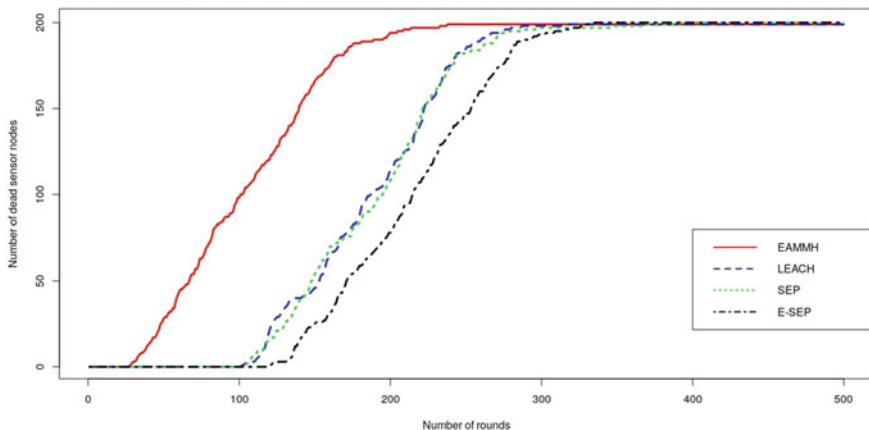
This section presents the simulation of some WSN routing protocols. We have compared LEACH, EAMMH, SEP, and E-SEP protocol. Parameters of simulation are presented in Table 1. In order to compare the selected protocols, two metrics have been chosen. The first one is the network lifetime, and the second is the residual energy at sensor nodes. Network lifetime has been estimating by the number of dead sensor nodes from the starting to the death of all the sensor nodes. Number of CH and number of packets sent to BS and CH at each round have been also generated for all protocols.

Radio hardware energy dissipation model has been used. This model allows us to model the energy dissipation for the transmitter in order to run its radio electronics and power amplifier. In addition, it models also the energy dissipation for the receiver to run its radio electronics. Based on the distance between sensor nodes, the channel models were used [18]. Models of transmission and reception are as follows:

$$E_{Tx}(l, d) = E_{Tx-elec}(l) + E_{Tx-amp}(l, d) = \begin{cases} lE_{elec} + l_{\epsilon_{fs}}d^2, & d < d_0 \\ lE_{elec} + l_{\epsilon_{mp}}d^4, & d > d_0 \end{cases} \quad (2)$$

**Table 1** Simulation parameters

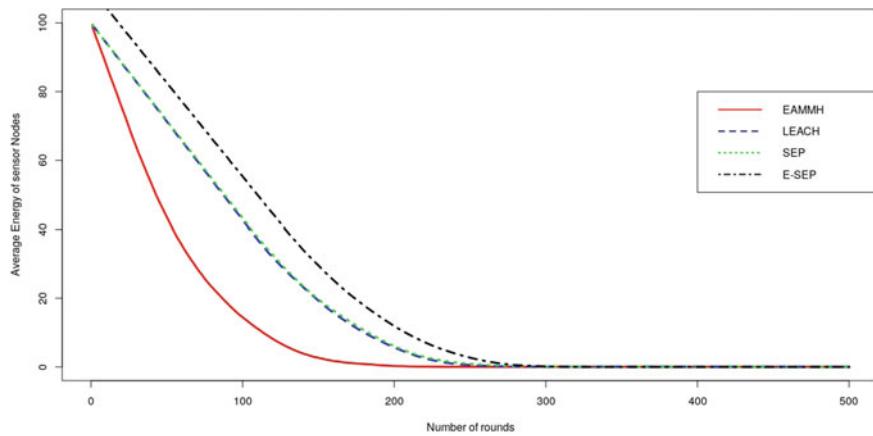
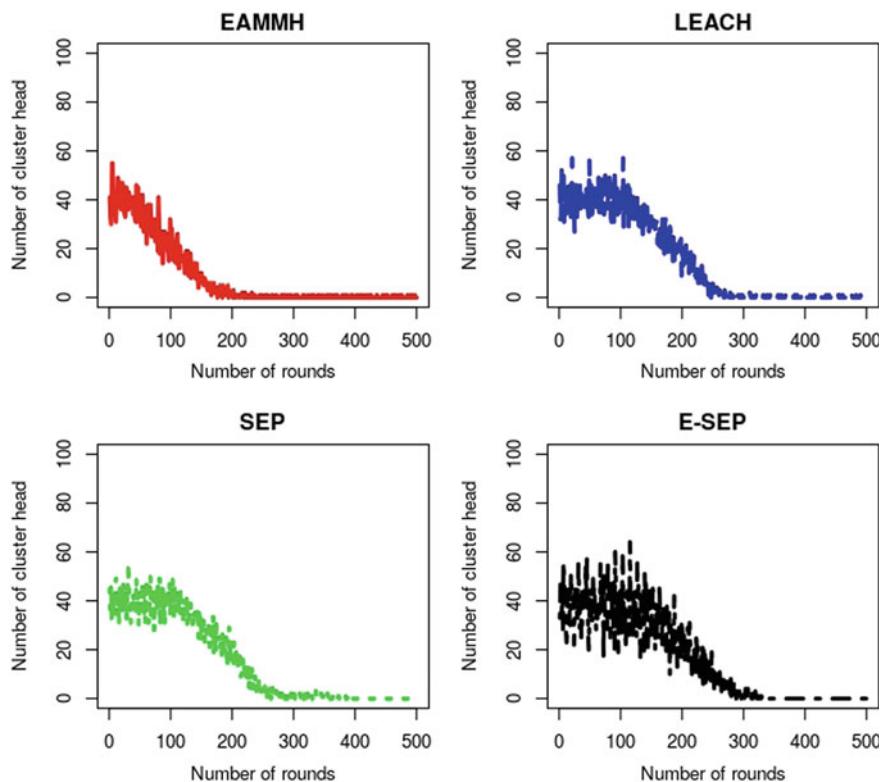
Parameter	Value
Area	100 m <sup>2</sup>
Sensor node number	200
Size of packet $l$	4000 bits
Data aggregation energy for 1 bit $E_{DA}$	5 nJ/bit
Dissipation of energy for transmitting or receiving 1 bit of data $E_{elec}$	50 nJ/bit
Energy dissipation coefficient in the free space (transmission coefficient amplifier $\epsilon_{fs}$ )	10 pJ/bit/m <sup>2</sup>
Energy dissipation coefficient in the multi-path attenuation model (transmission coefficient amplifier $\epsilon_{mp}$ )	0.0013 pJ/bit/m <sup>4</sup>
Initial energy of sensor nodes $E_o$	0.1 J

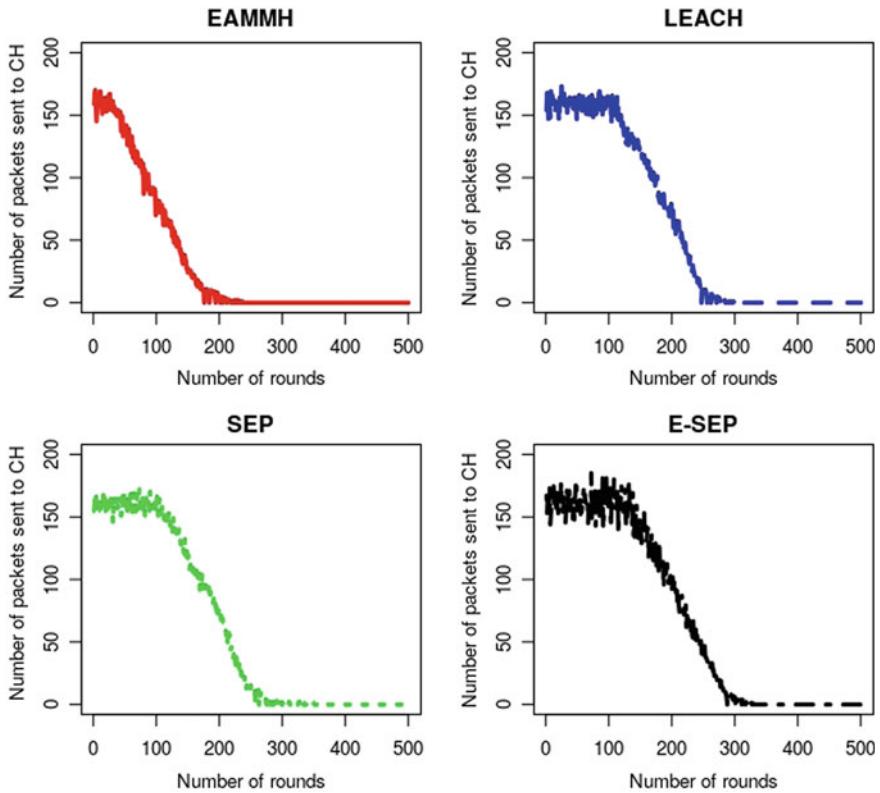
**Fig. 4** Network lifetime: dead sensor nodes

$$E_{Rx}(l) = E_{Rx-elec}(l) = lE_{elec} \quad (3)$$

Figure 4 shows the network lifetime of WSN for each protocol. From this figure, it is obvious to notice the routing protocol that performs better than the rest of protocols. E-SEP performs better than LEACH, EAMMH, SEP. Besides, sensor nodes start to die at 101, 27, 103, and 119 round for LEACH, EAMMH, SEP, and E-SEP, respectively.

The total residual energy of the network is estimated based on the residual energy level in sensor nodes at each round. Figure 5 presents the results of each protocol according to this metric. The total of residual energy has been represented in 100% at

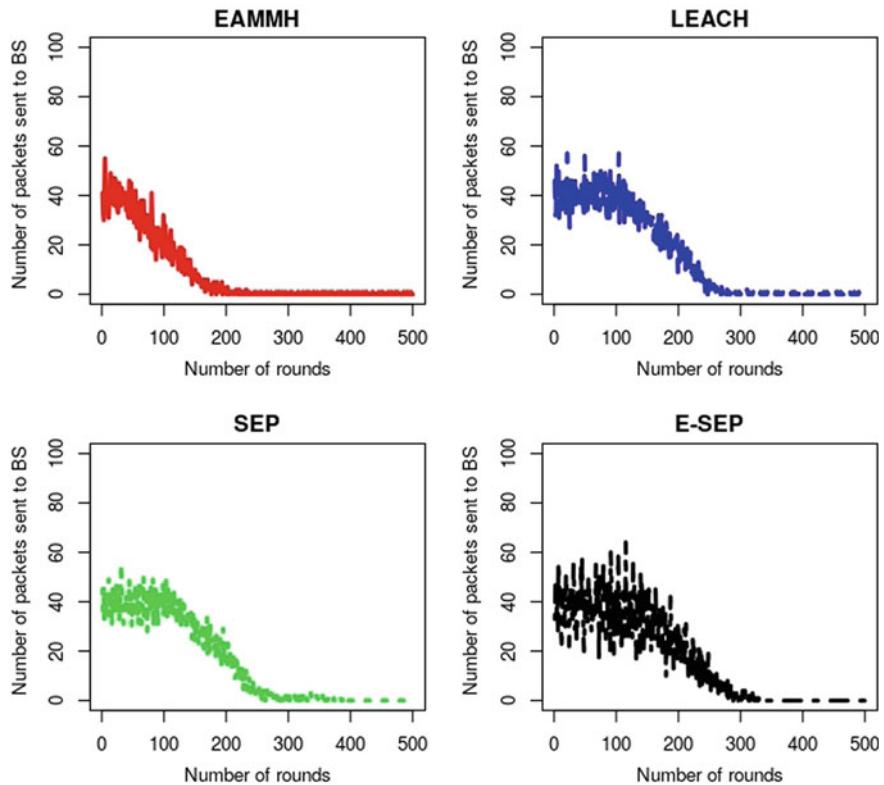
**Fig. 5** Total residual energy**Fig. 6** Number of cluster heads (CHs) at each round



**Fig. 7** Number of packets sent to CHs

the beginning for each sensor node; when the sensor node does not have any energy, it will be represented by 0%. E-SEP protocol succeeds to save more energy than rest of protocols.

From Fig. 6, we can see that protocol E-SEP generates many head clusters at each round. By using many CHs, the transmission of data will be assured by many sensor nodes to the BS. This strategy can minimize the consumption of energy in comparison with the other protocols. Figures 7 and 8 show the number of packets sent to CHs and to BS at each round, respectively. It is clear that E-SEP has more traffic than LEACH, EAMMH, SEP protocols.



**Fig. 8** Number of packets sent to BS (sink)

#### 4 Conclusion

WSNs are very used today in different domains. The routing task is very essential in WSN. Many protocols have been proposed over years. The most effective protocols are based on clustering, and the most known is LEACH protocol. In this paper, we have presented some WSN routing protocols. We have compared LEACH, EAMMH, SEP, and E-SEP protocols. Based on the two most important metrics which are network lifetime and residual energy, we could say that E-SEP performs very well than LEACH, EAMMH, and SEP.

## References

1. Kuorilehto, M., Hännikäinen, M., Hämäläinen, T.D.: A survey of application distribution in wireless sensor networks. *EURASIP J. Wirel. Commun. Netw.* **5**, 1–15 (2005)
2. Smys, S., Bashar, A., Haoxiang, W.: Taxonomy classification and comparison of routing protocol based on energy efficient rate. *J. ISMAC*. **3**(2), 96–110 (2021)
3. Kaur, L., Kaur, R.: A Survey on energy efficient routing techniques in WSNs focusing IoT applications and enhancing fog computing paradigm. *Glob. Transit, Proc* (2021)
4. Schwiebert, L., Gupta, S. K. S., Weinmann, J.: Research challenges in wireless networks of biomedical sensors. In: Proceedings of 7th ACM International Conference on Mobile Computing and Networking (MobiCom '01), pp. 151–165, Rome, Italy (2001)
5. Kafi, M.A., Challal, Y., Djenouri, D., Doudou, M., Bouabdallah, A., Badache, N.: A study of wireless sensor networks for urban traffic monitoring: applications and architectures. *Procedia comput. Sci.* **19**, 617–626 (2013)
6. Dowlatshahi, M.B., Rafsanjani, M.K., Gupta, B.B.: An energy aware grouping memetic algorithm to schedule the sensing activity in WSNs-based IoT for smart cities. *Appl. Soft Comput.* **108**, 107473 (2021)
7. WATERSENSE project consortium. [www.projectwatersense.nl](http://www.projectwatersense.nl). Last accessed 4 Sep 2021
8. Wang, H., Han, G., Zhu, C., Chan, S., Zhang, W.: TCSLP: a trace cost based source location privacy protection scheme in WSNs for smart cities. *Future Gener. Comput. Syst.* **107**, 965–974 (2020)
9. Chen, Y., Zhao, Q.: On the lifetime of wireless sensor networks. *IEEE Commun. Lett.* **9**, 976–978 (2005)
10. Verdone, R., Dardari, D., Mazzini, G., Conti, A.: Wireless sensor and actuator networks technologies, analysis and design. Academic Press, London (2008)
11. Wang, Q., Balasingham, I.: Wireless sensor networks—An introduction. In Tan, Y.K. (eds.) *Wireless Sensor Networks: Application-Centric Design*, pp. 1–13. InTech. (2010)
12. Jacob, I., Jeena, Ebby Darney, P.: Artificial bee colony optimization algorithm for enhancing routing in wireless networks. *J. Artif. Intell.* **3**(01), 62–71. (2021)
13. Heinzelman, W.R., Chandrakasan, A., Balakrishnan, H.: Energy-efficient communication protocol for wireless microsensor networks, in: Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, vol. 1, pp. 10, IEEE Computer Society, Maui, HI, USA (2000)
14. Heinzelman, W.B., Chandrakasan, A.P., Balakrishnan, H.: An application-specific protocol architecture for wireless microsensor networks. *IEEE Trans. Wirel. Commun.* **1**(4), 660–670 (2002)
15. Smaragdakis, G., Matta, I., Bestavros, A.: SEP: a stable election protocol for clustered heterogeneous wireless sensor networks. In Second International Workshop on Sensor and Actor Network Protocols and Applications (SANPA 2004), vol. 3 (2004)
16. Aderohunmu, F.A., Deng, J.D.: An Enhanced Stable Election Protocol (SEP) for Clustered Heterogeneous WSN. University of Otago, New Zealand, Department of Information Science (2009)
17. Mundada, M.R., CyrilRaj, V., Bhuvaneswari, T.: Energy aware multi-hop multi-path hierarchical (EAMMH) routing protocol for wireless sensor networks. *Eur. J. Sci. Res.* **88**(4), 520–530 (2012)
18. Rappaport, T.S.: *Wireless Communications: Principles and Practice*, vol. 2. Prentice Hall PTR, New Jersey (1996)

# Matching Forensic Composite Sketches with Digital Face Photos: A Bidirectional Local Binary Pattern-Based Approach



H. T. Chethana and Trisiladevi C. Nagavi

**Abstract** Facial sketches are extensively used by investigators in order to identify the suspects involved in criminal activities. The manual method of identifying suspects is slow and complex. To make the process automated, proposed method attempts to map the computer created composite sketches to face photos automatically. This research work focuses on searching for missing and wanted persons who are involved in criminal activities that in turn assist investigative agencies in locating suspects in a timely manner. Proposed method attempts to address the challenge of mapping composite sketch to facial photos using bidirectional local binary pattern (BLBP). In the proposed method, Viola–Jones algorithm is used to detect composite sketch; feature extraction is done using BLBP; classification and recognition are done using two-dimensional convolution neural networks (2D-CNNs). The experimental results show that BLBP and 2D-CNN combined approach achieves recognition accuracy of 90% in comparison with other existing methods (Han et al. in IEEE Trans. Inf. Forensics Secur. 8, 191–204, 2013; Hochreiter et al. in Neural Comput. 9, 1735–1780, 1997; Paritosh et al.: in International Conference on Biometrics, Phuket, Thailand, pp. 251–256, 2015; Roy, H., Bhattacharjee, D.: Adv. Intell. Syst. Comput. 883, 2019).

**Keywords** Two-dimensional convolution neural networks (2D-CNNs) · Bidirectional local binary pattern (BLBP) · Viola–Jones algorithm · Composite sketch · Digital face photos · Composite sketch with age variations dataset · Feature extraction · AdaBoost algorithm

---

H. T. Chethana (✉) · T. C. Nagavi

Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysore, Karnataka, India

e-mail: [chethanaht@vvce.ac.in](mailto:chethanaht@vvce.ac.in)

T. C. Nagavi

e-mail: [trisiladevi@sjce.ac.in](mailto:trisiladevi@sjce.ac.in)

H. T. Chethana

Department of Computer Science and Engineering, S. J. College of Engineering, JSS Science and Technology University, Mysore, Karnataka, India

## 1 Introduction

Facial composite sketches are extensively employed to assist identification of criminals when facial image of the suspect is not available at the time of the crime. This system is significant because in most cases the facial photograph of a suspect is not available [4, 20, 32]. Sketches are of two kinds: composite and hand drawn. Composite sketches are generated by taking the description from eyewitness with the help of software tools such as FACES and EvoFIT [18]. Alternatively, hand drawn sketches are generated by professional face drawing artists based on eye witness description. Preparation of composite sketches takes less time, effort, and cost when compared to hand drawn sketches [7].

The real-world examples of composite sketches with digital face photos are shown in Fig. 1. Figure 1a and c shows composite sketches of sample face 1 and 2, whereas Fig. 1b and d demonstrate the corresponding digital face photos. Composite sketches do not contain miniature feature details and lacks in texture information; as a result, these sketches look artificial. But, in digital images, representation of information is very rich. Due to this diversity, it is a highly challenging task to match composite sketch with corresponding digital face photos [5, 6].

Once a sketch of suspect's face is created, authorities will think that using these sketches someone will recognize the individual and provide relevant information which helps the investigating agencies to find the suspects quickly. The eyewitness or victim's description becomes the only form of evidence available [18]. But, this process is inefficient and does not leverage. In particular, law enforcement agencies maintain extensive mug shot databases.

**Fig. 1** Sample pair of composite sketch with digital faces photos [8, 18]



Sample 1: a) Composite sketch b) Digital face photo



Sample 2: c) Composite sketch d) Digital face photo

In accordance to that proposed system focuses on automatic system of mapping composite sketch with digital face photo using BLBP with 2D-CNN. The body of the paper is structured as follows. The exhaustive survey of literature available on the proposed method is discussed in Sect. 2. Further, proposed methodology is illustrated in Sect. 3. Experimental results and analysis is presented in Sect. 4. Finally, the paper is concluded with concluding remarks and future scope of the work.

## 2 Literature Survey

In this section, the exhaustive literature survey on various methods implemented for matching composite sketch to digital face photos is discussed. Plenty of research works are observed in the field of facial forensic image mapping.

An approach based on fine-tuning dual streams deep network (FTDSDN) with multi-scale pyramid decision (MsPD) was proposed by authors [14] for solving heterogeneous face recognition. The approach which combines CNN and FTDSDN removes nonlinear information and retains discriminative data by Rayleigh quotient. The MsPD adaptively regulates sub-structure weight and obtains strong classification performance. Experimental analysis is performed on CUHK FERET and CASIA NIR-VIS 2.0 databases.

In another work, authors [9, 31] have proposed heterogeneous face recognition (HFR) to recognize the person from facial visible and near-infrared images. The task is challenging as it contains very few training samples. Mutual component convolutional neural network (MC-CNN) is used to tackle these two issues. The MC-CNN incorporates mutual component analysis (MCA) and deep CNN having special fully connected (FC) layer [34]. This FC layer extracts modal-independent hidden factors and updates according to maximum likelihood rather than back propagation which reduces overfitting from limited data [35].

An automatic recognition of photo face from a sketch image was proposed by authors [27] for criminal investigation. Using nearest neighbor algorithm, the feature vectors of both composite sketch and photo were compared. The ‘n’ most similar photos were retrieved and matched using the L1-distance measure. The experiment was carried out on three datasets, namely CUHK, CUFS, and FERET. Proposed method provides promising results compared with other existing methods.

Similarly, a binary method for matching images between sketch and photos of heterogeneous faces was proposed by the authors [24]. The robust binary pattern of local quotient (RBPLQ) is used to extract illumination and noise invariant features. Local quotient (LQ) extracts illumination invariant information. Robust local binary pattern (RLBP) captures LQ variations. The experiments are carried out with NIR-VIS benchmark database, and recognition accuracy of 60.72% is achieved.

An approach based on counter propagation network (CPN) using biogeography particle swarm optimization (BPSO) was proposed by the authors [1] for face recognition based on sketches. Proposed method was used to calculate the mean square error (MSE) between the feature vectors of facial sketch and photo. The BPSO-CPN

method is demonstrated on CUHK and IIITD dataset with 1000 sketches and photos. The experimental results are promising and of high precision nature compared to other existing methods.

A face sketch synthesis method was proposed by the authors [30] to perform the representation of nonlinear mapping between face photos and sketches. The sketches synthesized from exemplar methods are blurred and affected with block artifacts. To improve synthesis performance, joint training scheme is proposed by considering sketches. A joint training is performed with a photo and sketch. The sample is constructed first by combining photo and its sketch using high-pass filtered (HPF) image. A random sampling approach is adopted for each test photo to choose the joint training photo and sketch in the neighboring region. The results of the experiments which are performed on public datasets reveal the superiority of the joint training model.

In another work, authors [22] have proposed the concept of multi-scale Markov random fields RF model and facial landmarks (MRF-FLs) method. The combined use of MRF with FL helps in reducing the distortions at the lower part of face contour. Experimental results are evaluated on CUHK and AR face sketch databases [28].

Most of the works [1, 9, 24] have been carried out on heterogeneous face recognition. As per the observation, few works are noted in the area of matching composite sketch with digital face photos on IIITD dataset. Even though these research works [22, 30] are based on matching composite sketch with digital face photos, they are addressing traditional features like illumination and pose variation.

The works proposed in the literature review have focused on matching composite sketch with digital face photos by considering noisy and synthetic sketches. It is observed that none of the works [24, 33] have employed BLBP approach with 2D-CNN. In the current work, an attempt is made to improve the face photo recognition results with composite sketches by employing BLBP approach with 2D-CNN.

### 3 Proposed Methodology

The proposed work aims to address the mapping of composite sketch with digital face photos using supervised learning approach. It uses Viola–Jones algorithm for detecting faces from composite sketches, BLBP for salient feature extraction, classification, and recognition using 2D-CNN. The reasons for adopting these methods are stated below.

Viola–Jones facilitates the detection of faces from the given image. It has high face detection rate, robust and can be used in real-time applications. It is one of the good and popular face detection algorithms. Hence, it is employed in the proposed work for detecting the face of a given composite sketch.

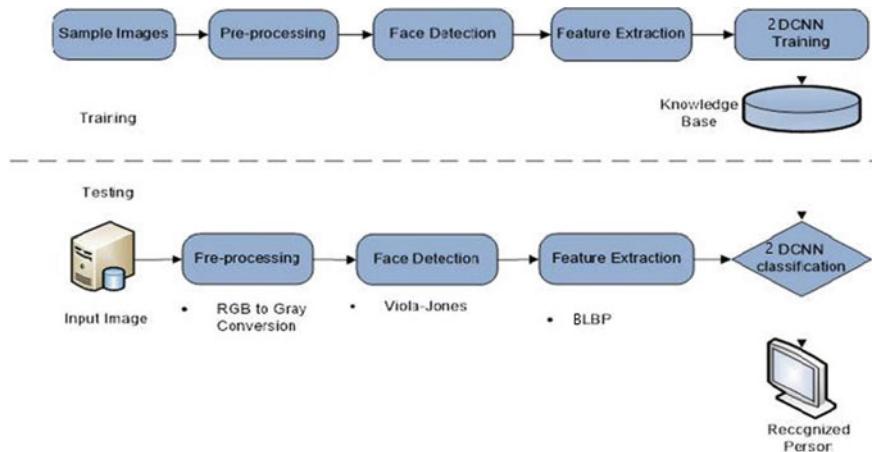
Local binary pattern (LBP) is an effective texture descriptor and helps to capture prominent features using spatial information. Further, the process of feature extraction is made robust by using bidirectional LBP coding [2].

CNNs are type of MLPs which are fully connected networks. The fully connected nature of these networks makes them susceptible to overfitting with the data. Typically, overfitting is avoided by adding some form of magnitude measurement of weights to the loss function. But, CNNs follow a distinctive means to avoid overfitting. They exploit the hierarchical patterns in the input data and congregate multifaceted patterns through simpler and smaller patterns.

Also, the connectivity mode between neurons has resemblance to connectivity between cortical neurons. Here, neurons respond to stimuli only through restricted region of the visual field known as the receptive field. The receptive fields of every neuron participating in CNN partially overlap to cover the whole visual field.

This nature of CNNs makes them simple and efficient to adopt. In the proposed system, two-dimensional convolutional kernels are applied to leverage the spatial and temporal features of the slice so that the classification accuracy improves.

The proposed architecture has training and testing phase. During training phase, preprocessed facial composite sketch is given as input for extracting facial information. Viola-Jones algorithm is one of the most suitable algorithms for facial region of interest extraction. Algorithm detects the face in the frame sequence from preprocessed facial composite sketch. In order to extract the features from the detected composite sketch, BLBP-based approach is used. These extracted features are fed as input to 2D-CNN. Later on, the generated features from the network are stored in database. Testing phase follows all the steps of the training phase on test image sequence. It makes use of the generated features generated during training for classifying the test sample. The architecture of proposed system is shown in Fig. 2.



**Fig. 2** Architecture of the proposed system

## Algorithm to Match Composite Sketch with Digital Face Photos

- Step 1:** Preprocessing technique is applied to composite sketch (input image) and corresponding digital face photo. In preprocessing stage, threshold is set by varying the illumination.
- Step 2:** Digital face photos of corresponding sketches are trained by using a combination of BLBP—a feature extraction technique and 2D-CNN.
- Step 3:** Composite sketch is fed as input to the proposed system in the testing phase.
- Step 4:** For every composite sketch, corresponding match is computed and digital face photo bearing highest accuracy is displayed.

### 3.1 Preprocessing

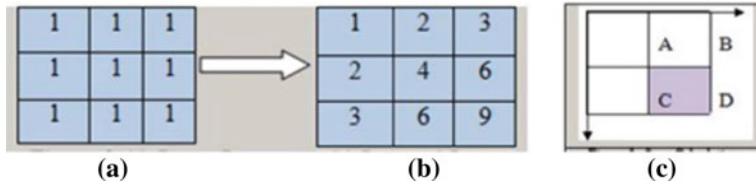
Preprocessing is the series of operations performed on image data which reduces noise, removes unwanted distortions, and enhances some image features which improves the quality of the image. This step is very important for further processing, and it is needed on color, gray level, or binary images. In this stage, each facial photo and composite sketch are resized and transformed into two dimensional grayscale images. In the proposed methodology, simple preprocessing is used because CNN learns through filters. This is the salient feature of the CNN, and it is a major advantage.

### 3.2 Face Detection

Viola–Jones algorithm [29] is used to detect face from preprocessed digital face image and composite sketch. This algorithm exhibits high detection accuracy. Even in composite sketch images, face is identified irrespective of the texture, color, motion, position, and size. The main goal of this approach is to simply differentiate faces from non-faces. It consists of four stages: integral image conversion, Haar feature selection, AdaBoost training, and cascading classifiers.

#### 3.2.1 Integral Image Conversion

It is the first step in Viola–Jones algorithm which converts the original image into an integral image. Integral image facilitates to calculate the Haar features quickly. Figure 3a, b, and c depicts input image, integral image, and area of rectangle, respectively. An integral image shown in Fig. 3b is obtained by converting all pixel values of input image shown in Fig. 3a equal to sum of the pixels left and above of the associated pixel values. It is depicted in Fig. 3b, and it helps to evaluate the rectangle



**Fig. 3** **a** Input image, **b** integral image representation, **c** area of the rectangle

features. Figure 3c denotes the overlapping of rectangle corners of the input image with the integral image pixel values.

Each rectangle area is computed by using the integral image. Figure 3c indicates the purple rectangular area =  $D + A - B - C$ , where rectangle  $B, C$  includes rectangle  $A$ ; hence,  $A$  is added in the calculation.

### 3.2.2 Haar Feature Selection

Haar features are evaluated in two ways (i) by computing each rectangle area, multiplying each rectangle area with their individual weights, then results are combined (ii) by taking the black and white rectangles difference. The obtained Haar feature values [17] are given as input to the cascaded classifier. Haar filters are applied onto one special area, and Haar feature values are calculated by taking the difference of pixels from black and white area as shown in Eq. (1).

$$\text{Haar Feature value} = \sum (\text{black pixels area}) - \sum (\text{white pixels area}) \quad (1)$$

### 3.2.3 AdaBoost Algorithm

In this step, weighted weak classifiers are combined to construct AdaBoost. A weak classifier is mathematically represented in Eq. (2).

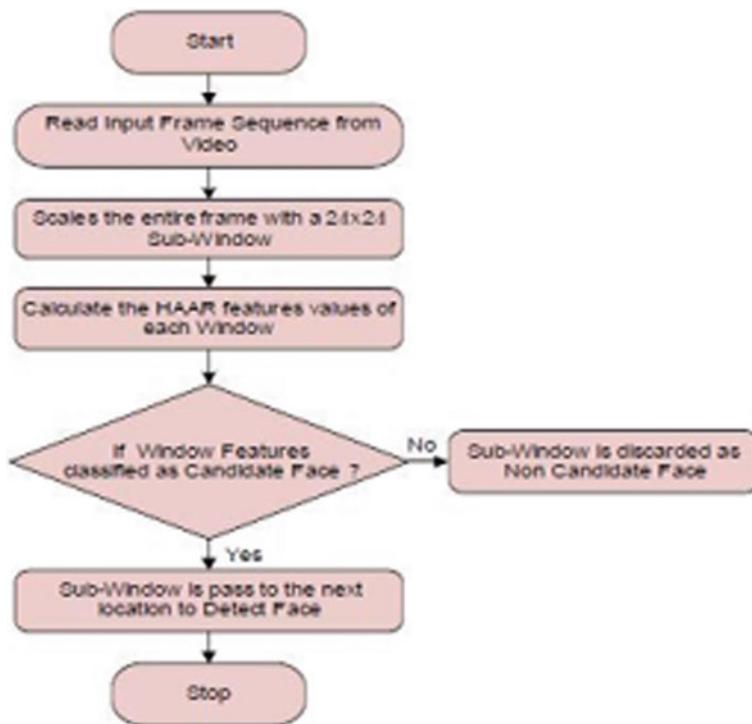
$$h(s, f, p, \theta) = \begin{cases} 1 & \text{if } pf(s) > p\theta \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where  $s$  denotes  $24 \times 24$  sub-window,  $f$  indicates the Haar feature,  $p$  denotes the polarity, and  $\theta$  indicates the threshold to classify whether  $s$  is face or non-face. From the obtained Haar feature values, best features are selected by modifying AdaBoost algorithm.

### 3.2.4 Cascading Classifier

Here, the obtained Haar feature values are given as input to the cascaded classifier. The cascading classifier is constructed with many stages, and each stage has a strong classifier whose work is to compute and classify whether the given input sub-window is positive (face) or negative (non-face). If the input sub-window for a given stage is classified as non-face, then immediately, the given sub-window is discarded. Conversely, if input sub-window is categorized as a face, then this face sub-window is passed to a subsequent stage of cascaded classifier.

Initially, frame sequences are taken as input than a  $24 \times 24$  sub-window which slides over the entire frame to calculate the Haar feature values. Using cascading classifier, each Haar feature is classified and used to detect whether it is composite face or not. If it is a composite face, then the sub-window is passed to the next location, otherwise sub-window is discarded. The flow diagram of face detection is shown in Fig. 4.



**Fig. 4** Flow diagram of face detection

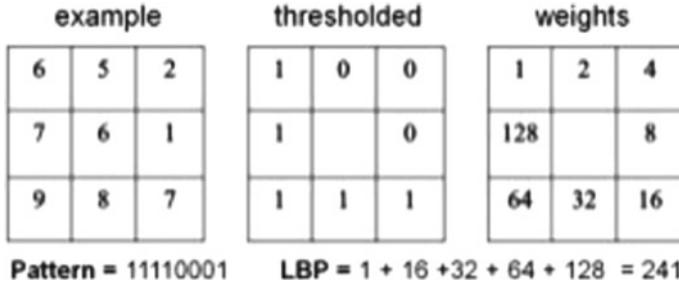


Fig. 5 Local binary pattern

### 3.3 Feature Extraction

The proposed approach uses BLBP to extract the features. Unidirectional LBP remembers only the past information from the input but to predict both past and future information a bidirectional LBP is implemented [11, 15, 16].

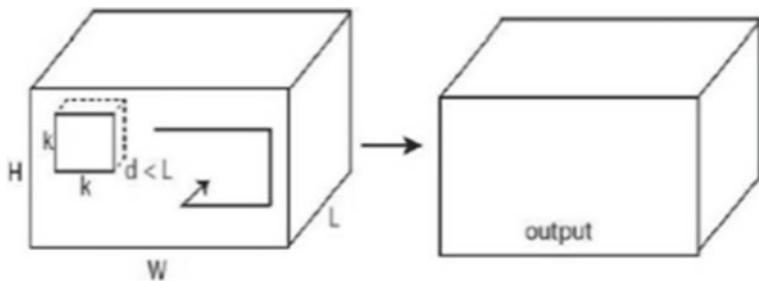
Spatial structure features of the local image such as texture are extracted by using LBP. For a center pixel value  $t_c$ , the neighboring pixels  $t_i$  are equally spaced. LBP works by thresholding the neighbors  $t_i$  with the center pixel value  $t_c$  to generate a bit binary pattern. The output of LBP for  $t_c$  is expressed in decimal form. Output binary pattern is obtained as  $d_i = t_i - t_c$  the difference between neighboring and center pixel,  $d_i = 1$  if  $d_i \geq 0$  and  $d_i = 0$  if  $d_i < 0$ . The LBP uses only the magnitude information while ignoring sign information [10]. The pictorial representation of LBP is given in Fig. 5.

### 3.4 Two-Dimensional Convolutional Neural Networks

The basic CNN is used to collect only spatial information but ignores the temporal information. To overcome this problem, the 2D-CNN approach uses a 2D object as input which collects both spatial information and temporal relations [25, 26].

The 2D-CNN filter/kernel slides on the temporal and spatial dimension of input which enables the neural network to collect more feature vectors. The extracted features map contains the temporal information essential for labeling the frame sequence. It contains 2D feature values for input data. The convolution and pooling layers of 2D-CNN work in cubic style [19]. Figure 6 depicts the 2D convolution of multiple frames. Height ( $H$ ) represents the height of each frame; length ( $L$ ) represents the depth of stack of frames, and  $W$  represents width of each frame.

The input layers of 2D-CNN accept the image as input to represent it in the matrix form. The array of the pixel represented in matrix form is given as input to the convolution layer. This layer uses a convolutional function defined as  $y = \text{conv}(x, w)$  where  $x$  denotes input data,  $w$  denotes the convolution filter, and  $y$  represents the



**Fig. 6** 2D convolution on multiple frames

convolution filter output. Convolution filters are also referred to as kernel or neuron. For 2D input data ' $x$ ', the dimensions are  $Z \times M \times N \times K \times S$ , where  $Z$  defines the feature length values,  $M$  depicts height,  $N$  depicts width of the feature maps,  $K$  represents the image channel number, and  $S$  defines the batch size. These dimensions are same for convolution filter with height, width, and length.

The feature maps are extracted by applying these convolution filters. Rectified linear unit (ReLU) is an activation function in neural network which will not trigger all the neurons at the same time [3, 23]. It works six times faster when generally compared with other activation functions such as sigmoid and tangent hyperbolic (tanh). The rectified output features are given as input to pooling layer to decrease the amount of parameters, spatial size, and computation of the network [36].

## 4 Experimental Results and Analysis

In this section, a discussion is made on the results of the experiments conducted to study the performance of the proposed method based on composite sketch with age variations (CSA) dataset [8, 18]. The dataset is collected from IIIT Delhi which consists of three age variations: same, old, and young. Fifty pairs of composite sketch with same age variations and its corresponding digital face photos are considered for experimentation. The training and testing set consists of digital face photos and composite sketches, respectively. Few sample composite sketches are shown in Fig. 7.

During testing phase, composite sketches are fed as input to the proposed system. It is compared with all the digital face photos stored in the training set. Proposed system extracts the matching digital face photo having highest percentage of matching. Sample composite sketch, its matching digital photo, and corresponding recognition accuracy are shown in Figs. 8 and 9. The plot of recognition accuracy for matching the given input composite sketch with all the digital face photos is shown in Fig. 10. In this plot, values along X- axis denote various digital face photos, and Y-axis values denote the matching accuracy in percentage.



**Fig. 7** Sample composite sketches used for testing



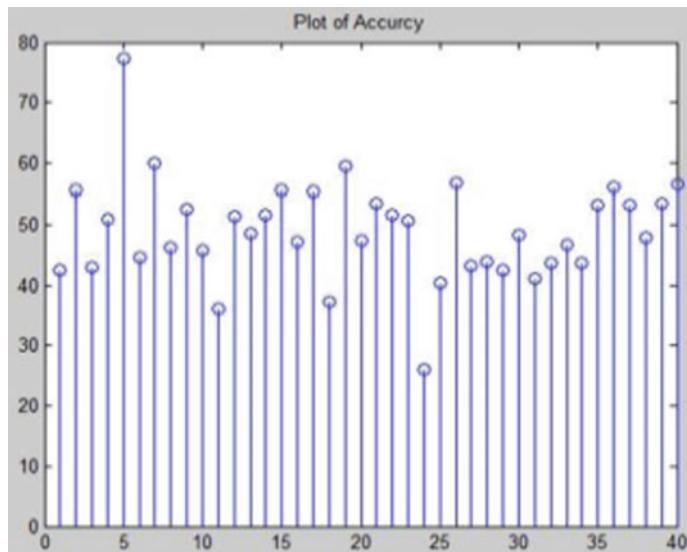
**Fig. 8** Sample input and its corresponding output image

```

Select Query
DCNN Classification
Percentage of Matching:
77.3639

```

**Fig. 9** Sample percentage of matching



**Fig. 10** Sample plot of accuracy in percentage

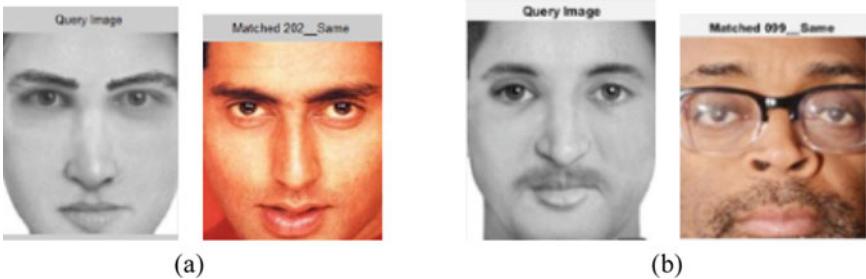
The performance analysis of the proposed method is carried out using the confusion matrix. Values are shown in Table 1. The standard yardsticks used for analyzing the results are true positive (TP) and false positive (FP) which are shown in Table 1.

Fifty pairs of composite sketches with digital face photos were considered for experimentation. Out of 50 pairs, 45 pairs were correctly matched and predicted as positive class (TP). Remaining five pairs were incorrectly matched and predicted as negative class (FP). Results of the sample cases are portrayed in Fig. 11 for correct and incorrect matches.

The accuracy of our proposed system is 90%, and it is calculated using Eq. (3).

**Table 1** Confusion matrix

	True positive	False positive	Accuracy
Results	45	05	90%



**Fig. 11** Sample results using the proposed method **a** correct match **b** incorrect match

$$\text{Accuracy} = \text{TP}/(\text{TP} + \text{FP}) \quad (3)$$

Details of recognition accuracy for fifty pairs of composite sketch and digital face photo are shown in Table 2. Same is pictorially represented in the graph presented in Fig. 12. Here, values along X-axis represent sample numbers, and values across Y-axis denote corresponding accuracy. In Fig. 10, the plot is drawn based on the values obtained for matching one composite sketch with all the digital face photos. Conversely, in Fig. 12, values in X-axis denote various composite sketches, and Y-axis represents recognition accuracy obtained for matching its corresponding digital face photos.

Comparative study is conducted on existing methods and proposed method. It is observed that proposed work provides a recognition accuracy of 90% compared with other existing methods indicated in Table 3. Same is pictorially represented in Fig. 13.

The system performance is appraised through exhaustive testing and results analysis. The experimental results prove that BLBP and 2D-CNN combined approach achieves recognition accuracy of 90% in comparison with other existing methods as shown in Fig. 13. Here,  $x$  denotes names of various methods, and y-axis portrays recognition accuracy. The system can be made scalable for large database of composite sketch-based recognition.

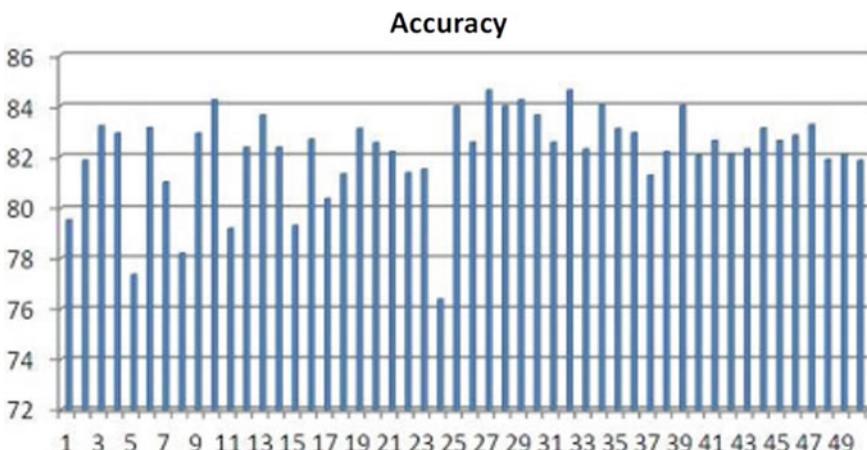
## 5 Conclusion

This research paper aims to solve the problem faced by forensic facial experts in manual identification of suspects. To address this issue, a new approach for facial photo identification based on composite sketch is proposed which is helpful for forensic study. The approach uses BLBP and 2D-CNN for feature extraction and classification, respectively. The experimental results of the proposed approach envisage an appreciable recognition accuracy of 90% in comparison with other existing methods [12, 13, 21, 24]. The task of suspect's face analysis based on the composite

**Table 2** Details of recognition accuracy for fifty pairs of composite sketch and digital face photo

S. No.	Image No.	Accuracy	Output image	S. No.	Image No.	Accuracy	Output image
1	85	79.53	085_Same	26	211	82.5516	211_Same
2	86	81.84	086_Same	27	212	84.6122	212_Same
3	87	83.2013	087_Same	28	213	83.9789	213_Same
4	88	82.9092	088_Same	29	214	84.2316	214_Same
5	89	77.3639	089_Same	30	215	83.6245	215_Same
6	90	83.1291	090_Same	31	216	82.545	216_Same
7	91	81.0389	091_Same	32	217	84.6122	217_Same
8	92	78.2006	092_Same	33	218	82.2792	218_Same
9	93	82.9092	093_Same	34	219	84.0183	219_Same
10	94	84.2316	094_Same	35	220	83.093	220_Same
11	95	79.1883	095_Same	36	221	82.9256	221_Same
12	96	82.3481	096_Same	37	222	81.3014	222_Same
13	97	83.6245	097_Same	38	223	82.1709	223_Same
14	98	82.3481	098_Same	39	224	83.9986	224_Same
15	99	79.3064	099_Same	40	225	82.0233	225_Same
16	201	82.6664	201_Same	41	226	82.6205	226_Same
17	202	80.3827	202_Same	42	227	82.0987	227_Same
18	203	81.3539	203_Same	43	228	82.2891	214_Same
19	204	83.093	204_Same	44	229	83.1061	229_Same
20	205	82.5319	225_Same	45	230	82.6205	230_Same
21	206	82.1709	206_Same	46	231	82.8305	204_Same
22	207	81.4064	207_Same	47	232	83.2537	204_Same
23	208	81.5409	208_Same	48	233	81.8723	233_Same
24	209	76.3828	209_Same	49	234	82.0528	211_Same
25	210	83.9789	210_Same	50	235	81.8297	214_Same

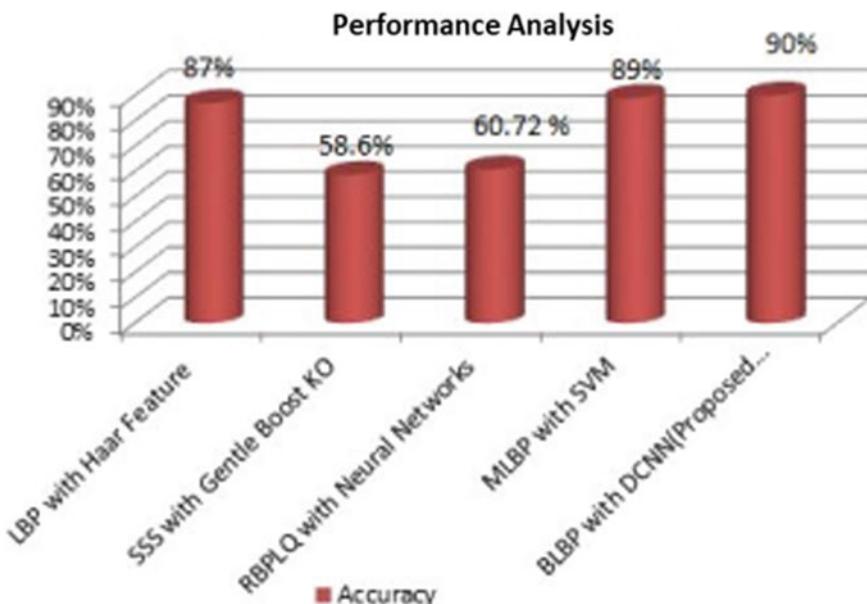
sketch is addressed and explored for the purpose of criminal identification. The proposed BLBP and 2D-CNN combined algorithm significantly outperforms the state-of-the-art criminal face analysis techniques. The system can be made scalable for large database of composite sketch-based face recognition.



**Fig. 12** Sample plot of accuracy for 50 samples pairs

**Table 3** Comparison of proposed method with existing approaches

S. No.	Author	Feature extraction	Classifier	Accuracy (%)
1	Hochreiter et al. [11]	LBP	Haar feature	87
2	Paritosh et al. [19]	Single shot detector	Gentle boost KO	58.6
3	Roy et al. [21]	RBPLQ with LQ	Neural networks	60.72
4	Han et al. [10]	MLBP	Support vector machine	89
<b>5</b>	<b>Proposed work</b>	<b>BLBP</b>	<b>2D-CNN</b>	<b>90</b>



**Fig. 13** Comparison of proposed method with other existing approaches

## References

1. Agrawal, S., Singh, R.K., Singh, U.P., Jain, S.: Biogeography particle swarm optimization based counter propagation network for sketch based face recognition. *Multimedia Tools Appl.* **78**, 9801–9825 (2018)
2. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: application to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 2037–2041 (2006)
3. Cambria, E., Hazarik, D., Poria, S., Hussain, A., Subramanyam, R.B.V.: Benchmarking Multimodal Crime Analysis, pp. 166–179. Springer Nature (2017)
4. Chethana, H.T., Trisiladevi, C.N.: Face recognition methods for facial image analysis in forensics. In: Proceedings of 3rd International Conference on Electrical, Electronics, Communication, Computer Technologies & Optimization Techniques, p. 56. Mysuru, India (2018)
5. Chethana H.T., Nagavi, T.C.: Face recognition for criminal analysis using Haar Classifier. *i-Manager's J. Comput. Sci.* **8**(1) (2020)
6. Chethana, H.T., Nagavi, T.C.: A new framework for matching forensic composite sketches with the digital images, IJDGF Special Issue Submission: Advanced Digital Forensic Techniques for Digital Traces, vol. 13, Issue 5, Article 1 (2021)
7. Chugh, T., Bhatt, H.S., Singh, R., Vatsa, M.: Matching age separated composite sketches and digital face images. In: Proceedings of 6th International Conference on Biometrics: Theory, Applications & Systems. Arlington, VA, USA (2018)
8. Chugh, T., Singh, M., Nagpal, S., Vatsa, M.: Transfer learning based evolutionary algorithm for composite face sketch recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA (2017)
9. Deng, Z., Peng, X., Li, Y., Qiao, Y.: Mutual component convolutional neural networks for heterogeneous face recognition. *IEEE Trans. Image Process.* **28**, 3102–3114 (2019)

10. Frinken, V., Uchida, S.: Deep BLBP neural networks for unconstrained continuous handwritten text recognition. In: Proceedings of 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 911–915. IEEE, NW Washington, DC, United States (2015)
11. Graves, A., Jaitly, N., Mohamed, A.R.: Hybrid speech recognition with deep bidirectional LSTM. In: Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 273–278 (2013)
12. Han, H., Klare, B.F., Bonnen, K., Jain A.K.: Matching composite sketches to face photos: a component based approach. *IEEE Trans. Inf. Forensics Secu.* **8**, 191–204 (2013)
13. Hochreiter, S., Schmidhuber, J., Mehrl, K.: Long short-term memory neural computation. *Neural Comput.* **9**, 1735–1780 (1997)
14. Hu, W., Hu, H.: Fine tuning dual streams deep network with multi-scale pyramid decision for heterogeneous face recognition. *Neural Process. Lett.* **50**, 1465–1483 (2019)
15. Karim, F., Majumdar, S., Darabi, H., Chen, S.: LBP fully convolutional networks for time series classification. *IEEE Access* **6**, 1662–1669 (2018)
16. KaaeSonderby, S., KaaeSonderby, C., Nielsen, H., Winther, O.: Convolutional LBP networks for subcellular localization of proteins. In: Proceedings of International Conference on Algorithms for Computational Biology, pp. 68–80. Springer, Cham (2015)
17. Ma, S., Bai, L.: A face detection algorithm based on Adaboost and new Haar-like feature. In: Proceedings of 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), pp. 651–654. Beijing (2016)
18. Nagpal, S., Singh, M., Singh, R., Noore, A., Majumder: A face sketch matching via coupled deep transform learning. In: Proceedings of International Conference on Computer Vision (ICCV), pp. 5419–5428. Venice, Italy (2017)
19. Ogawa, A., Hori, T.: Error detection and accuracy estimation in automatic speech recognition using deep bidirectional recurrent neural networks. *Speech Commun.* **89**, 70–83 (2017)
20. Patil, S., Shibhangi, D.C.: Composite sketch based face recognition using ANN classification. *Int. J. Sci. Technol. Res.* **9**, 42–50 (2020)
21. Paritosh, M., Vatsa, M., Singh, R.: Composite sketch recognition via deep network—a transfer learning approach. In: International Conference on Biometrics, pp. 251–256. Phuket, Thailand (2015)
22. Radman, A., Suandi, S.A.: Markov random fields and facial landmarks for handling uncontrolled images of face sketch synthesis. *Pattern Anal. Appl.* **22**, 259–271 (2019)
23. Rosas, V.P., Mihalcea, R., Morency, L.P.: Multimodal crime analysis of Spanish online images. *IEEE Intell. Syst.* **28**, 38–45 (2013)
24. Roy, H., Bhattacharjee, D.: Heterogeneous face matching using robust binary pattern of local quotient: RBPLQ. *Adv. Intell. Syst. Comput.* 883 (2019)
25. Sainath, T.N., Vinyals, O., Senior, A., Sak, H.: Convolution long short-term memory, fully connected deep neural networks. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4580–4584. United States (2015)
26. Salama, S.E., Shoman, M.E., WahbyShalaby, M.A.: EEG-based emotion recognition using 2D convolutional neural networks. *Int. J. Adv. Comput. Sci. Appl.* **9**, 329–337 (2018)
27. Setumin, S., Suandi, S.A.: Cascaded static and dynamic local feature extractions for face sketch to photo matching. *IEEE Access* **7**, 27135–27145 (2019)
28. Trisiladevi, C.N., Bhajantri, N.U.: Overview of automatic Indian music information recognition, classification and retrieval systems. In: Proceedings of International Conference on Recent Trends in Information Systems (ReTIS). Kolkata, India (2011)
29. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. I-I. Kauai, HI, USA (2011)
30. Wan, W., Lee, H.J.: A joint training model for face sketch synthesis. *Appl. Sci.* **9**, 1731 (2019)
31. Wang, J., Yang, Y., Mao, J., Haung, Z., Haung, C., Xu, W.: CNN-RNN a unified framework for multi-label image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2285–2294 (2016)

32. Xu, J., Xue, X., Wu, Y., Mao, X.: Matching a composite sketch to a photographed face using fused HOG and deep feature models. *The Visual Computer* (2020). <https://doi.org/10.1007/s00371-020-01976-5>
33. Xu, X., Li, Y., Jin, Y.: Hierarchical discriminant feature learning for cross-modal face recognition. *Multimedia Tools Appl.* (2019)
34. Zhao, F.P., Li, Q.N., Chen, W.K., Liu, Y.F.: An efficient sparse quadratic programming relaxation based algorithm for large-scale MIMO detection. *arXiv e-prints*, [arXiv:2006.12123](https://arxiv.org/abs/2006.12123) (2016)
35. Zhang, Y., Gao, S., Xia, J., Liu, Y.F.: Hematopoietic hierarchy: an updated roadmap. *Trends Cell Biol.* **28**, 976–986 (2018)
36. Zhang, M., Wang, N., Li, Y., Gao, X.: Neural probabilistic graphical model for face sketch synthesis. *IEEE Trans. Neural Netw. Learn. Syst.* **31**, 2623–2637 (2019)

# Reduced Complexity of LDPC Codes using Hard Decision Decoder



Allu Swamy Naidu, Appala Naidu Tenu, and Ajeet Singh

**Abstract** Low-density parity-check (LDPC) codes, invented in the 1960s by Gallager, are a significant topic in coding theory and have large number of practical applications in various scientific domains. Different variants of LDPC codes are developed in the literature which are having better performing encoding as well as decoding procedures. The design of these LDPC coding algorithms is in such a way that it can get back the original message codeword even in the case of massive amount of noise in the communication channel or missing some codeword bits. Therefore, LDPC codes are perfect candidates for real-time applications. This paper presents an efficient message passing procedure to perform the decoding of LDPC codes. The proposed procedure results in the reduction of computational complexity. This procedure exploits hard decision decoding method. The experimental simulation is also performed using additive white Gaussian noise (AWGN) channel through which we obtained different bit error rate (BER) values corresponding to different threshold and signal-to-noise ratio (SNR) values. The adaptability of this simulation is in practical usage in wireless sensor network scenario. In the experiment results, it is observed that with the increase of SNR values in the 1–30 db, BER values are decreased.

**Keywords** LDPC · Binary erasure channel · Binary symmetric channel · Hard decision decoding · Bit error rate · Tanner

---

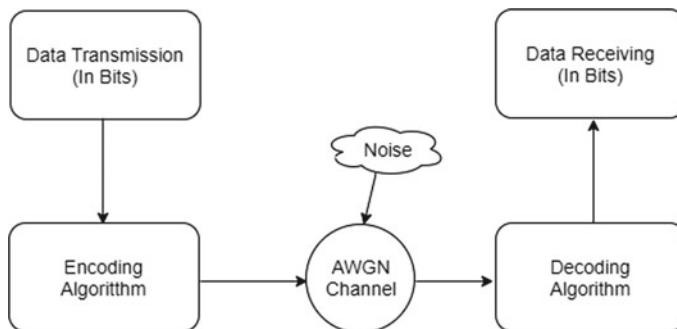
A. S. Naidu (✉)  
Acharya Nagarjuna University, Guntur 522510, India  
e-mail: [alluswamynaidu33@gmail.com](mailto:alluswamynaidu33@gmail.com)

A. S. Naidu · A. N. Tenu · A. Singh  
C.R. Rao Advanced Institute of Mathematics, Statistics, and Computer Science, University of Hyderabad Campus, Hyderabad 500046, India  
e-mail: [ajeetcs@uohyd.ac.in](mailto:ajeetcs@uohyd.ac.in)

## 1 Introduction

Hamming code, proposed by Richard [1], can detect two-bit of error, and it can correct one bit of error because of bounded redundancy while decoding. In other words, hamming code can only perform detection and correction of error in the case of lower error rates. For multiple error correction (BCH), codes are used, given by Bose et al. [2] in 1960.

The limitation of BCH code is that the code length cannot be selected randomly. BCH codes are less practical for longer codewords. To overcome this drawback, LDPC codes were invented in 1962 by Gallegar [3]. These codes are forward error-correcting codes. These codes fall in the category of linear block codes and further classified into regular and irregular LDPC codes [4–10]. In general, messages are transmitted in the form of zeros and ones over the communication channel. In these binary string of codewords, number of zeros are more as compared to ones. LDPC codes provide some computational benefits which are—comparatively better block error performance, support to higher data rates, higher block lengths. Along these benefits, LDPC codes give almost linear time complexity for decoding algorithms. Due to these benefits, LDPC codes are majorly used in high-speed telecommunication applications such as 4G/5G networks and digital video broadcasting systems (DVBS). Decoding of LDPC codes is performed by bit-flipping method which was invented by Gallager [3], that gives low complexity. Bit-flipping method provides lesser bit error rate performance as well as facilitates to design of the simpler decoder. Figure 1 shows the basic data communication system. In this data transmission system, data in bits format is input to the encoding algorithm. After encoding, this data goes through AWGN channel (some noise will be added). The decoding algorithm removes the noise and recovers original data in bits.



**Fig. 1** Data transmission system

## 1.1 Related Work

This section presents some significant developments in this domain over past years. In 1999, Fossorier et al. [11, 12] were given a simplified version of belief propagation decoding algorithm using the mean sum approximation technique. Even though it performs the reduction in computational complexity, the performance while decoding is degraded. Authors in [13–17] proposed various techniques which result into an optimal performance while decoding process. Authors in [18, 19] discussed the different methods about the LDPC codes. Authors in [20] explain the systematic LDPC codes with hard decision message passing algorithm for non-binary symbols. [21] addresses how the information storage bit-flipping decoder for LDPC codes is discussed. In [22], solving the NAND flash using non-binary low-density parity-check hard decision technology. In [23] for systematic LDPC on Raspberry Pi boards, how to implement non-binary encoder and decoder. In [24] explains the capable architecture for stochastic LDPC decoder with good bit error rate (BER) performance. In [25, 26], using symbol flipping approach, solving the problem of decoding non-binary LDPC over finite field GF( $q$ ). Chen et al. [27] proposed a scheme to investigate the block length and suitable BER in their design. Their simulation results show that the performance of LDPC codes is comparatively better than the other codes having shorter block length. Authors in [28] have compared their proposed work on G.9959 protocol in wireless sensor network with other trust routing protocols.

## 1.2 Organization of the Paper

Rest of the paper is organized as—Sect. 2 provides an overview of LDPC codes representation. LDPC construction methods, encoding, and computational complexity analysis are given in Sect. 3. Message passing decoding algorithms such as hard decision decoding and bit-flipping decoding are given in Sect. 4. The obtained simulation results are covered in Sect. 5. Finally, Sect. 6 concludes the paper.

## 2 LDPC Codes Representation

LDPC codes represent in mainly two different forms. One is characterized using matrices and another is using graph's representation.

## 2.1 Through Matrix Description

A parity-check matrix is represented in Eq. 1.

$$H_R = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \end{pmatrix} \quad (1)$$

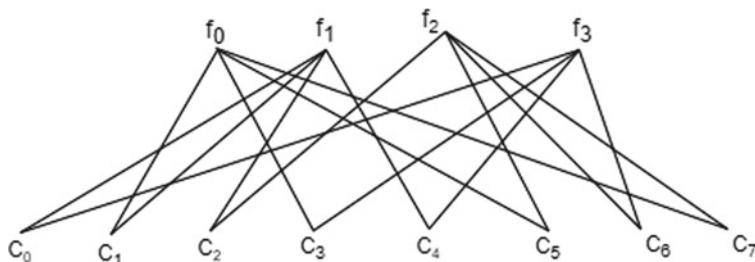
It is having its dimension ( $n \times m$ ) as 8 columns and 4 rows. In this represented matrix,  $W_r$ ,  $W_c$  denotes—total number of 1's in  $r$ th row and  $c$ th column, respectively. If  $W_c$  is much lesser than  $n$  and  $W_r$  is also much lesser than  $m$ , then given matrix in Eq. 1 is called as low-density matrix.

## 2.2 Through Graphical Description

An adequate graphical description for LDPC codes was given by Tanner [29] in 1981. Tanner graphs follow the property of bigraph. Tanner graphs not only depict an overall description of the code, in addition to this, they also assist to illustrate the decoding algorithm.

The shortest cycle in Fig. 2 is as follows:  $c_1 \rightarrow f_1 \rightarrow c_2 \rightarrow f_2 \rightarrow c_5 \rightarrow f_0 \rightarrow c_1$ . Tanner graphs having shortest cycles paths must be ignored because they reflect worst performance while decoding process. In Fig. 2,  $f_0$  to  $f_3$  represent 4 check nodes (C-nodes) and  $c_0$  to  $c_7$  represent 8 variable (V-nodes). Each element entry in  $H$  matrix is represented by  $h_{ij}$  and denoted as follows:

$$h_{ij} = \begin{cases} 1, & \text{if } f_i \text{ is connected to } c_j \\ 0, & \text{otherwise} \end{cases}$$



**Fig. 2** Tanner graph representation for Eq. 1 matrix

### 2.3 Regular and Irregular LDPC Codes

Suppose in  $H$  matrix, if total ones count in individual column is same and  $mW_r = nW_c$  then  $H$  is called regular LDPC codes. The example representation in equation 1 is regular LDPC code. Through graphical description, we can observe that each c-node has same vertex degree as 4 and each v-node has same vertex degree 2.

In case of irregular LDPC codes, total ones count in individual column is not same. For irregular code  $m \left( \sum_i h_{i,i} \right) = n \left( \sum_i v_{i,i} \right)$ , where  $v_{i,i}$ -fraction of columns of weight and  $h_{i,i}$ -fraction of rows of weight. The following matrix  $H_I$  is an example of irregular LDPC codes.

$$H_I = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 \end{pmatrix} \quad (2)$$

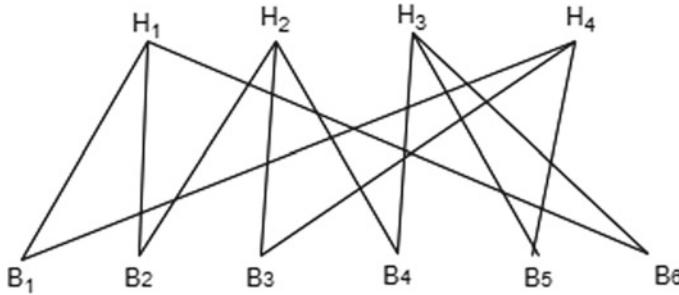
## 3 LDPC Construction Methods

In 1996, MacKay and Neal [30] proposed another construction method for which the steps are as below

- Assigned all zero's in the parity-check matrix  $H$ . After that bits are flipped randomly but bits not distinct.
- Columns in  $H$  matrix are being added from left to right direction once at a time.
- In the matrix  $H$ , having weight  $j$  per column and equal weight in each individual row and none of the two columns must be joined to at most one. This will avoid four-cycle length.

These LDPC codes show better performance converging to Shannon's limit. An example construction with length 12 (3,4)-regular parity-check matrix is shown below.

$$H = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ \hline 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ \hline 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$



**Fig. 3** Tanner graph

### 3.1 Encoding

Consider a code  $C = [c_1, c_2, c_3, c_4, c_5, c_6]$  which is having the string of length six satisfies the following parity-check equations (Fig. 3):

$$\begin{aligned} c_1 \oplus c_2 \oplus c_4 &= 0 \\ c_1 \oplus c_3 \oplus c_5 &= 0 \\ c_1 \oplus c_2 \oplus c_3 \oplus c_6 &= 0 \end{aligned}$$

Above three parity-check equations can be represented in the matrix form as below:

$$\begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \end{pmatrix} = 0$$

Any binary code having  $m$ -parity-check equations and  $n$ -codewords represents  $m \times n$  binary parity-check matrix. In the above parity-check matrix  $H$ , 3 rows represent a parity-check equations and 6 columns represent codewords. Consider  $Y$  with code length as six will be called as a valid codeword corresponding to  $H$  if and only if it satisfies  $H.Y^T = 0$ . The above parity-check equations can also be represented as

$$\begin{aligned} c_4 &= c_1 \oplus c_2 \\ c_5 &= c_1 \oplus c_3 \\ c_6 &= c_1 \oplus c_2 \oplus c_3 \end{aligned}$$

Here codeword bits  $c_1, c_2, c_3$  hold three bit message and  $c_4, c_5, c_6$  hold three parity-check bits. This manner code word is encoded as follows.

$$(c_1 \ c_2 \ c_3 \ c_4 \ c_5 \ c_6) = (c_1 \ c_2 \ c_3) \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}$$

above matrix  $G$  is the generator matrix for the given code. In general,  $k$  message bits can be represented by the vector  $v = (v_1, v_2, \dots, v_k)$  then the codeword can be found using matrix equation  $C = v.G$ . For such binary code having  $k$  number of message bits and codewords of length  $n$ , matrix  $G$  is binary matrix of dimension  $k \times n$ . Rate of the code is calculated as  $\frac{k}{n}$ . There are  $2^k$  possible codewords for  $k$  message bits, which is subset of all possible  $2^n$  vectors of length  $n$ .

The code is said to be systematic if  $k$  codeword bits from starting consist of message bits. For such systematic codes, first  $k$  columns of the generator matrix are in the form of identity matrix of size  $k$ . On the parity-check matrix  $H = [A, I_{n-k}]$  (here A: Binary matrix of dimension  $(n - k) \times k$ ;  $I_{n-k}$ : Identity matrix of  $(n - k)$  dimension) and apply the Gauss–Jordan elimination method to obtained generator matrix  $G = [I_k, A^T]$ . Any generator matrix  $G$  and parity-check matrix  $H$  follow the orthogonality property, i.e.,  $G.H^T = 0$ .

Parity-check matrix is an important choice for construction of LDPC codes. For a code, out of all parity-check constraints,  $n - k$  are linearly independent. Parity-check matrix rank is  $n - k$ .  $r_2(H)$  represents number of rows in  $H$ , and these number of rows are linearly dependent on  $GF(2)$ . Above procedure can be shown using example as below.

$$H = \begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

In this above example  $n = 10$ ,  $k = 5$ , rate =  $\frac{5}{10}$  and we need to encode the LDPC code.

- As first step matrix  $H$  is converted into row-echelon form. This is achieved by applying some elementary row operations. In the  $H$  matrix first and second row is kept same as one's are already present in the diagonal. Interchanging  $3^{rd}$  and  $5^{th}$  rows. Now perform sum of  $1^{st}$  and  $4^{th}$  rows under  $GF(2)$  and replaced by  $4^{th}$  row. Interchanging  $4^{th}$  and  $5^{th}$  rows. Now resultant matrix  $H_{r_e}$  is given as follows.

$$H_{r_e} = \begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}$$

- Next we perform row operations on above matrix to get reduced row-echelon  $H_{rr_e} = (I|A)$

$$H_{rr_e} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}$$

- At last standard form of parity-check matrix is obtained as  $H_{\text{std}} = (A|I)$

$$H_{\text{std}} = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

- Finally, the generator matrix  $G$  derived from  $H_{\text{std}}$  where  $G = (I|A^T)$

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

- Matrix  $G$  is given to the encoder and matrix  $H$  is given to the decoder

## 4 Applications of LDPC Codes

The usefulness and practicality of LDPC codes have attracted the scientific community and researchers to apply them in various application domains. Some of its applications are mentioned below.

1. LDPC code in Wiretap channel
2. LDPC code in Code-Based Cryptosystems
3. LDPC code in MIMO-OFDM system

Some of the other applications of LDPC codes are - DVB-S2 standard for the satellite transmission of digital television, 10GBase-T Ethernet, Wi-Fi 802.11 standard, FPGA-based decoding systems.

## 5 Message Passing Decoding Algorithms

In this type of decoding, messages are being passed both forward and backward directions between the bit nodes and check nodes repeatedly till the result is obtained. These algorithms are also called as iterative decoding algorithms. Some of these algorithms are bit-flip decoding and belief propagation decoding. In bit-flip decod-

ing, messages are in binary format. Whereas belief propagation decoding algorithm message is in the form of probabilities that represent the acceptance level of codeword bits.

### 5.1 Message Passing Procedure on Binary Erasure Communication Channel

For a transmitted bit in the BEC channel, there are two possibilities—either bit is received in the correct manner or the bit is erased with probability  $p$ . The prime function of the decoder is to find the erased bit and flipped bit. The message  $M'$  is passed among bit nodes and check nodes of the Tanner graph. This message  $M'_i$  corresponding to the  $i$ th bit node acknowledges the bit value as 1 or 0 in case if it is known else  $x$  if it is unknown. If only single ‘ $x$ ’ bit of the message is received by check node, then parity-check equations are used to calculate the value of ‘ $x$ ’; otherwise, check node sends it back to the remaining connected bit nodes and calculate the unknown value. This message is denoted as  $E_{ij}$ ,  $i$ th check node to  $j$ th bit node. This procedure continues iteratively till all the unknown bit values are recovered. When some channels like AWGN and binary symmetric introduced the errors at receiver ends. Then the messages in this decoding process give perfect guesses for the bit values of the codeword depending upon the present information at every node.

**Example:** Consider the following Tanner graph in Fig. 4.

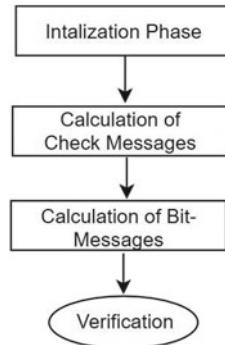
Consider a code  $C = [c_1, c_2, c_3, c_4, c_5, c_6]$  which is having the string of length six satisfies the following parity-check equations:

$$\begin{aligned} c_1 \oplus c_2 \oplus c_4 &= 0 \\ c_2 \oplus c_3 \oplus c_4 &= 0 \\ c_4 \oplus c_5 \oplus c_6 &= 0 \\ c_1 \oplus c_3 \oplus c_5 &= 0 \end{aligned}$$

Notation  $H_i$  denotes set of bits in the  $i$ th parity-check equation.  $H_1 = \{1, 2, 6\}$ ,  $H_2 = \{2, 3, 4\}$ ,  $H_3 = \{4, 5, 6\}$ ,  $H_4 = \{1, 3, 5\}$ . In similar way, notation  $B_j$  denotes set of parity-check equations corresponding to  $j$ th bit.  $B_1 = \{1, 4\}$ ,  $B_2 = \{1, 2\}$ ,  $B_3 = \{2, 4\}$ ,  $B_4 = \{2, 3\}$ ,  $B_5 = \{3, 4\}$ ,  $B_6 = \{1, 3\}$ .

Consider a codeword which we want to encode is  $C = [1 0 1 1 0 1]$ . This codeword was sent over the erasure communication channel and received the codeword  $R = [1 0 1 x x x]$ . Now to recover these erased bits, we use the above-discussed message passing procedure. In this method,

- Since first check node is connected with 1st, 2nd, and 6th bit nodes and so have incoming message 1, 0,  $x$
- Calculate the value of erased bit  $x$  using  $E_{1,6} = R_1 \oplus R_2 = 1 \oplus 0 = 1$
- Now updated received code vector  $R = [1 0 1 x x 1]$
- Second check node is connected with 2nd, 3rd, and 4th bit nodes and so have incoming message 0, 1,  $x$

**Fig. 4** Flowchart

- Calculate the value of erased bit  $x$  using  $E_{2,4} = R_2 \oplus R_3 = 0 \oplus 1 = 1$
- Now updated received code vector  $R = [1 0 1 1 x 1]$
- Continue the same process until remaining erased bits are recovered
- Finally algorithm returns the decoded codeword  $R = [1 0 1 1 0 1]$
- It can be observed that erased bits are recovered correctly.

## 5.2 Hard Decision Decoding

In this method, fixed values in binary format (0, 1) are taken as input. This decoding process can be explained in the following example. In Fig. 1, consider a code  $C = [c_0, c_1, c_2, c_3, c_4, c_5, c_6, c_7]$  which is having the string of length eight. Suppose original message which was send through BHC channel is  $C = [0 0 1 1 1 0 0]$  and received codeword is  $R = [0 0 0 1 1 1 0 0]$ .

- In the first step, each v-node( $f_i$ 's) receives the values from c-nodes as Table 1
- In the second step, as v-node  $f_0$  is connected to  $c_1, c_3, c_5, c_7$  then the value sent from  $f_0$  to  $c_1$  is obtained by XOR operation.  $c_3 \oplus c_5 \oplus c_7 = 1 \oplus 1 \oplus 0 = 0$ . Similarly, remaining values are obtained which are representing in below table
- In the third step, for the v-node  $c_0$ , we perform a look up into Table 2.  $c_0$  is presented in  $f_1$  and  $f_3$ . Therefore, the values assigned to this v-node are  $f_1 \rightarrow 1, f_3 \rightarrow 0$ . Continue this procedure for all the remaining node values.
- Now in the final step, majority voting is performed on the bit values
- In this manner obtained the decoded value, i.e.,  $R = [0 0 1 1 1 0 0]$

**Table 1** Received values from c-nodes

	Received values from c-nodes
$f_0$	$c_1 \rightarrow 0, c_3 \rightarrow 1, c_5 \rightarrow 1, c_7 \rightarrow 0$
$f_1$	$c_1 \rightarrow 0, c_2 \rightarrow 0, c_4 \rightarrow 1, c_0 \rightarrow 0$
$f_2$	$c_2 \rightarrow 0, c_5 \rightarrow 1, c_6 \rightarrow 0, c_7 \rightarrow 0$
$f_3$	$c_0 \rightarrow 0, c_3 \rightarrow 1, c_4 \rightarrow 1, c_6 \rightarrow 0$

**Table 2** Sent values from v-nodes

	sent values from v-nodes
$f_0$	$0 \rightarrow c_1, 1 \rightarrow c_3, 1 \rightarrow c_5, 0 \rightarrow c_7$
$f_1$	$1 \rightarrow c_0, 1 \rightarrow c_1, 1 \rightarrow c_2, 0 \rightarrow c_4$
$f_2$	$1 \rightarrow c_2, 0 \rightarrow c_5, 1 \rightarrow c_6, 1 \rightarrow c_7$
$f_3$	$0 \rightarrow c_0, 1 \rightarrow c_3, 1 \rightarrow c_4, 0 \rightarrow c_6$

c-node	Connected to v-node	received $R$	values assigned to v-nodes	Majority vote
$c_0$	1,3	0	$f_1 \rightarrow 1, f_3 \rightarrow 0$	0
$c_1$	0,1	0	$f_0 \rightarrow 0, f_1 \rightarrow 1$	0
$c_2$	1,2	0	$f_1 \rightarrow 1, f_2 \rightarrow 1$	1
$c_3$	0,3	1	$f_0 \rightarrow 1, f_3 \rightarrow 1$	1
$c_4$	1,3	1	$f_1 \rightarrow 0, f_3 \rightarrow 1$	1
$c_5$	0,2	1	$f_0 \rightarrow 1, f_2 \rightarrow 0$	1
$c_6$	2,3	0	$f_1 \rightarrow 1, f_3 \rightarrow 0$	0
$c_7$	0,2	0	$f_0 \rightarrow 0, f_2 \rightarrow 1$	0

### 5.3 Bit-flipping Decoding

Bit-flipping decoder is a hard decision information passing procedure used in LDPC codes. As soon as valid codeword is obtained by scrutinizing the satisfiability of parity-check equations, the decoder can be instantly stopped the process. It has two significant advantages: one is additional iterative steps can be skipped as soon as the solution is obtained. Second is that it is able to detect the failure while concur to codeword. Algorithm 1 represents the bit-flipping process. Flowchart is given Fig. 4.

### 5.4 Decoding Process

Consider the example discussed in Section 2.1. A codeword which we want to encode is  $C = [1\ 0\ 1\ 1\ 0\ 1]$ . This codeword sent over the BSC channel and received the codeword  $R = [0\ 0\ 1\ 1\ 0\ 1]$ .

**Algorithm 1:** Bit-flipping Algorithm

---

**Result:** Decode  $R$

initialization;

**for**  $i=1$  to  $n$  **do**

- |  $M_i = r_i$ ;
- end**

repeat ;

Calculations Check messages;

**for**  $j=1$  to  $m$  **do**

- | **for**  $i=1$  to  $n$  **do**

  - | |  $E_{j,i} = \sum_{i' \in B_j, i' \neq i} (M_{i'} \bmod 2)$
  - | **end**

- | **end**

Calculations Bit messages;

**for**  $i=1$  to  $n$  **do**

- | **if** the messages  $E_{j,i}$  disagree with  $y_i$  **then**
- | |  $M_i = (r_i + 1 \bmod 2)$ ;
- | **else**
- | | go to next value
- | **end**

**end**

Verification;

**for**  $j=1$  to  $m$  **do**

- |  $L_j = \sum_{i' \in B_j} (M_{i'} \bmod 2)$

**end**

**if**  $L_j = 0$  or  $I = I_{max}$  **then**

- | finished

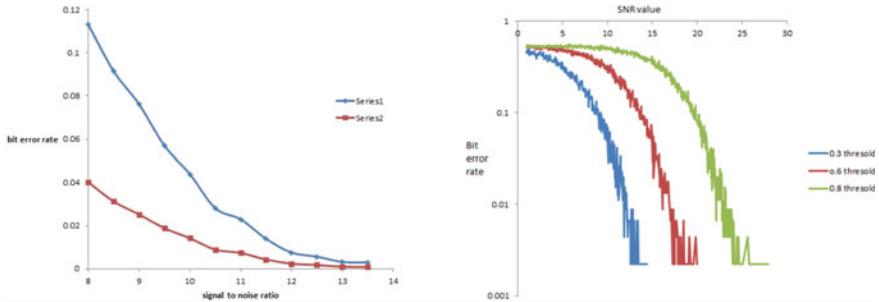
**else**

- | |  $I = I + 1$

**end**

---

- In step 1, the messages for check nodes are computed. For the first check node, the messages are  $E_{1,1} = 1, E_{1,2} = 1, E_{1,6} = 0$ . Repeating the same process, we can obtain the values for remaining check nodes are  $E_{2,2} = 0, E_{2,3} = 1, E_{2,4} = 1, E_{3,4} = 1, E_{3,5} = 0, E_{3,6} = 1, E_{4,1} = 1, E_{4,3} = 0, E_{4,5} = 1$ ,
- In the step 2, the majority voting is performed and value 1 is obtained and flips the bit value as  $0 \rightarrow 1$ .
- Now received codeword is  $R = [1\ 0\ 1\ 1\ 0\ 1]$ . These values will be checked in the parity-check equations.
- If it satisfies all the parity-check equations, then procedure can be stopped else continue the step 2.



(a) The bit error rate performance of BFA decoding on an AWGN channel.  
Here threshold 1(series 1), 0.5(series 2)

(b) The bit error rate performance a  
of bit-flip de-coding on an AWGN channel

**Fig. 5 a** Bit error rate performance of BFA decoding on an AWGN channel. Here threshold 1(series 1), 0.5 (series 2). **b** The bit error rate performance of bit-ip de- coding on an AWGN channel

## 5.5 Computational Complexity Analysis

This subsection summarizes the computational complexity in terms of practicality of hard decision decoding algorithm. We also discussed the drawbacks of non-binary coding methods and how the proposed decoding-based algorithm is comparatively efficient in terms of less computational complexity and feasibility in real-time scenarios. The complexity of a codeword which is multiplied with a matrix mainly relies upon number of 1's present in the matrix. If sparse matrix  $H$  has its form  $(S^T | I)$ . After performing Gaussian elimination, the matrix  $G$  is computed as  $G = (I | S)$ . Since  $S$  is not sparse matrix in general, therefore the computational complexity is much higher. If the block length of the code is much higher, then sparse matrices will also not show the better performance due to increase complexity of  $O(n^2)$ . To overcome this drawback, we use iterative decoding and iterative encoding algorithms. These algorithms carry out internal calculations and then send these results through messages.

## 6 Simulation Results

The above-presented algorithm is simulated using *C* language. For the simplicity, we assume all the data bits as zero, and we assumed channel as AWGN. After bits are transmitted through channel, we performed hard decision threshold on received bits. After that, we performed BFA on received bits. By changing the SNR ratio, we observed the error behavior in terms of bit error rate of the decoding algorithm is calculated. The bit error rate can be calculated as the number of 1's in the received data to the total number of bits received.

Different threshold values we observe the SNR and BER values. Here we fixed the SNR values range which is 0–30 db with increment of 0.05 db, and number of

**Table 3** Comparison between SNR and BER values

Threshold value	SNR ratio (dB)	BER values
0.3	1	0.406666667
	2	0.390555556
	3	0.365
	.	.
	14.45	0.0016666667
0.6	1	0.498888889
	2.05	0.477222222
	3	0.474444444
	.	.
0.8	1	0.498888889
	2.05	0.477222222
	3	0.474444444
	.	.
	19.95	0.00222222222

simulations is 300. This output observes in the following table. In the case of increase of the SNR value, BER value is decreased. In this Table 3, we observe that for the 0.3, 0.6, 0.8 threshold values increase the SNR values from 1 to 30 db in increment of 0.05 the BER value be decreasing. Here maximum iterations taken into consideration are 100.

## 7 Conclusion

This paper presents an optimal message passing procedure to perform the decoding of LDPC codes. The major purpose is to correct the bits from error channel. As we are receiving the error bits from channel, we first perform hard decision decoding and bit-flip algorithm on these received bits. As the SNR value increases, the bit error rate decreases. The implementation complexity is very low compared to turbo codes. The complexity is reduced because of hard decision thresholding. At lower SNR values, the algorithm gives less performance as compared to higher SNR values (Fig. 5).

## References

1. Hamming, R.W.: Error detecting and error correcting codes. *Bell Syst. Tech. J.* **29**(2), 147–160 (1950)
2. Bose, R.C., Ray-Chaudhuri, D.K.: *Inf. Control. On a class of error correcting binary group codes* **3**(1), 68–79 (1960)
3. Gallager, R.G.: *Low-Density Parity-Check Codes*. MIT Press, Cambridge, MA (1963)
4. Chang, T.C., Wang, P., Su, Y.T.: IEEE Commun. Lett. Multi-stage bit-flipping decoding algorithms for LDPC codes **23**(9), 1524–1528 (2019)
5. Wang, L., Wang, D., Ni, Y., Chen, X., Cui, M., Yang, F.: China Commun. Design of irregular QC-LDPC code based multi-level coded modulation scheme for high speed optical communication systems **16**(5), 106–120 (2019)
6. Wyner, A.D.: The wire-tap channel,. *B.S.T.J.*, vol. 54, no. 8, pp. 1355–1387 (1975)
7. Thangaraj, A., Dihidar, S., Calderbank, A.R., McLaughlin, S., Merolla, J.-M.: Applications of LDPC codes to the wiretap channel. *IEEE Trans. Inf. Theory* **53**(8), 2933–2945 (2007)
8. McEliece, Robert J.: DSN Prog. Rep. A public-key cryptosystem based On algebraic coding theory (PDF) **44**, 114–116 (1978)
9. Niederreiter, H.: Knapsack-type cryptosystems and algebraic coding theory. *Prob. Control Inf. Theory. Problemy Upravlenija i Teorii Informacii* **15**, 159–166 (1986)
10. Liu, R., Zeng, B., Chen, T., Liu, N., Yin, N.: The application of LDPC code in MIMO-OFDM system. *IOP Conf. Ser.: Mater. Sci. Eng.* **322**(7) (2018)
11. Fossorier, M.P.C., Mihaljevic, M., Imai, H.: IEEE Trans. Commun. Reduced complexity iterative decoding of low density parity check codes based on belief propagation **47**(5), 673–680 (1999)
12. Wiberg, N.: Codes and decoding on general graphs. Ph.D. dissertation. Linköping University, Linköping, Sweden (1996)
13. Eleftheriou, E., Mittelholzer, T., Dholakia, A.: Reduced-complexity decoding algorithm for low-density parity-check codes. *IEE Electron. Lett.* **37**, 102–104 (2001)
14. Chen, J., Fossorier, M.P.C.: Decoding low-density parity-check codes with normalized APP-based algorithm. In: Proceedings, pp. 1026–1030. San Antonio, TX. IEEE Globecom (2001)
15. Hu, X.-Y., Eleftheriou, E., Arnold, D.-M., Dholakia, A.: Efficient implementation of the sum-product algorithm for decoding LDPC codes. In: Proceedings, pp. 1036–1036E. San Antonio, TX. IEEE Globecom (2001)
16. Chen, J., Fossorier, M.P.C.: IEEE Trans. Commun. Near-optimum universal belief-propagation-based decoding of low-density parity-check codes **50**(3), 406–414 (2002)
17. Chen, J., Fossorier, M.P.C.: Density evolution for two improved BP-based decoding algorithms of LDPC codes. *IEEE Commun. Lett.* **6**(5), 208–210 (2002)
18. Li, P., Leung, W.K.: IEEE Commun. Lett. Decoding low-density parity-check codes with finite quantization bits **4**(2), 62–64 (2000)
19. Chen, J., Fossorier, M.P.C.: Density evolution for BP-based decoding algorithms of LDPC codes and their quantized versions. In: Proceedings of IEEE Globecom, Taipei, Taiwan, R.O.C., Nov. 2002, pp. 1026–1030
20. Tootoolavest, U., Manthamkarn, V., Maheshwari, A.: Systematic low density parity check codes with hard decision message passing algorithm for non-binary symbols. In: 2020 8th International Electrical Engineering Congress (iEECON), pp. 1–4 (2020). <https://doi.org/10.1109/iEECON48109.2020.929467>
21. Cui, H., Lin, J., Wang, Z.: Information storage bit-flipping decoder for LDPC codes. In: IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 28, no. 11, pp. 2464–2468, Nov. 2020, <https://doi.org/10.1109/TVLSI.2020.3009270>
22. Ren, J., Ding, X., Xin, X.-N., Chen, H.-H.: An NB-LDPC decoder algorithm combined using channel information for Storage Application. In: 2020 IEEE 5th International Conference on Integrated Circuits and Microsystems (ICICM), pp. 306–309 (2020). <https://doi.org/10.1109/ICICM50929.2020.9292206>

23. Maheshwari, A., Tuntoolavest, U., Fukawa, K.: Implementation of the nonbinary encoder and decoder for systematic low density parity check codes on raspberry-PI boards. In: 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 751–0756 (2020). <https://doi.org/10.1109/IEMCON51383.2020.9284943>
24. Zhang, Q., Chen, Y., Li, S., Zeng, X., Parhi, K.K.: IEEE Trans. Circ. Syst. I: Regular Pap. A high-performance stochastic LDPC decoder architecture designed via correlation analysis **67**(12), 5429–5442 (2020). <https://doi.org/10.1109/TCSI.2020.3003457>
25. Ullah, W., Cheng, L., Takawira, F.: IEEE Access. Predictive syndrome based low complexity joint iterative detection-decoding algorithm for non-binary LDPC codes **9**, 33464–33477 (2021). <https://doi.org/10.1109/ACCESS.2021.3060806>
26. Deng, K., Cui, H., Lin, J., Wang, Z.: Counter random gradient descent bit-flipping decoder for LDPC codes. In: IEEE Computer Society Annual Symposium on VLSI (ISVLSI), pp. 55–60 (2021). <https://doi.org/10.1109/ISVLSI51109.2021.00021>
27. Chen, J., Iong, Z.: 5G systems with low density parity check based channel coding for enhanced mobile broadband scheme. IRO J. Sustain. Wirel. Syst. **2**(1), 42–49 (2020)
28. Kumar, D., Smyr, S.: An efficient packet delivery scheme using trust routing in G. 9959 protocol in a wireless sensor network. J. Ubiquitous Comput. Commun. Technol. (UCCT) **2**(03), 118–125 (2020)
29. Tanner, R.: IEEE Trans. Inf. Theory. A recursive approach to low complexity codes **27**(5), 533–547 (1981)
30. MacKay, D.J.C., Neal, R.M.: Electron. Lett. Near Shannon limit performance of low density parity check codes **32**(18), 1645–1646 (1996)

# A Review on Video Sharing over Content Centric Networks



C. Victoria Priscilla and A. R. Charulatha

**Abstract** The rise in wireless communication technologies has enabled the wireless mobile networks to play a key role in various Internet applications like social networks, mobile sensing, video streaming, etc. With this rising trend in mobile networks, many researchers have moved their focus on metrics to enhance quality of service and experience to the users on video streaming. The network providers need to provide hassle-free streaming, with high-quality resolution, instant start-up, less jitter to stay in competition and to increase their consumer base. This paper gives an overview of metrics required for achieving quality of viewing experience and providing service from the perspective of users, network provider, respectively. Various solutions proposed to overcome the technical hitch faced by network providers in video sharing in mobile networks is reviewed.

**Keywords** Wireless networks virtualization · Video sharing · Quality of experience · Quality of service · Content centric networking

## 1 Introduction

The developments of creating, distributing of available videos and user-created video contents through services like WhatsApp, YouTube and other social media services have now risen to greater extent. With the COVID pandemic, the streaming service has raised dramatically with the video traffic increase associated with the streaming services as user-created video contents has increased to new levels. The existing distribution networks are under huge pressure to afford exponential increase in bandwidth demands, chiefly because of very huge traffic demands linked with television services like Internet protocol television as well as the emergence of new video

---

C. Victoria Priscilla

Department of Computer Science, Shrimathi Devkunvar Nanalal Bhatt Vaishnav College for Women, Chrompet, Chennai 600044, India

A. R. Charulatha (✉)

Stella Maris College for Women(A), University of Madras, Chennai 600086, India

e-mail: [charulatha@stellamariscollege.edu.in](mailto:charulatha@stellamariscollege.edu.in)

streaming services such as Netflix, Amazon and other Skype-like video communications. The new developments in adapting to the advanced video formats, high-bandwidth requests and uninterrupted service have further complicated the situation [1].

‘Global Mobile Data Traffic Forecast 2017–2022’ conducted Cisco—Visual Networking Index (VNI) released on February 2019 which predicts the mobile data traffic to raise by 46% out of which the traffic of Internet video will comprise of 33%. The live video traffic will rise 15 times by 2022.

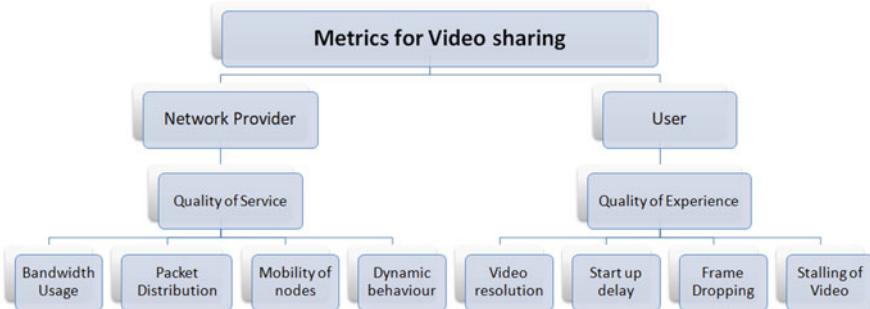
The different formats of IP video will persist to be above 80% of whole of IP traffic. Of the entire mobile data traffic, the IP video traffic would account up to 82% by the year 2022.

Internet video traffic itself is diversified. The live video over Internet is likely to draw huge amount of traffic as it replaces the usual broadcast screening hours. Also, the growth of dropcams, the video surveillance traffic, is to be considered as this kind of traffic is different than the video streamed on demand or live videos and represents a constant stream uploaded continuously to the cloud from small businesses and homes. In recent years, video surveillance has increased to detect unusual events and has become an essential requirement in public places where there is less manpower. The surveillance is automated to reduce the manual labour workload. This has increased the installation of lots of cameras and video streaming [2]. With the exemption of short-form video or video conferencing, the other formats of Internet video do not seem to have a large upstream factor. Initially, whilst content generated or created by users became popular, many expected the traffic to be symmetric, but it is not. The surfacing of subscribers themselves becoming as content producers, but the subscribers still enjoy watching much more videos than they upload. Upstream traffic proportion seems to be declining as a slight percentage for many years [3].

Customers expect advanced video services to be available anytime on diverse devices that too with less energy consumption. Also, the content providers try to operate through many interfaces across networks. Many research challenges like control signalling, isolation, discovery and allocation of resources, mobility of nodes management, network operation and management and security. Also, non-technical issues such as governance regulations remain to be areas of concern in wireless network virtualization before its wide deployment Though research has contributed number of new architect such as content-centric networks (CCNs) [4] and software-defined networking (SDN) [5] which focusses addressing issues like management of extensive growth in bandwidth, the migration to wireless networks and to resolve related security issues, video sharing has been given less attention.

## 2 Metrics for Video Sharing in Wireless Network

Video streaming services mainly depend on the providing high-quality visual content and easy access to the user with less latency, jitter and hassle-free access using mobile devices. The main areas of concern for any network provider are maintaining quality



**Fig. 1** Metrics for video sharing in wireless networks

of experience (QoE) for the user by maintaining quality of service (QoS). Quality of service and quality of experience are the major issues to attract and sustain the customers, apart from having a rise in the profits of network providers and optimizing network resources (see Fig. 1).

For achieving quality of experience, from a user perspective, video streaming sessions must be available anytime and anywhere. The user's viewing experience will be disturbed if the video resolution is low, or many frames are dropped or re-buffering of video which adds a lot of loads on the bandwidth. So, in order to guarantee high-user-perceived quality, the streaming protocols optimize the video resolution (or bit rate) within available bandwidth, avoid video stalling/re-buffering events and too much of frame drops and minimize the start-up delay [6]. For achieving quality of service, from the network services point of view, challenges arise due to lack of control techniques which are efficient in packet distribution, effective usage of bandwidth, mobility of nodes and the wireless resource's dynamic characteristic. Several algorithms and models were proposed with the objective of ensuring the user and Internet with QoE and QoS. Each model states its own parameters and mechanisms [7]. A network that can provide best bandwidth with less latency to many users simultaneously and also securely is the primary requirement.

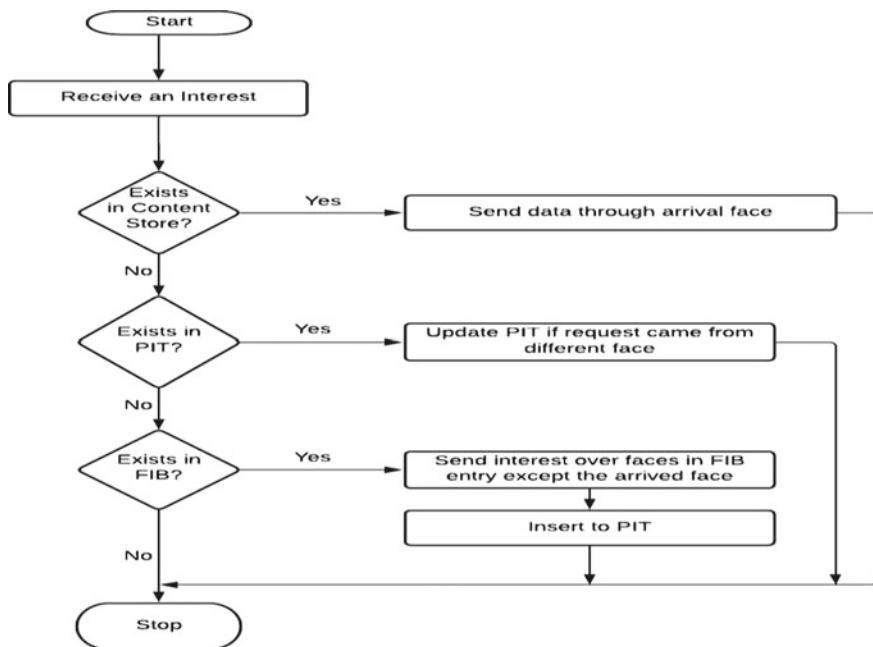
## 2.1 IP-Based Networking versus Content Centric Networking

Internet is a packet-switched network. Every piece of online content is kept on a user's server (host), most often that of the creator. Other users in the network must approach to that server with their requests to retrieve that particular content. The packets must then be sent as a response to the requestor. The all-to-cache method which is conventional caching method in mobile CCN (MCCN)—utilizes large amount of storage resources of each node in order to maintain content fetching from nearby nodes to offload the traffic in the network. The low storage capacities and the energy consumption of mobile devices does not support the high level of consumption. Also, the nodes being the content carriers, the movement of carriers moves the

content in geographic area, thereby leading to undesirable effect for efficiency on content caching. In addition, the conventional broadcasting-based content lookup approach wastes the bandwidth to a greater extent. The network congestion due to a greater number of request results in start-up delay. To address these problems encountered with traditional methods, a new framework content centric networking which focusses on content instead of IP/host-oriented architecture is developed, and different caching technique algorithms are designed.

## 2.2 Content Centric Networking

Content centric networking (CCN) is a networking approach that enables network to self-organize and push relevant content at anywhere, anytime. It provides accessibility, integrity and security of data. CCN stores the name and data item mappings. CCN contains two packet types encoded in efficient binary XML which are interest packet and data packet (similar to get and response in http). Each CCN entity operates on three data structures, and they are the content store (CS), the pending interest table (PIT) and the forwarding information base (FIB). The CCN uses longest prefix matching lookup for content names. The working of CCN is represented as flow chart in Fig. 2.



**Fig. 2** CCN process flow

### 3 Literature Review

The review is organized based on neighbour node selection strategies, topology and caching strategies that were used in sharing videos for achieving quality of service and quality of experience in content centric networks. Selection of neighbour based on their common level of interest, the amount of contribution by each of the neighbouring nodes and the maintenance cost is considered. Topology of network has a significant role in determining QoS, and algorithms are designed to accommodate the dynamic topology in heterogeneous networks. Caching strategies and cache replacement policies are keys to have higher throughput, high cache hit ratio and less turnaround time. Algorithms designed on caching and replacement strategies improve energy efficiency of nodes in the networks. Bio-inspired algorithms have also recently come into play for identifying the best route achieving energy efficiency and high throughput.

#### 3.1 *Neighbouring Node Selection Strategies*

Zong et al. [8] proposed a novel solution, CNVD—contribution-aware neighbour assist video delivery where the nodes over a network are allowed to build and retain their relationship along with neighbouring nodes and to exchange the cached videos. In order to achieve maintenance for the video content in CCNs economically as many of the existing solutions were maintaining redundant links between nodes and to have an well-organized lookup and distribution of videos, CNVD proposes a method in which the contribution of the neighbouring nodes is estimated based on its lookup success rate, the total number of videos stored in that node, delay in forwarding the interest packets and video data and the exact distance of the chosen neighbour nodes in both the lookup and transmission paths. A strategy for maintenance of neighbour nodes composed of construction and removal of neighbour relationship in terms of contribution levels of neighbour nodes is proposed, thereby reducing the maintenance cost of neighbour nodes. A unicast video lookup algorithm is proposed which selects the next hop nodes by estimating their levels of contribution like reducing the delay of look up.

Jia et al. [9] proposed a solution ‘demand-aware resource caching optimization’ (VDRCO) for video sharing. The variation of interest amongst users and mobility of nodes is the primary areas to be addressed in promoting the performance of the videos shared in wireless mobile networks. The authors identify that the user demand is based on the influence of the number of initial providers of a particular video, and the attractive content in the videos decides the amount of distribution in the procedure of video circulation.

The entire overlaid network is partitioned into numerous communities. Each community containing nodes with common interests. The clustering of communities with similar interests creates a stable link between nodes in that community,

thereby reducing cost involved for maintenance in those networks, and also achieves efficiently sharing the resources for video systems achieving QoS. A primary issue is how to exactly estimate the interests between nodes. Node similarity is reflected based on the videos watched by the users. When two users watch similar videos, then they are considered to have common interests in viewing videos. In conventional method, this similarity is identified based on the file name which reflects the entire video content. In this method, the accuracy in estimating the similarity of interests is low which leads to weak logical link between nodes. Description of video or short text can be used in estimating interest similarity, but the noise words and data sparseness present in those does not cater the need rather brings in negative influence in estimation accuracy. Using expanded description consumes more energy and the estimation calculation also results in high complexity.

To efficiently promote estimation accuracy of similarity of content in the video, thereby developing a space for video content with rich semantic-based description. Zhang and Xiong [10] in their paper proposed a novel algorithm ‘video community discovery scheme using ontology-based semantical interest capture’ (VCOSI). In order to describe the video content, VCOSI uses a semantical extension approach which is ontology-based and implements three classification methods which are used to select the key words to be added in the video content’s extended description. An estimation method where keywords are used to calculate the similarity amongst video contents. In order to bring down the complexity in calculation of similarity and to bring down the calculation load, the authors further propose an estimation algorithm based on prefix-filtering.

### 3.2 *Topology*

As an alternative of developing a network to reach out an estimated maximum demand, Martin et al. [11] proposed a network topology which is dynamically provisioned to accommodate the changing demands. This paper deals with an algorithm to provide autonomous network management with awareness on QoE which contributes to the network resource allocator its design, implementation and deployment with capability to configure, optimize and heal itself through machine learning, SDN and NFV technologies. The capability to analyze the topology of network, the levels of optimization of resources which are required for 5G media streaming services. The network resource allocator predicts traffic demands using machine learning algorithms and converts them into operational thresholds, and a topology is identified to deliver incoming traffic according to a service level agreement, the operational costs and eventually, to deploy it through the SDN controller.

Han et al. [12] proposed dynamic adaptive streaming via HTTP (DASH) which seemed to be trending in both academy and industry. The main objective of DASH algorithm is to distribute or share the video with high QoE in spite of dynamic network topology conditions. The main idea is that the video encoding is done at different bit rates and different resolutions. The encoding is divided into video segments of

2–30 s in length. The client node takes note of information on the available audio and video streams, its encodings and the duration of segment. Then, the client node requests a video one segment at a time using HTTP. The available bandwidth for the session is detected based on an algorithm on rate adaptation, and the video quality is attuned accordingly.

Meyerson et al. [13] introduced a server-side solution. This solution executes over a 5G system. A dynamic encoding to efficiently utilize memory is applied depending on the user's demand, and also the video encodings are not retained beforehand. This solution is achieved using network function virtualization and software-defined networks. NFV and SDN enable and provide the centralized management of network, and also NFV focusses mainly on optimizing the network services, and a virtual network function is accountable for deploying and running of virtual machines. These virtual network functions can be combined to offer a network service.

Wang et al. [14] in their paper proposed a content distribution algorithm which is highly efficient using dynamic adaptive streaming. The bit rate of the video and the rate of transmission of each node in network is determined. This enables improved throughput and the stalling time of the video is reduced adapting to the dynamic network topology.

### 3.3 *Content Caching Policies*

Noh and Song [15] proposed a progressive caching mechanism based on metafile on content centric network tree with faultless video streaming services. The major contributions of this paper are (i) the caching algorithm creates a metafile intended to reduce the bandwidth required for uninterrupted video streaming services in accordance to the cached data along the route of content delivery is employed in this architecture. (ii) An effective caching information exchanging protocol with no additional network overhead is designed amongst CCN nodes by using the reserved header area of the CCN interest/data packets. The caching scheme proposed modifies the header structure of the interest and data packets and shares the cached information amongst the nodes which is not done in traditional CCN caching method. This technique aids to avoid caching of redundant data in neighbouring nodes.

Sun et al. [16] developed a delay-aware content distribution framework for various nodes in a virtualized mobile network. The authors developed a content placement algorithm to avoid redundancy of content and also minimizing the total visiting time of users across all slices is developed. A clustering algorithm which maximizes the intra-cluster signal to interference plus noise ratio (SINR) which is of low-complexity is used to reduce the downloading time.

Kumar and Tiwari [17] developed a novel content caching strategy which determines the content popularity based on the number of times, and the content is requested and the number of unique contents in the network. This algorithm seems to efficiently cache the popular contents.

Situmorang et al. [18] proposed a combination of least recently used replacement policy with least frequently used policy which is called as caching everything everywhere. This strategy improves the cache hit ratio for smaller cache sizes.

Ji et al. [19] in their paper proposed an algorithm primarily focussing on the content replacement strategy which plays a significant role in enhancing the QoE. The PGR algorithm calculates the content popularity based on the number of the times, and the content is requested by the consumer and the time gap between each requests. This algorithm shows significant improvement in utilization of space in the nodes and high cache hit ratio.

Alotaibi and Alahmadi [20] suggested a caching strategy which decides the node in which the content has to be cached based on the topology of the network and cache replacement strategy which will decide the content to be replaced when requirement for space arises based on the number of resources that are held by the content to stay in that node. The content which utilizes more resources will be considered for replacement.

Li et al. [21] proposed an ant colony optimization inspired routing algorithm. A strategy on content management which stores the content type along with the content name, thereby enabling faster look up and continuous monitoring of content concentration enabling a high cache hit ratio on dynamic environments.

Zhou et al. [22] propose an ant colony inspired caching strategy. The proposed algorithm aims to optimize the network performance with cache allocation solution which is achieved by this ant colony algorithm. A best path is chosen which has the high-pheromone concentration. This algorithm achieves high cache hit ratio, low throughput and good energy efficiency but all these are achieved when the user requests are static. This algorithm is not suitable for dynamic user requests.

Wang and Ding et al. [23] created a novel energy efficient routing algorithm which utilizes artificial bee colony optimization implemented on fuzzy c-means clustering algorithm to find the optimal method for clustering and identifying cluster heads. Communication amongst clusters is achieved by the polling mechanism. This algorithm improves the energy efficiency and produces better throughput.

Zhang et al. [24] suggest using artificial bee colony algorithm which generates a routing table and the routing path which is planned based on the location of sink node. This algorithm also adopts a method to choose cluster heads based on its residual energy and the energy of its surrounding nodes. Sub-cluster heads are also marked, and after maximum replacements in cluster heads, the sub-cluster heads will be promoted as cluster heads. This algorithm is suitable for low-power ad hoc wireless networks.

## 4 Inferences

With respect to maintaining neighbouring node relationships a check for optimality in large network, impact of mobility and application of this solution over dynamic topology is yet to be achieved. With regard to topology, the energy efficiency and

the adaptability to the link quality are yet to be achieved, and the video transcoding services need to be optimized. With regard to caching strategy, it is based on the various factors like popularity of content, bandwidth and latency under certain set of constraints like availability of cache space, content and also careful eviction of the existing cache contents in order to increase content availability and reducing total network utilization and reduced latency. The content of the video is cached based on content popularity. The content caching needs to be addressed over dynamic topology and mobility of nodes. The major limitations on video sharing over content centric networking are the adaptability to the link quality and the topology of network. There is also a drawback of optimization of space consumption of the content delivery networks.

## 5 Summary of Review

The summary of the literature review is presented in Table 1.

## 6 Conclusion

Network virtualization is perceived as an evolutionary paradigm change, but still faces several challenges. This research review's purpose is to present different aspects posed by the research on the video sharing in wireless networks. This important because mobile devices have become an integral portion of everyone's life. Sharing of videos, creating video content, social media sharing, watching videos YouTube, Netflix, etc., whilst travelling has become more common. The covered areas of challenges and effects provide a general view of what is the expectation of the consumer and service provided by the network provider. However, the continuous increase of users and their requirements lets the providers confronted with the necessity to rethink the utilization of existing network technologies to be competitive and at the same time profitable. As the literature review has revealed, there has been much research and discussion conducted on the metrics on video sharing. Most of the research found was based on trying to achieve one metrics or two. It is important to conduct more studies on satisfying additional metrics and achieving QoS and QoE.

**Table 1** Summary of the literature survey

Title	Metrics/issues addressed	Methodology/contributions
A novel contribution-aware neighbour assist video delivery solution over mobile content centric networks	Redundant links between nodes	Efficient content delivery solution which builds neighbour relationship based on level of interest and the lookup capacity. Reduces delay of lookup and transmission of videos by selection of appropriate next hop nodes in terms of contribution levels of neighbour nodes
Video sharing solution based on demand-aware resource caching optimization	Redundant caching, trend-caching and far-end video fetching	Clustering of nodes with analogous video demand to enable distribution optimization and efficient video sharing. Balances load supply and demand using near-end video fetching
Video community discovery scheme using ontology-based semantical interest capture	Delay in transmission, high bandwidth and other resource utilization	A novel solution to achieve efficient resource sharing to promote QOS. To reduce the consumption of energy by mobile nodes. Estimation method makes of selected keywords to calculate similarity of video content
Network resource allocation system for QOE aware delivery of media services in 5G networks	Dynamic topology and efficient resource utilization	Integrates machine learning in SDN controller to forecast resource demands and enables autonomic infrastructure management
AMVS-NDN: Adaptive mobile video streaming and sharing in wireless named data networking	High-video quality and less traffic	Mobile station uses Wi-Fi link opportunistically and share content directly bypassing the base stations
Virtualized dynamic transcoding service for adaptive streaming video over HTI'P in SG systems	Bandwidth consumption and adaptive streaming over HTI'P	SONATA framework for reliable and scalable SG networks
Decentralized asynchronous optimization for dynamic adaptive multimedia streaming over information-centric networking	Distributed asynchronous optimization algorithm	Transmission rate is optimized within the constraints of the available network and preventing stalling of playbacks

(continued)

**Table 1** (continued)

Title	Metrics/issues addressed	Methodology/contributions
Progressive caching system for video streaming services over content centric network	Redundant caching, chunks are stored in the order of priority	Metafile created by scalable caching algorithm to reduce required peak bandwidth required for seamless streaming services. A progressive caching system in which chunks of video are selectively stored based on priority and hop distance. CCN packet header is modified to accommodate distance and priority
Delay-aware content distribution via cell clustering and content placement for multiple tenants	Latency and faster download speed	The scheme on cell clustering along with graph partitioning is proposed to reduce the user download time. Content placement algorithm based on minimum vertex coverage is proposed to avoid replication of content and improved QoE
Optimized content centric networking for future Internet: dynamic popularity window-based caching scheme	Dynamic popularity window-based caching	Content cached based on the uniqueness in the network and the frequency of requests
A Simulation of cache replacement strategy on named data network	Cache everything everywhere	A combination of least recently used and least frequently used policies is applied. Efficient for small cache sizes
replacement-based content popularity and cache gain for 6G content centric network	Popularity and cache gain replacement algorithm	Content popularity and the frequency of content being requested is considered
Efficient caching and replacement strategy in content centric network (CCN) based on Xon-path and hop count	Path and hop count algorithm	The topology of network is considered for choosing a node for caching and the content which holds lots of resources is considered for replacement
AGO-inspired information-centric networking routing mechanism	Ant colony optimization algorithm	Content concentration and content type are considered in caching the content

(continued)

**Table 1** (continued)

Title	Metrics/issues addressed	Methodology/contributions
An ant colony inspired cache allocation mechanism for heterogeneous ICN	Energy efficiency	Implements ant colony optimization: Minimizing energy consumption and CRs load as well as maximizing cache hit ratio and throughput
An energy efficient routing protocol based on improved artificial bee colony algorithm for wireless sensor networks	Artificial bee colony algorithm	Cluster heads and the optimal clustering of nodes are implemented, thereby improving energy efficiency
Seamless clustering multi-hop routing protocol based on improved artificial bee colony algorithm	Energy efficiency	Implements artificial bee colony optimization

## References

- Popescu, A., Yao, Y., Ilie, D.: Video distribution networks: architectures and system requirements. In: *Greening Video Distribution Networks*, pp. 1–23. Springer, Cham (2018)
- Sharma, R., Sungheetha, A.: An efficient dimension reduction based fusion of CNN and SVM model for detection of abnormal incident in video surveillance. *J. Soft Comput. Paradigm (JSCP)* **3**(2), 55–69 (2021)
- <https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html>
- Kurose, J.: Content-centric networking: technical perspective. *Commun. ACM* **55**(1), 116–116 (2012)
- Software Defined Networks. <https://sdn.ieee.org/>
- Pakha, C., Chowdhery, A., Jiang, J.: Reinventing video streaming for distributed vision analytics. In: 10th {USENIX} Workshop on Hot Topics in Cloud Computing (HotCloud 18) (2018)
- Rodrigues, D., Cerqueira, E., Monteiro, E.: Quality of service and quality of experience in video streaming. In: Proceedings of the International Workshop on Traffic Management and Traffic Engineering for the Future Internet (FITraMEn2008). EuroNF NoE, Porto, Portugal (2008)
- Zhong, L., Jia, S., Wang, M.: A novel contribution-aware neighbor-assist video delivery solution over mobile content-centric networks. *Mobile Inf. Syst.* **2016** (2016)
- Jia, S., et al.: A novel video sharing solution based on demand-aware resource caching optimization in wireless mobile networks. *Mobile Inf. Syst.* **2017** (2017)
- Zhang, R., Xiong, S.: A novel mobile video community discovery scheme using ontology-based semantical interest capture. *Mobile Inf. Syst.* **2016** (2016)
- Martin, A., et al.: Network resource allocation system for QoE-aware delivery of media services in 5G networks. *IEEE Trans. Broadcast.* **64**(2), 561–574 (2018)
- Han, B., et al.: AMVS-NDN: Adaptive mobile video streaming and sharing in wireless named data networking. In: 2013 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). IEEE (2013)
- Meyerson, E., et al.: Virtualized dynamic transcoding service for adaptive streaming video over HTTP in 5G systems. In: 2018 International Conference on Smart Communications in Network Technologies (SaCoNeT). IEEE (2018)
- Wang, M., et al.: Decentralized asynchronous optimization for dynamic adaptive multimedia streaming over information centric networking. *J. Netw. Comput. Appl.* **157**, 102574 (2020)

15. Noh, H., Song, H.: Progressive caching system for video streaming services over content centric network. *IEEE Access* **7**, 47079–47089 (2019)
16. Sun, G., et al.: Delay-aware content distribution via cell clustering and content placement for multiple tenants. *J. Netw. Comput. Appl.* **137**, 112–126 (2019)
17. Kumar, S., Tiwari, R.: Optimized content centric networking for future internet: dynamic popularity window based caching scheme. *Comput. Netw.* **179**, 107434 (2020)
18. Situmorang, H., et al.: A simulation of cache replacement strategy on named data network. In: 2018 12th International Conference on Telecommunication Systems, Services, and Applications (TSSA). IEEE (2018)
19. Ji, Y., et al.: Replacement based content popularity and cache gain for 6G content-centric network. *Phys. Commun.* **44**, 101238 (2021)
20. Alotaibi, B., Alahmadi, S.: Efficient Caching and Replacement Strategy in Content Centric Network (CCN) Based on Xon-Path and Hop Count
21. Lv, J., et al.: ACO-inspired information-centric networking routing mechanism. *Comput. Netw.* **126**, 200–217 (2017)
22. Zhou, P., et al.: An ant colony inspired cache allocation mechanism for heterogeneous information centric network. *IEEE Access* **9**, 55485–55496 (2021)
23. Wang, Z., et al.: An energy efficient routing protocol based on improved artificial bee colony algorithm for wireless sensor networks. *IEEE Access* **8**, 133577–133596 (2020)
24. Zhang, T., et al.: Seamless clustering multi-hop routing protocol based on improved artificial bee colony algorithm. *EURASIP J. Wirel. Commun.* **2020**(1), 1–20 (2020)

# An Overview of Augmenting AI Application in Healthcare



Aarthi Chellasamy and Aishwarya Nagarathinam

**Abstract** Artificial intelligence (AI) is showing a paradigm shift in all spheres of the world by mimicking human cognitive behavior. The application of AI in healthcare is noteworthy because of availability of voluminous data and mushrooming analytics techniques. The various applications of AI, especially, machine learning and neural networks are used across different areas in the healthcare industry. Healthcare disruptors are leveraging this opportunity and are innovating in various fields such as drug discovery, robotic surgery, medical imaging, and the like. The authors have discussed the application of AI techniques in a few areas like diagnosis, prediction, personal care, and surgeries. Usage of AI is noteworthy in this COVID-19 pandemic situation too where it assists physicians in resource allocation, predicting death rate, patient tracing, and life expectancy of patients. The other side of the coin is the ethical issues faced while using this technology like data transparency, bias, security, and privacy of data becomes unanswered. This can be handled better if strict policy measures are imposed for safe handling of data and educating the public about how treatment can be improved by using this technology which will tend to build trust factor in near future.

**Keywords** Healthcare · Artificial intelligence · COVID-19 · Neural networks · Ethics

## 1 Introduction: AI—The New Age of Healthcare

Artificial intelligence (AI), machine learning (ML), blockchain, cloud computing, and automation—these are the buzzwords in the new world of twenty-first century. Further, advancement in information technologies such as mobile computing, augmented reality, smart cities, mobile health, wearable devices, and Internet of Things also paves the way implementing novel techniques and practices in each industry. Artificial intelligence (AI) has been applied in several industries like

---

A. Chellasamy · A. Nagarathinam (✉)

School of Business and Management, CHRIST University, Bangalore, India  
e-mail: [aishwarya.n@christuniversity.in](mailto:aishwarya.n@christuniversity.in)

telecommunication, retail, finance, transportation, and much more. The pandemic covid-19 has clearly shown that any industry can take a break from its functioning, but not farming and healthcare. Healthcare has become the worry of most countries because of its crumbling medical infrastructure, aging population, shift in lifestyle choices, and changing patient expectations. Even though innovation is on a tenfold rise in this industry, the demand for their services, raising costs, and meeting patient's complex needs is still a concern. Major structural and transformational changes are happening in this industry, which would create 40 million health sector-related jobs by 2030, says the World Health Organization [1]. This growth predicts that the need for physicians, nurses, and technicians will be on a greater amount worldwide in the future. On taking a note on healthcare automation, artificial intelligence will be greatly helpful to prevail over these shortcomings and in revolutionizing the way healthcare is delivered. In artificial intelligence, all the data generated from the customers are used for detailed personal profiling which is of a great value to behavior prediction and analysis. Health organizations have a large number of datasets in the form of patient records, clinical trials, and research data. AI technologies are very well suited to analyze and process this data and uncover hidden information, recognize patterns, and make predictions. There are a huge number of instances which show that artificial intelligence algorithms are performing on par or in a better way than us in resolving complex problems. AI could have a triple role in the healthcare industry impacting customer experience, business processes, and cost management. In fact, Business Insider Intelligence report stated that spending on AI all across the world is projected to grow at a rate of 48% between 2019 and 2023. Also, medical institutions acquiring AI startups will also rise in the year 2021 leading to a health AI market valuation of \$6.6B [2].

In this covid-19 pandemic, the demand for medical practitioners has enabled on-demand healthcare services using tracking apps and medical search platforms. These services help the patients to connect with their doctors anytime and anywhere, thus reducing the costs and unnecessary exposure to contagious diseases. Several AI-powered bots like GYANT, Florence, Izzy, Buoy Health, Sensely, and Cancer Chatbot are in the market now which can analyze the symptoms of a health condition much like a medical practitioner and educate patients to make better health-related decisions. The healthcare infrastructure has evolved with the assistance of intelligent healthcare techniques and data analytics. The architecture of intelligent healthcare systems can be segregated as three layers, namely data collection, data management, and the service layer [3]. The data collection and the data management layer can be further divided as visualization layer, processing, and analytics layer [4]. The applications of AI are in the fields of early detection and diagnosis, treatment and prediction, and prognosis evaluation [5]. In a study discussing AI relevance in healthcare and its implication on healthcare workers, the authors state that so far no jobs have been eliminated in the healthcare industry by AI; even in jobs like radiologist and pathologist, the penetration of AI is very slow [6]. This research paper presents a healthcare AI framework with various technologies and the relevant data used in each technology.

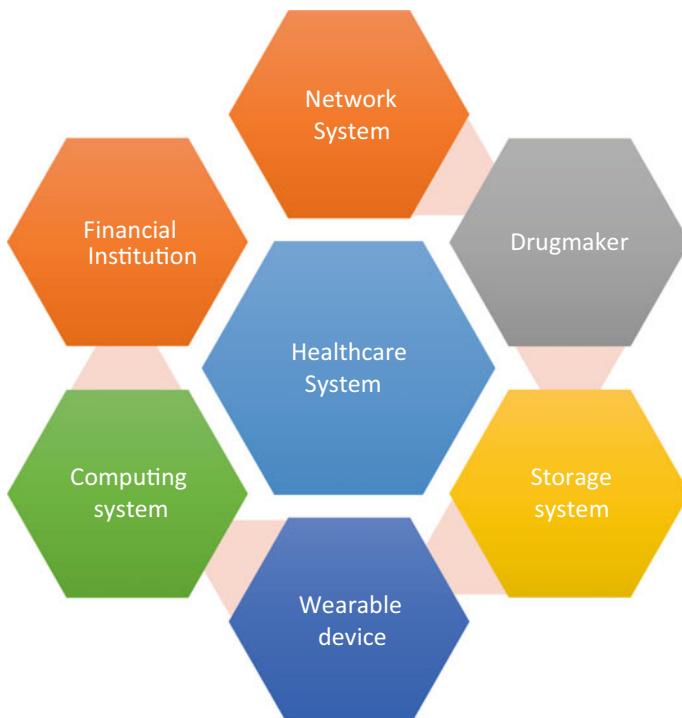
Big data works with datasets that are too complex or too big to analyze by traditional data processing applications. Big data works around four principles, volume, variety, velocity, and veracity [7]. Healthcare industry provides an abundance of data within a multi-dimensional system. With the efficient use of AI algorithms and data analytics, this industry can thrive and develop the ability to tackle large amount of health-related data. Big data in healthcare works with patient, medical, research, physiological, and consumer data. This can be segregated in the form of patients' medical history, clinical data, and medical data. Due to digitization, electronic health records (EHRs) and electronic medical records (EMRs) have become essential across medical institutions [8]. The new healthcare data dynamics uses new technologies such as Hadoop lakes, HIVE, and MapReduce to extract information from their database. The impact of big data has more significance in this industry due to increasing costs, huge amounts of data availability and digitization. Big data and predictive analytics can improve basic to advanced data decisions on patient care systems. It can contribute to precision public health by improving public health surveillance and assessment. Big data can bring new insights into risk factors that lead to disease. By importing data from patient's wearable devices and mobile health applications, real-time data could be studied, and risk could be predicted [9]. New drug development involves lengthy trial and error experiments and needs more time, effort, and money. Big data removes much of the guesswork and allows researchers to arrive at results faster and in a more accurate way. During the Covid-19 pandemic, big data played a major role in the rapid development of Covid-19 vaccines for the public [10]. Novartis Genomics used Hadoop and Apache Spark to design a workflow system for researching next generation sequencing (NGS) by analyzing their multi-dimensional diverse data. Flatiron Health used billions of data points from their cancer patients to enhance their research and get new insights. Pieces technologies based on Dallas use a clinical engine to make recommendations and decisions based on their lab reports and patient's vitals. Health fidelity uses natural language processing to extract information from clinical charts, which helps the practitioners to identify the risk and make appropriate assessment.

## 2 Technological Advancements in Healthcare

Artificial intelligence (AI) refers to a machine that solves complex problems and mimics human capabilities. The growth of big data in AI helps in the fields of medical research, precision medical initiatives, drug development, connected machines, automated image diagnosis, fraud detection, administrative workflow assistance, and insurance. There are numerous companies who are pioneers in the area of artificial intelligence in the healthcare industry including IBM Watson and Google's DeepMind. IBM Watson ranks number one in the International Data Corporation's (IDC) AI market for 2020, and 70% of the global banking institutions uses Watson [11]. They have several applications such as blockchain healthcare, clinical workflows, diagnostic imaging, radiology, clinical decision support, and the like. Google's

DeepMind works in the areas of cancer diagnosis, averting blindness, and predicting patient outcomes. In a quest for a breast cancer solution, the company's algorithm outperformed all human radiologists it competed against on an average of 11.5%. Also, Verily, the life sciences branch of Alphabet, has patented a digital contact lens which could detect blood sugar levels. Some of the health AI disruptors are iCarbonX, Ginger.io, Nuritas, Ovuline, Gauss Surgical, Medalogix, GNS Healthcare, Medal, Zephyr Health, Baylabs, Zebra, Tao and Atomwise to name a few. Figure 1 shows the expanded healthcare system with traditional contemporary roles.

Machine learning (ML) is a subset of AI that learns from the data and builds computational models to make precise decisions. Robotic surgery and medical imaging diagnosis are some applications of ML. Pfizer uses ML for immunotherapy research to understand how the body's immune system can fight cancer [12]. PathAI a Cambridge-based machine learning company helps pathologists to make quicker and accurate diagnosis on treatments and therapies. Quantitative INsights, a Chicago-based AI company, researches breast cancer diagnosis with its computer-assisted breast MRI workstation Quantx. Microsoft's project InnerEye uses machine learning to develop 3D imaging on the location of tumors to assist in radiotherapy and surgical procedure. The most commonly used machine learning algorithms are supervised learning, unsupervised learning, reinforcement learning, and deep learning.



**Fig. 1** Expanded healthcare system [3]

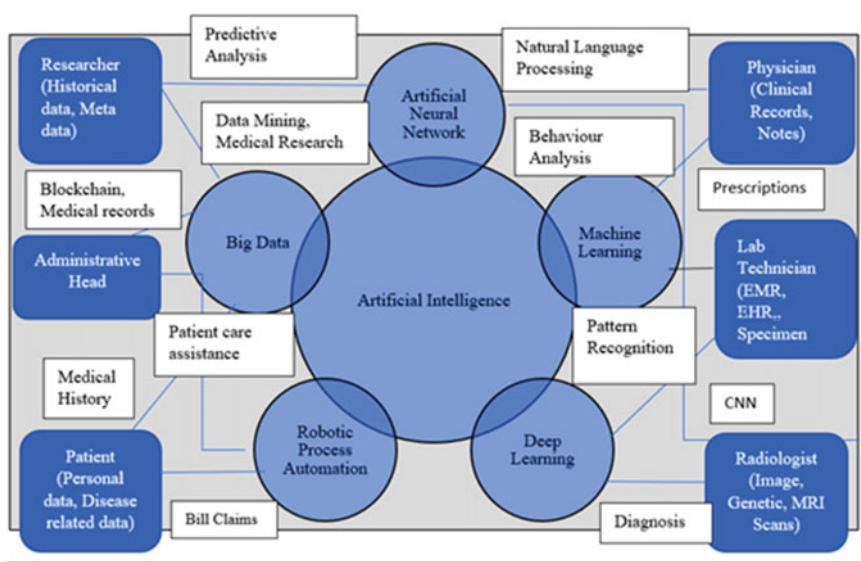
Supervised learning provides data-driven clinical decisions for discrete variables and does predictive analytics. Unsupervised learning is used for exploratory variables such as pattern recognition, anomaly detection, and clustering. However, in real time, both supervised and unsupervised algorithms are often used in combination for leveraging the advantages of both. Deep learning enhances the capacity of supervised and unsupervised learning by adding hidden layers of data through artificial neural networks. Genomics, drug discovery, NLP, Go games are some of the frontiers which had a breakthrough from deep learning algorithms.

Artificial neural networks (ANNs), which is a complex subset of machine learning, work in the same way as the neurons in the human brain operate; hence, computers learn to process raw data from examples with minimal human direction. Artificial neural networks gather knowledge by detecting patterns and relationships in data and ‘learn’ through experience [13]. This technology has been used in various industries for several decades; however, in the healthcare industry, they are used to predict whether a patient may be infected with a particular disease or not. Convolutional neural network (CNN) impacts the healthcare industry in the form of image classification. HealthPNX, a product of Zebra Medical vision, an Israeli medical imaging analytics company has got FDA approval for AI pneumothorax alert which is considered as a successful use case of neural networks. Another type of neural network in the healthcare industry is recurrent neural network (RNN) which is used to build voice recognition applications. Natural language processing is a common technique in RNN. Short-term automation through AI helps medical practitioners to dictate prescription and record meetings which helps them in having quality time with their patients rather than typing on the keyboard. Reference [14] designed deep neural network architectures with CNN and RNN components for Parkinson’s disease for efficient patient-specific analysis and treatment selection.

Robotic process automation (RPA) helps hospitals to streamline their business processes. This technology enables them to automate redundant tasks, save money, time, and improve operational efficiency [15]. RPA robots are capable of mimicking most human interactions, log files, and fill forms thus freeing up clinicians time and effort. Some of the use cases of RPA are appointment scheduling, billing process, claim management, audit management, inventory management, data entry, regulatory compliance, and the like [16]. Max Healthcare, an Insurance company, was able to reduce turnaround time by at least 50 percent by using UiPath Enterprise RPA. The institute adopted RPA for claims processing, data reconciliation for Central Government Healthcare Scheme and Ex-servicemen Contributory Healthcare Scheme [14].

### 3 Healthcare AI Framework

The healthcare industry generates a huge amount of information on a day-to-day basis. The use of advanced technology to stratify and leverage this data to ensure accessibility is the need of the hour. The data available in the form of electronic



**Fig. 2** Intelligent AI healthcare framework. *Source* Author's depiction

health records (EHRs), electronic medical records (EMRs), scans, ECGs, and all lab-related data can be fed to the AI algorithms for training/supervised learning. This equips the AI to create their own ‘logic’ and build predictive models to empower medical personnel. Figure 2 denotes the multifaceted usage of AI in Medicare and its applications. Various personnel involved in the intelligent healthcare system are identified, and the different data they generate are associated with it. It is evident from the framework that for improved health outcomes in the future, AI and machine learning are unavoidable. It is high time that the healthcare industry recognizes each application of AI and applies them with deep understanding at relevant hospital departments.

#### 4 AI Applications in Healthcare

The major objective of a good healthcare system is to predict, personalize, and prevent disease with a participatory approach, and nowadays, AI brings these paradigm shifts with the help of analytics and availability of healthcare data. AI has potential applications in healthcare in broader scope of treatment, diagnosis, surgeries, managing hospital records, clinical data interpretation, and connected healthcare devices. AI has become a game changer in the healthcare industry because of the following reason: First, developed countries are focusing on improved outcomes; second might be the explosion of data availability due to connected IoT devices and up gradation in

software and hardware which make data handling an easy task. Following are the few applications where healthcare professionals profoundly use AI for betterment of patient care.

#### ***4.1 Personal Care Through Chatbots***

Not all patients need a consultation with a physician; AI with the help of virtual assistants can screen patients who just need medical advice and provide information to physicians, and patients can further be assisted through chatbot which is 24/7 available to multiple patients for any query. AI systems can be used as an audit tool to reduce prescription errors and can be used as a pregnancy monitor. There is an AI nurse assists robot by name Moxi which monitors refilling stocks, retrieving patient information from lab results and medical history [17]. Looking at AI's potential use, BCG expects that by 2022 the healthcare industry may spend \$2.1 billion/year on implementing AI tools exclusively on remote prevention [18].

#### ***4.2 Disease Diagnosis and Prevention***

AI helps the health professionals for early diagnosis of diseases, and one such technique is in vitro diagnosis which uses biosensors and chips and another one named as gene expression where ML is implemented and AI identifies a set of data and detects abnormalities [19]. AI also assists medical practitioners in diagnosing one and two dimensional medical imaging in case of prediction, identification, and managing illness. Adding on to this AI can assist in decision support system (DSS) to improve diagnostic accuracy and disease management [20, 21].

#### ***4.3 Administrative, R&D Application***

AI can also be used for extensive application in managing administrative tasks in hospitals for claiming insurance, processing bills, managing hospital records, and patients' documents. ML helps to pair the data from different databases and aids insurers to come to a decision that whether claims are authentic and proceed for further payment process which saves everyone's time and reduces fraud and healthcare data breaches. AI is also capable of finding new drugs based on the history of data and medical intelligence. A company, by name NuMedii, has come with artificial intelligence for drug discovery (AIDD) technology along with big data to rapidly discover the link between disease and drug.

#### **4.4 AI-Assisted Surgeries**

AI along with robots performs surgery with more precision and at a faster pace. As the robots are less prone to fatigue, the prolonged operation can be performed better in a more efficient and accurate way. A robot by name Vicarious Surgical which is embedded with virtual reality and AI performs minimally invasive operations and another miniature mobile robot by name Heartlander developed by Carnegie Mellon University facilitated heart therapy [22]. A Harvard Business Review on Technology by Kallis et al. [23] says that AI-assisted orthopedic surgery has reduced surgical complications, reduced stay of patients at hospital, and reduced error.

#### **4.5 AI-Battling the Coronavirus Pandemic**

The beginning of 2020 has transformed our way of life with unprecedented term COVID-19, colloquially termed as coronavirus. A decade of digital transformation has been quickly witnessed in this short span of one year with accelerated usage of AI in this outbreak of pandemic. Thanks to the massive growth of the Internet and increasing computation of data, AI has shown outstanding performance in detecting, extracting, predicting (BlueDot prediction during first outbreak in Wuhan) the patterns of data and judging the impact of this disease. Deep learning and machine learning technologies were deployed in image recognition, segmentation, time series forecasting [24, 25], and involving robots for operations which has proven success in these social distancing norms. Piccialli et al. [26] have studied AI application in pandemic based on four focus areas as diagnosis, predictions, treatment, and tracing. In treating the patient, there occurs a situation to prioritize resources because there may be shortage of equipment, and in such cases, practitioners will be clueless in choosing the right patient to give these facilities, [27] and also, there occurs confusion in choosing the donor for plasma transfusion. So, to sort this out, Albahri et al. [28] proposed a multi-criteria decision analysis algorithm which prioritizes patients based on their health condition and results obtained from the laboratory. The ML feature has helped Taiwan government for contact tracing by enabling GPS option in infected people's phones so that they are under monitor without further movement. AI algorithms can also diagnose the symptoms of COVID through patient CT scan image, exposure history, lab results, and classify data based on respiratory patients as well [27]. In South Korea, AI along with neural networks is used to collect patients' data which predicts recovery and death rate [29]. This COVID has also led to a lot of misinterpretation of data leading to false communication which may create a panicky situation among the public. Machine learning algorithms could be used to identify trends, sentiment analysis, and deliver information about the root cause of false data and help in preventing rumors and misinformation [30]. Though AI takes forefronts in dealing with this pandemic, there should be a legal framework, approved by the

government and availability of epidemic data and ethical consideration in dealing with those data is in upfront.

## 5 Ethical Considerations

Augmenting AI in healthcare though brings in a lot of advantages, it also has its other side of testing its ethical strand in usage of these technologies in the medical field. Clinicians do have responsibility in educating patients about usage of AI technology and exposure of data thereafter. Usage of health apps involves AI and chatbots where again the user data are stored at back end and consumers are really not aware and not very keen in reading user agreements. The application of these technologies also test the safety and transparency of data handling where AI should ensure the validity of dataset and the algorithm used for data handling should generate refined accurate results [31]. References [32–35] have stated that there may be a situation for human bias inherited by AI leading to gender bias or race bias which is the result of improper dataset training. Another important consideration is accountability, i.e., who takes ownership at the time of failure if any discrepancies occur as this can be described as “problem of many hands” from programmer, end user or owner of data. So, it is not a question about “good/bad;” with respect to technology usage in healthcare, the right direction with legal guidance by building trust among the public may resolve these constraints, and wider usage of AI in society can be expected.

## 6 Conclusion

In this paper, the authors have reviewed the new age of healthcare technology by using artificial intelligence and its major application in healthcare. A framework has been developed to denote the various applications of AI and the medical personnel involved in it. The framework works on the input-process model which discusses the data generated by each medical personnel and the process involved. By and large the applications are many fold but not limited to personal care, diagnosis, prevention, prediction, and surgeries. The growing interest of research in AI application in widespread areas like education, business, banking, autonomous vehicles, social media, and healthcare is increasing to a greater extent. The AI in healthcare gained more attention as it could easily solve many complex problems that are expected to arise in this field, and effective handling of coronavirus is one such evidence. The development of AI provides a wide range of solutions to healthcare, and in turn, healthcare demands more capable AI. This supply and demand ratio will encourage these two sectors to advance in the near future which will improve the quality of life of society.

## References

1. Global Strategy on Human Resources for Health: Workforce 2030. World Health Organization. [https://www.who.int/hrh/resources/pub\\_globstrathrh-2030/en/](https://www.who.int/hrh/resources/pub_globstrathrh-2030/en/) (2016). Accessed on 12 June 2021
2. Business Insider Intelligence Report: AI in Healthcare in 2021: Medical Benefits Examples. Accessed on 2 Jan 2021
3. Ma, X., Wang, Z., Zhou, S., Wen, H., Zhang, Y.: Intelligent healthcare systems assisted by data analytics and mobile computing. *Wirel. Commun. Mobile Comput.* **16** (2018)
4. El Aboudi, N., Benhlima, L.: Big data management for healthcare systems: architecture, requirements, and implementation. *Adv. Bioinf.* (2018). <https://doi.org/10.1155/2018/4059018>
5. Jiang, F., Jiang, Y., Zhi, H. et al.: Artificial intelligence in healthcare: past, present and future. *Stroke Vasc. Neurol.* **1** (2017)
6. Davenport, T., Kalakota, R.: The potential for artificial intelligence in healthcare. *Future Healthc. J.* **6**(2), 94 (2019)
7. Laney, D.: 3D Data Management: Controlling Data Volume, Velocity, and Variety, Application Delivery Strategies. META Group Inc., Stamford (2001)
8. Pastorino, R., De Vito, C., Migliara, G., Glockner, K., Binenbaum, I., Ricciardi, W., Boccia, S.: Benefits and challenges of Big Data in healthcare: an overview of the European initiatives. *Eur. J. Public Health* **29**(3) (2019)
9. Dash, S., Shakyawar, S.K., Sharma, M. et al.: Big data in healthcare: management, analysis and future prospects. *J. Big Data* **6**(54) (2019)
10. Zoro: How Big Data Improves Efficiency, Costs, and Patient Outcomes. Applications and Examples of Big Data in Healthcare (2021).
11. IDC Futurescape Report: Global AI Market in 2021. IDC (2021)
12. Everson, T.: The Promising Role of the Immune System in Cancer. Get Healthy Stay Healthy (2016)
13. Agatonovic-Kustrin, S., Beresford, R.: Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *J. Pharm. Biomed. Anal.* **22**(5), 717 (2000)
14. HT: RPA in Healthcare: Robotics to the Pink of Health, Tech Plus (2020).
15. Kantarci, A.: RPA in Healthcare: Benefits, Use Cases & Case Studies. AI Multiple (2021)
16. Bhatnagar, R., Jain, R.: Robotic process automation in healthcare-a review. *Int. Robot. Autom. J.* **5**, 12 (2019)
17. Spatharou, A., Hieronimus, S., Jenkins, J.: Transforming Healthcare with AI: The Impact on the Workforce and Organizations. McKinsey & Company (2020)
18. Aboshiha, A., Gallagher, R., Gallagher, R., Gargan, L.: Chasing Value as AI Transforms Healthcare. BCG (2019)
19. Rong, G., Mendez, A., Assi, E.B., Zhao, B., Sawan, M.: Artificial intelligence in healthcare: review and prediction case studies. *Engineering* **6**, 291–301 (2020). <https://doi.org/10.1016/j.eng.2019.08.015>
20. Elkin, P.L., Schlegel, D.R., Anderson, M., Komm, J., Ficheur, G., Bisson, L.: Artificial intelligence: Bayesian vs Heuristic method for diagnostic decision support. *Appl. Clin. Inf.* **9**(2), 432–439 (2018)
21. Safdar, S., Zafar, S., Zafar, N., et al.: Machine learning based decision support systems (DSS) for heart disease diagnosis: a review. *Artif. Intell. Rev.* **50**, 597–623 (2018)
22. Rangaiah, M.: Artificial Intelligence in Healthcare: Applications and Threats Analytics Steps (2020)
23. Kalis, B., Collier, M., Fu, R.: 10 Promising AI applications in healthcare. *Harvard Bus. Rev.* 1–5 (2018). Accessed on 10 Jun 2021
24. Shorten, C., Khoshgoftaar, T.M., Furht, B.: Deep learning applications for COVID-19. *J. Big Data* **8**(1), 18 (2021)
25. Nayak, J., Naik, B., Dinesh, P., Vakula, K., Rao, B.K., Ding, W., Pelusi, D.: Intelligent system for COVID-19 prognosis: a state-of-the-art survey. *Appl. Intell.* (2021)

26. Piccialli, F., di Cola, V., Giampaolo, F. et al.: The role of artificial intelligence in fighting the COVID-19 pandemic. *Inf. Syst. Front.* (2021)
27. Mohammad, H., Tayarani, N.: Applications of artificial intelligence in battling against COVID19: a literature review. *Chaos, Solitons Fractals* 142 (2021)
28. Albahri, A., Al-Obaidi, J.R., Zaidan, A., Albahri, O., Hamid, R.A., Zaidan, B.: Multi-biological laboratory examination framework for the prioritisation of patients with COVID-19 based on integrated AHP and Group VIKOR methods. *Int. J. Inf. Technol. Decis. Making* **19**, 1247–1269 (2020)
29. Al-Najjar, H., Al-Rousan, N.: A classifier prediction model to predict the status of Coronavirus COVID-19 patients in South Korea. *Eur. Rev. Med. Pharmacol. Sci.* **24**, 3400–3403 (2020)
30. Khan, R., Shrivastava, P., Kapoor, A., Tiwari, A., Mittal, A.: Social media analysis with AI: sentiment analysis techniques for the analysis of Twitter COVID-19 data. *J. Crit. Rev.* **7**, 2761–2774 (2020)
31. Gerke, S., Minssen, T., Cohen, G.: Ethical and legal challenges of artificial intelligence-driven healthcare. In: *Artificial Intelligence in Healthcare*, pp. 295–336 (2020)
32. Brian, L.: Gender as a variable in natural-language processing: ethical considerations. In: *Proceedings of the First {ACL} Workshop on Ethics in Natural Language Processing*, pp. 1–11 (2017)
33. Koolen, C., van Cranenburgh, A.: These are not the stereotypes you are looking for: bias and fairness in authorial gender attribution. In: *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pp. 12–22 (2017)
34. Vijayakumar, T., Vinothkanna, R., Duraipandian, M.: Fusion based feature extraction analysis of ECG signal interpretation—a systematic approach. *J. Artif. Intell.* **3**(1), 1–16 (2021)
35. Chen, J.I.Z., Hengjinda, P.: Early prediction of coronary artery disease (CAD) by machine learning method-a comparative study. *J. Artif. Intell.* **3**(1), 17–33 (2021)

# Development of an Android Application to Carry Out Tourist Visits in Madrid as a Value-Added Service



Antonio Sarasa-Cabezuelo

**Abstract** Currently, it is possible to access huge amounts of information generated by public and private institutions. This information is available in many cases for free so that it can be processed and exploited to give it a different utility than the original one. This article presents an example of the exploitation of a repository of open data and information from Twitter to create a value-added service aimed at facilitating tourist visits around the city of Madrid. This service has been implemented through an Android mobile application.

**Keywords** Twitter · Open data repositories · Mobile app · Android · Data analysis

## 1 Introduction

In recent decades, there has been the so-called data revolution [8], which refers to the technological possibility of being able to access huge amounts of information from various fields. This in turn has made it possible to exploit the information available to obtain a strategic advantage or obtain an economic benefit. In this context, two initiatives [18] that have made this phenomenon possible have been open data repositories and linked data repositories. Each initiative arises in a context and with a different motivation. Open data repositories [7] have been the mechanism used by public and private institutions to publish information about their activity. Access to information is normally free and is done in many cases using REST-type Web service APIs, through which query requests are made. The result of these requests is returned in some data exchange file in XML, CSV, JSON, or similar format. Thus, using the recovered data, the information can be exploited for different purposes for which the data were initially created. The objective of this type of repositories is to provide the raw information so that third parties can process it to give it a different utility. For their part, linked data repositories [19] arise in the field of the semantic Web

---

A. Sarasa-Cabezuelo (✉)

Facultad de Informática, Universidad Complutense de Madrid, Calle Profesor José García Santesmases 9, 28040 Madrid, Spain  
e-mail: [asarasa@ucm.es](mailto:asarasa@ucm.es)

and consist in the creation of large networks of interlaced knowledge. Access to the information is free and is done through a specific query language called SPARQL. The result of these requests is returned in a data exchange file in XML, CSV, JSON format, or those belonging to the semantic Web scope such as RDF. The objective in this case is the creation of a large interrelated knowledge base on which information searches can be carried out in a more semantic way than those that can be carried out with traditional information search engines.

In any of the cases, from the available data, value-added services can be created. To do this, the data are processed and given a different utility for which they were created, which is valuable for a certain set of people. In the world of mobile applications, you can find numerous examples of apps that implement value-added services [3] from data retrieved from open or linked data repositories [10] such as transport applications or traffic situation, applications of the meteorological state, on stock market data, and others. In all cases, they retrieve the raw information from a repository and create services on that data. A very interesting area to create value-added services is tourism [2] since it has very peculiar characteristics [5]. First, a large number of information sources are available in the form of both public and private open data repositories [16]. Second, these types of services are very popular with people who travel [17]. And thirdly, the service they provide in general is usually very useful for the traveler since they usually cover some type of common need (for example, finding accommodation, finding a flight).

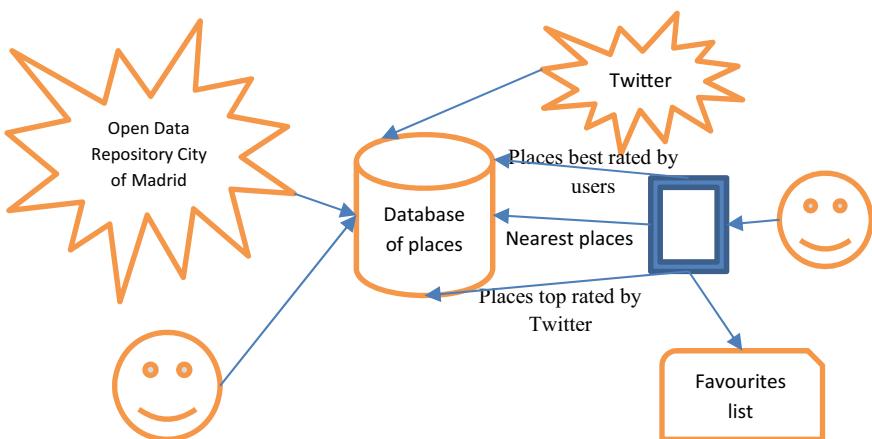
This article presents an Android application that implements a set of value-added services aimed at offering help to tourists with respect to the places they can visit in the city of Madrid. There are numerous applications with similar purposes for many of the great cities of the world, including Madrid, such as [13] 101 trips, Enjoy Madrid, Madrid Tourism, or Madrid City Guide [21]. However, the implemented application presents two important differences with respect to other similar ones. First of all, the searches for places are not carried out in a static way based on keywords. The search is carried out dynamically based on three criteria. On the one hand, it can be searched taking into account the opinions, and evaluations are made by the users of the application of the places they have visited; on the other hand, it can be searched by proximity using the geolocation information provided by the mobile and using the geographical information of the places. And finally, a search can be carried out taking into account the evaluations made by users in the tweets they publish on Twitter [9]. Second, the application presents another difference in its social aspect. In this sense, it is possible to implement a very simple social network that consists of creating lists of friendly users among which recommendations of places that can be visited can be made. Likewise, the application has other differentiating characteristics such as the ability to manage a list of places to visit or the ability to search by voice.

The structure of the paper is as follows: Section 2 describes the functionality of the application as well as its general structure. Section 3 presents the application architecture and a brief description of the database. Section 4 describes how to collect and process data about places, and Sect. 5 describes how to retrieve and process Twitter tweets. Next, Sect. 6 shows part of the implementation made. And finally, Sect. 7 presents the conclusions and a set of lines of future work.

## 2 Purpose and Functionality

The objective of the mobile application is to create a value-added service that helps tourists who visit the city of Madrid to select interesting places to visit. To build it, data from the open data repository of the Madrid city council have been used. This repository has datasets with information on the places that can be visited in the city such as museums, temples, parks, monuments, and others. Furthermore, these datasets can be retrieved for free through a REST-like Web services API. On the other hand, the application can also be enriched by places that users can enter manually. Thus, through these two sources, a large database can be created with places that can be visited in the city of Madrid. On this information, a set of place recommendation services can be built for registered users. First, using the geolocation of the users' mobile, nearby places can be recommended by calculating the distance to each place registered in the database and filtering by a minimum distance. Second, places can be recommended based on the evaluations made by the users themselves of those places they have visited. And finally, a recommendation of places can be made taking into account the evaluations extracted from tweets retrieved from Twitter and that speak about the places stored in the database. Likewise, the application allows each user to select sets of registered users with whom they can exchange recommendations of places to visit, and maintain a list of favorite or pending places to visit. This list is made up of places selected by a user as well as places recommended by other users. In Fig. 1, the application schematic is shown.

To implement the application, three types of users have been defined: unregistered users, registered users, and administrator. Non-registered users are those who do not have an account in the application and who can only invoke information query functions about the places stored in the application. The administrator is the user in charge of maintaining the application so that he can delete users, consult or modify



**Fig. 1** General scheme of the application

the information of any user, and consult/modify/delete information about the places. Finally, the registered user is the one who has an account in the application and can perform operations on the account such as logging in or out, modifying the profile or deleting the account, operations on places such as consulting a place, adding/modifying/delete a place, comment/rate/recommend a place or accept/reject a recommendation, and social operations such as adding/removing friends who use the application to be able to recommend places as well as consult lists of friends. In this way, the functionality of the application has been divided into three modules:

(a) Basic Functions Module

The basic functions of any registered user are the registration of a user in the application, the authentication and exit of the application, the access and modification of the user profile, and the withdrawal of the user account. These functions are shared with the administrator, who also has the possibility of consulting all registered users, deleting user accounts, or modifying the data of any registered user account.

(b) Place Management Module

This module includes all the functions that allow to manage and exploit the information about the places that the application stores. In this sense, a registered user can add new places other than those that the application has automatically collected; it is able to modify or delete the information of a place added by the user himself; it is able to rate or comment on any place managed by the application; it is able to recommend a place to any registered user, and it is able to accept or reject recommendations of places made by any registered user. On the other hand, any user can make three types of inquiries about the places managed by the application: retrieve places near a user, retrieve places ordered by the rating that registered users have made, and retrieve places ordered by the rating they have received in the Twitter tweets. In the first case, to perform the recovery of the places by proximity, the application will use the geolocation of the user and the places so that the distance of the user to the different places will be calculated and those that are at a distance will be displayed less than 5 km (this search distance is configurable). With respect to the search by evaluation of registered users, the application retrieves those places that have been best rated by other registered users and the samples ordered from best to worst score. Finally, the place search functionality using twitter rating works similarly to the user rating search functionality. The only difference is the sort and search criteria. In this case, the criterion is the rating received by the places in the tweets collected automatically by the application.

(iii) Social Module

In this module, all the social network type functions are included. In this sense, a registered user can create a list of registered users friends to whom it possible to recommend places managed by the application. When a user recommends a place to another user, the place appears in a recommendation tab that the application has. Thus, the user can accept the recommendation and add it to a set of recommended places pending visit or can reject the recommendation. Similarly, a registered user can consult the list of friendly users or delete a friendly user from this list.

### 3 Architecture and Data Model

In the architecture of the application, the MVVM [11] pattern has been used, which consists of defining three different parts: model, view, and view model. The view shows the presentation of the data and reacts to changes in the view model. On the other hand, the view model defines the presentation logic and facilitates communication between the view and the model. And the model defines the business logic in charge of obtaining the data through an API or a connection to a database. The interaction between the components is carried out by observers so that the view observes the view model, which observes the model so that if there is any change in it, then the view models will be informed, and each model of the view will inform their respective view. In this way, any change that occurs is transferred to view.

To implement the architecture described, a REST-type Web services API [12] has been defined, and a data model has been implemented using a MySQL-type relational database. Through the Web services implemented in the API, it interacts with the database to perform CRUD operations. For example, 4 parameters are required to invoke the web service that retrieves the list of the places best valued by the application users: the page number, the number of places (by default they are three places), the client user who has requested the response (in case of not being registered, it is anonymous), and the text of the search integrated in the application in case the user uses the text or voice search. As a result, a query is executed ordering the results by the evaluation of the place and stores them in a JSON file that is returned to the Android application (which will process the file to visually display its content). Among the implemented Web services, there is a special one in charge of managing the locations of the users who accept that the application stores the geolocation information of the mobile. To do this, the application itself invokes this Web service automatically every 5 s, storing the coordinates of the user's mobile in the database. Thus, with this geolocation information, it is possible to achieve greater precision when searching for the places closest to the user. On the other hand, a script that runs automatically has also been implemented that eliminates all geolocation information of users that have been stored for more than 1 h.

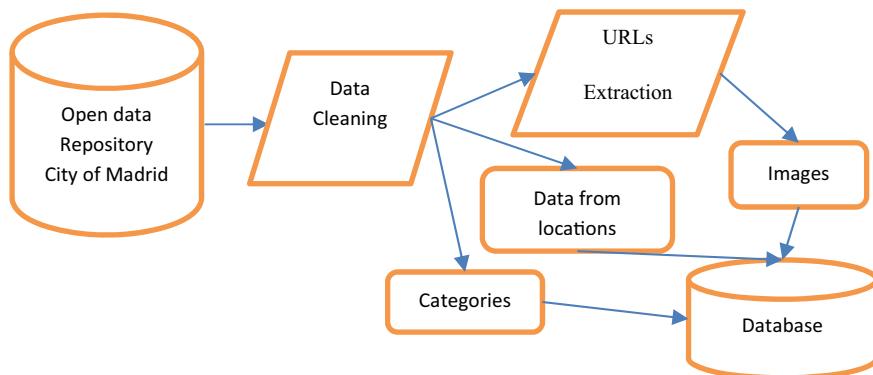
Finally, with respect to the data model, a MySQL-type relational database has been created with 11 tables: the user table that stores information on all registered users, the friends table that stores the information on friendship relationships between users, the tracking table that stores the geolocation information of each user who has given permission to access the mobile data, the places table that stores the information of the places managed by the application, the Location\_images table that stores all the images that it has a place in the form of urls, the Type\_of\_place table that stores the types of places, the comments table that stores the opinions expressed by registered users within the application, the Twitter\_ratings table that stores the comments extracted from tweets about the places, the visited table that stores the places visited or pending to be visited by users (in particular the places recommended by other registered users), the Location\_favorites table that stores the favorite places

of registered users, the recommendations table that stores the recommendations that have been sent between the users of the application.

## 4 Collection of Data on Places

The data of the places that are shown by the mobile application are retrieved from the open data portal of the Madrid City Council [1]. A total of nine datasets are retrieved in XML and CSV formats that contain information about accommodation, museums, parks, restaurants, monuments, shops, temples, clubs, and tourist information about the city of Madrid. In order to be used, the data must be previously processed through the Pandas [14] and ElementTree libraries [6], so that a four-stage processing is performed (see Fig. 2).

In the first stage, the datasets are cleaned: Irrelevant columns are removed; null values are substituted; special characters are replaced, and some transformations are performed on the data. In the data of each place, there is a column that contains a URL to a set of images about the place although it is not always filled. Thus, in the second stage of processing, the URLs of each of the places are accessed, if they exist, and the images of the places are retrieved through the BeautifulSoup [22] Web scraping library. However, for those places for which there is no URL with images, then the photos are extracted from Google Photos by performing a direct search using the Selenium Web scraping tool [4]. In the third stage, the information is retrieved from an information field that appears in all the data that contains a set of words that represent categories to classify the places. The goal is to create the set of all the place classification categories that have been used in the retrieved datasets. Finally, in the fourth stage, all the information that has been retrieved and processed in the previous stages is inserted into the database tables. There is a Web service in the API



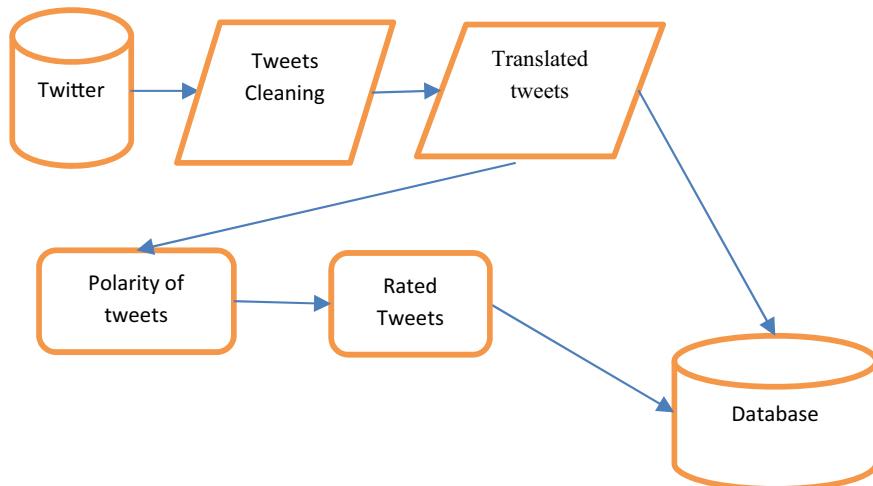
**Fig. 2** Open data portal recovery

that runs every week to perform a new data collection by deleting the existing ones and replacing them with the new ones.

## 5 Collecting Tweets About Places

Another data source that is used to generate the content displayed on the mobile is Twitter. The information contained in the tweets is used to establish the ranking and popularity of the places that can be visited in the city of Madrid. Based on the opinion reflected, each place is scored. To do this, it is necessary to retrieve tweets that describe opinions of the places in the database and then carry out an analysis of the sentiment of the tweets to know if the opinion of the users is negative or positive. Tweets are processed in several stages (see Fig. 3).

In the first stage, sets of tweets containing information from the different places are retrieved using the Twitter API. Then, in the next phase, the text contained in the tweets is cleaned and prepared: Hashtags, retweets, or emojis that the text may contain are eliminated. The goal is to get the clear text of the tweet. The next step would be to assess each text through a sentiment analysis [20]. For this, the Textblob library has been used, which allows natural language processing such as classification, translation, or sentiment analysis. However, in order to use this library, it is necessary that the text to be processed is in English. For this reason, the third phase of this processing consists of performing an automatic translation of the clear text of the tweets using the Google Translate API [15]. In the last phase of this processing, the sentiment analysis of the plaintext in English of the retrieved tweets is carried out. This calculation consists of calculating the polarity of a tweet, which



**Fig. 3** Tweets recovery

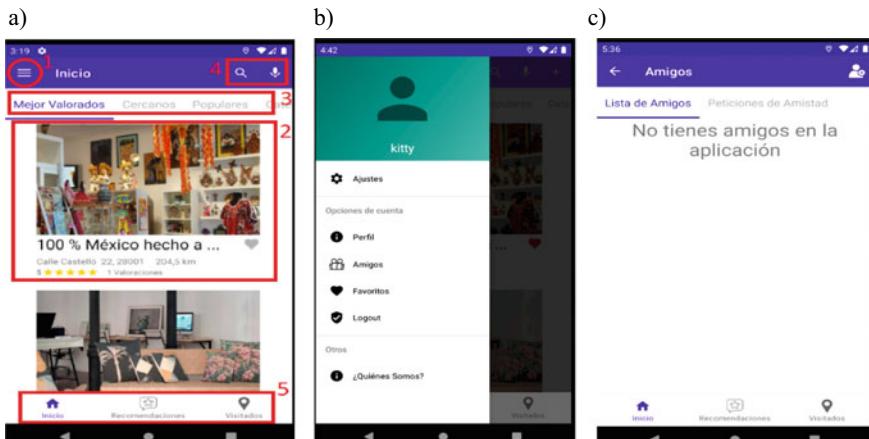
defines how positive or negative a tweet is with respect to a previously preset topic. The polarity result is a number between  $-1$  and  $1$ , with  $1$  being very positive and  $-1$  being very negative. From the polarity obtained, each tweet is scored with values between  $1$  and  $5$  according to the following assignment:

- If polarity between  $-1$  and  $-0.5$ , then value  $1$  is assigned
- If polarity between  $-0.5$  and  $0$ , then value  $2$  is assigned
- If polarity equal to  $0$ , then value  $3$  is assigned
- If polarity between  $0$  and  $0.5$ , then value  $4$  is assigned
- If polarity between  $0.5$  and  $1$ , then value  $5$  is assigned.

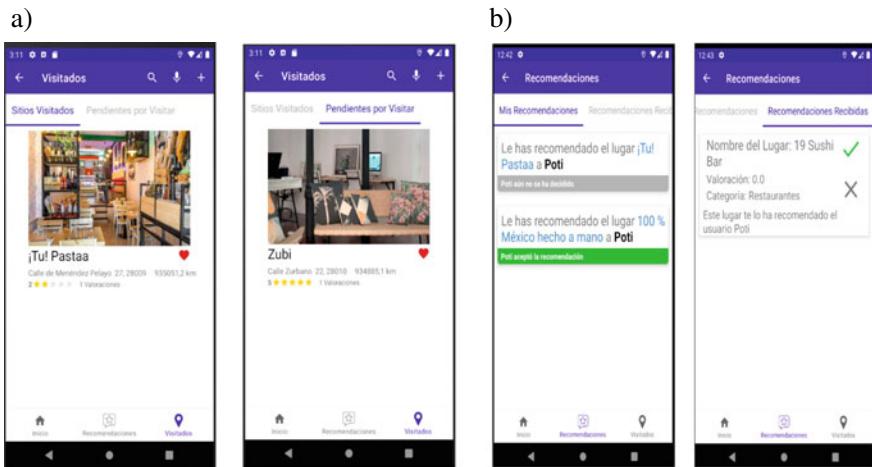
In this way, it is possible to classify the tweets that speak of a place as positive, negative, or neutral. Finally, the clear text of the tweets is stored in the database together with the score obtained. The stored tweets and their score will be used to generate automatic positive, negative, or neutral comments that will be displayed next to the information of each place. Likewise, an average score of the place is generated for the place by taking the average of the scores obtained by the tweets that speak of that place. This average score is used to classify the place with respect to the rest of the places. There is a Web service in the API that runs every week to perform a new data collection by deleting the existing ones and replacing them with the new ones.

## 6 Implementation

Next, some of the implemented functions will be shown. Figure 4a shows the main interface of the application from which the functionalities can be accessed. From zone



**Fig. 4** a Main interface, b options menu, c friends list



**Fig. 5** a Visited screen, b recommendations screen

number 1, a menu with options can be accessed (Fig. 4b): app configuration, user profile, friend management (Fig. 4c), favorites list, exit the application, or information about the application.

In zone 2, the list of places that have been obtained from one of the four possible queries that can be done from the zone numbered 3 is shown: best rated, nearby, popular, or categories. From zone 4, it is able to search for places by text or by voice. And finally, from zone 5, it is able to access the list of recommended places and the list of places visited.

Figure 5a shows the screen of visited places, in which there are two tabs: visited sites and pending sites to visit. And in Fig. 5b, the recommendations screen is shown, which has two tabs: places recommended by the user and places recommended to the user.

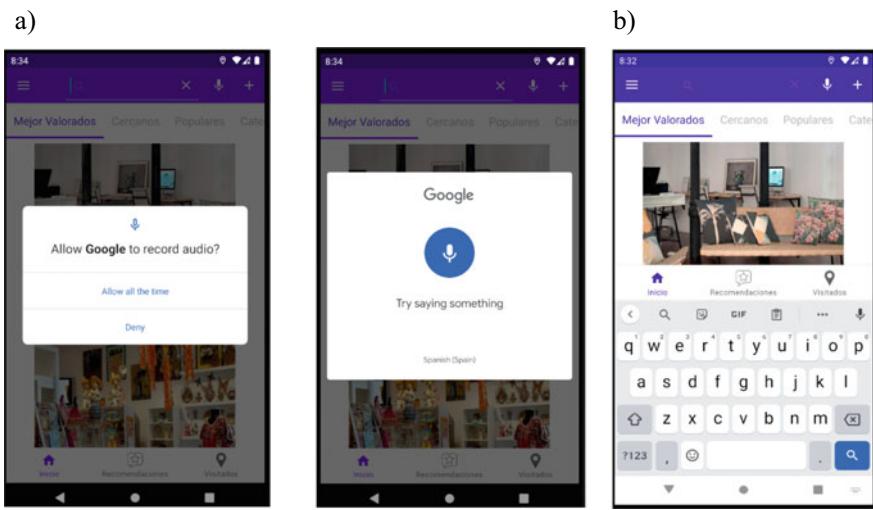
In Fig. 6a, the search for places by voice is shown, and in Fig. 6b, the search by text is shown.

The information accessible when a place is selected is shown below: information on the place (Fig. 7a), route to the place (Fig. 7b), and valuation of the place (Fig. 7c).

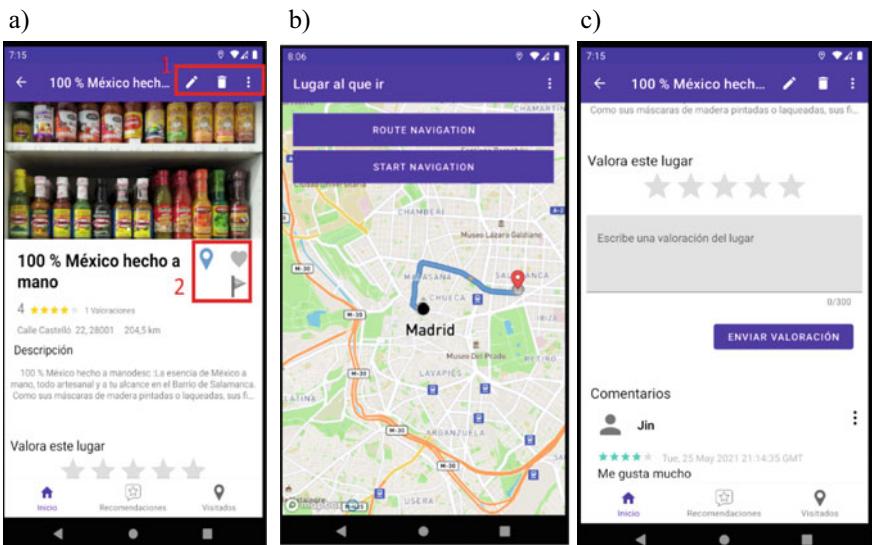
And finally, Fig. 8a shows the screens of the favorites list; Fig. 8b shows the list of popular places, and Fig. 8c shows the search by categories.

## 7 Conclusions and Future Work

In this article, a mobile application has been presented that implements a value-added service aimed at tourists visiting the city of Madrid. Using data from the Madrid city council's open data repository and information obtained from Twitter, a service is provided that allows tourists to recommend places they can visit under four different

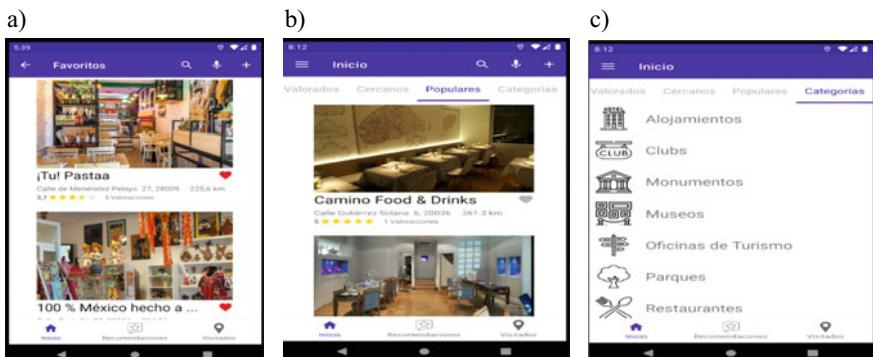


**Fig. 6** **a** Voice search, **b** text search



**Fig. 7** **a** Information of a place, **b** route to a place, **c** assessment of a place

criteria: by proximity, by user assessment, by assessment of Twitter users, and by categories. Likewise, the application allows to manage lists of favorite and pending places to visit, as well as to share places they have visited with other users. The implementation is an example of how it is possible to create value from data that was created for other purposes and thus gives it a new utility.



**Fig. 8** a List of favorite places, b list of popular places, and c list of navigation by categories

However, there are some improvements in the application such as:

- Expand the data sources used.
- Expand the scope of application to other cities.
- Add a map to show nearby places.
- Provide an interface for other languages.
- Expand the information offered for each place.

**Acknowledgements** I would like to thank Juan Antonio Escobar de los Ángeles, Daniel Nasim Santos Ouafki, Jin Tao Peng Zhou, and Jin Wang Xu for implementing the described application.

## References

1. Alejo, H.E.C.D.: Portal de Datos Abiertos del Ayuntamiento de Madrid. Consultor Ayuntamientos Juzgados: Revista Técnica Especializada Administración Local Justicia Municipal **3**, 128–144 (2020)
2. Ammirato, S., Felicetti, A.M., Linzalone, R., Carlucci, D.: Digital business models in cultural tourism. *Int. J. Entrepreneurial Behav. Res.* (2021)
3. Belhadi, A., Djennouri, Y., Lin, J.C.W., Cano, A.: A data-driven approach for Twitter hashtag recommendation. *IEEE Access* **8**, 79182–79191 (2020)
4. Bruns, A., Kornstadt, A., Wichmann, D.: Web application tests with selenium. *IEEE Softw.* **26**(5), 88–91 (2009)
5. Chakrabarty, N.: A regression approach to distribution and trend analysis of quarterly foreign tourist arrivals in India. *J. Soft Comput. Paradigm (JSCP)* **2**(1), 57–82 (2020)
6. Garabík, R.: Processing XML text with Python and ElementTree—a practical experience. *Insight into the Slovak and Czech Corpus Linguistics*, 160 (2006)
7. Goben, A., Sandusky, R.J.: Open data repositories: current risks and opportunities. *Coll. Res. Libr. News* **81**(2), 62 (2020)
8. Hernández-Pérez, T.: En la era de la web de los datos: primero datos abiertos, después datos masivos. *Profesional Información* **25**(4), 517–525 (2016)

9. Karami, A., Lundy, M., Webb, F., Dwivedi, Y.K.: Twitter and research: a systematic literature review through text mining. *IEEE Access* **8**, 67698–67717 (2020)
10. Kontogianni, A., Alepis, E.: Smart tourism: state of the art and literature review for the last six years. *Array* **6**, 100020 (2020)
11. Li, X., Chang, D., Pen, H., Zhang, X., Liu, Y., Yao, Y.: Application of MVVM design pattern in MES. In: 2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), pp. 1374–1378. IEEE (2015)
12. Masse, M.: REST API Design Rulebook: Designing Consistent RESTful Web Service Interfaces. O'Reilly Media, Inc. (2011)
13. Mazeda, M., Teixeira, L.: Mobile applications for cultural tourism. *St. James Way. Rev. Lusófona Estudos Culturais/Lusophone J. Cult. Stud.* **7**(1), 165–184 (2020)
14. Molin, S.: Hands-On Data Analysis with Pandas: Efficiently Perform Data Collection, Wrangling, Analysis, and Visualization Using Python. Packt Publishing Ltd. (2019)
15. Prates, M.O., Avelar, P.H., Lamb, L.C.: Assessing gender bias in machine translation: a case study with google translate. *Neural Comput. Appl.* **32**(10), 6363–6381 (2020)
16. Piedra, N., Chicaiza, J., López, J., Caro, E.T.: A rating system that open-data repositories must satisfy to be considered OER: reusing open data resources in teaching. In: 2017 IEEE Global Engineering Education Conference (EDUCON), pp. 1768–1777. IEEE (2017)
17. Rehman Khan, H.U., Lim, C.K., Ahmed, M.F., Tan, K.L., Bin Mokhtar, M.: Systematic review of contextual suggestion and recommendation systems for sustainable e-tourism. *Sustainability* **13**(15), 8141 (2021)
18. Sarasa-Cabezuelo, A.: Using open data repositories and geolocation to create value-added services for tourism. In: 2020 24th International Conference Information Visualisation (IV), pp. 126–131. IEEE (2020)
19. Sarasa-Cabezuelo, A.: Creation of value-added services by retrieving information from linked and open data portals. In: Advanced Concepts, Methods, and Applications in Semantic Computing, pp. 147–165. IGI Global (2021)
20. Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.F., Pantic, M.: A survey of multimodal sentiment analysis. *Image Vis. Comput.* **65**, 3–14 (2017)
21. Vijesh, J.C., Raj, J.S.: Location-based orientation context dependent recommender system for users. *J. Trends Comput. Sci. Smart Technol. (TCSST)* **3**(1), 14–23 (2021)
22. Zheng, C., He, G., Peng, Z.: A study of web information extraction technology based on beautiful soup. *J. Comput.* **10**(6), 381–387 (2015)

# MFES Framework for Efficient Feature Selection Among Subsystems in Intelligent Building



Abba Babakura, Abubakar Roko, Aminu Bui, Ibrahim Saidu,  
and Mahmud Ahmad Yusuf

**Abstract** The increasing trend of problem representation and high-dimensional data collection calls for the utilization of feature selection in many machines learning tasks and big data representations. However, identifying meaningful features from thousands of related features in the smart home data which are dissimilar in nature remains a nontrivial task. This has prompted for the deployment of a feature selection algorithm (FSA) that provides two possible solutions. First, to provide an efficient scheme that best optimizes the features for subsystem decisions and second, tackles feature subset selection bias problem. In this paper, a MFES framework for feature selection is proposed that uses a hybrid mechanism to tackle the problem of feature subset selection bias in intelligent building data. The mechanism uses the effectiveness of filters and accuracy of wrappers to obtain significant features for prediction. The proposed MFES framework resulted in 92.17% of accuracy as compared to the baseline approach resulting in 87.21% of accuracy. The experimental results show that efficient and better prediction accuracy can be achieved with a smaller feature set.

**Keywords** Feature selection · Machine learning · Simulated annealing algorithm (SAA) · F-score · Info-gain

---

A. Babakura (✉) · A. Roko · A. Bui

Department of Computer Science, Usman Danfodio University, Sokoto, Nigeria  
e-mail: [abba\\_babakura@yahoo.com](mailto:abba_babakura@yahoo.com)

I. Saidu

Department of Information and Communication Technology, Usman Danfodio University, Sokoto, Nigeria

M. A. Yusuf

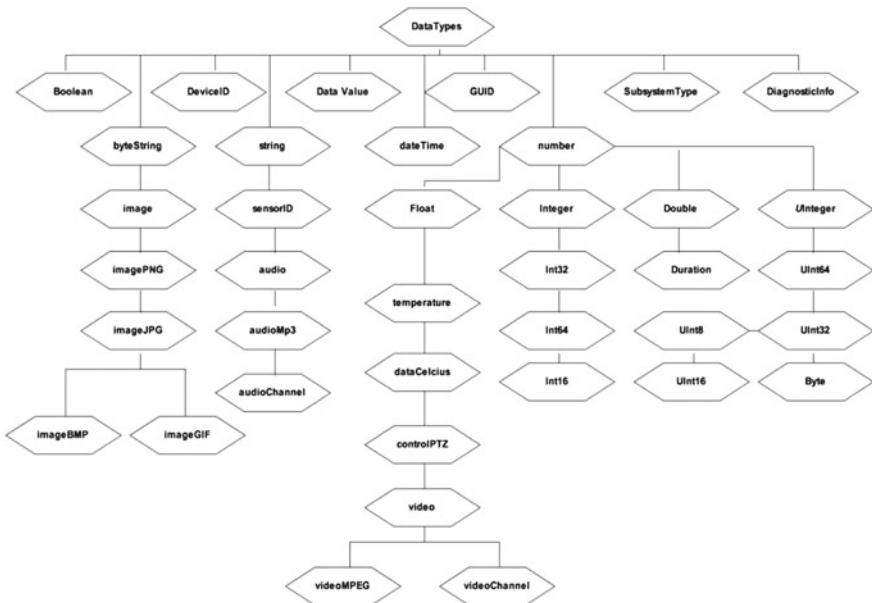
Department of Computer Science, Bayero University Kano, Kano, Nigeria

## 1 Introduction

Feature selection has been a widely utilized technique for the dimensionality reduction, becoming the main point of research area in big data, data mining as well as machine learning [1, 2]. In machine learning, as data dimensionality increases, the necessary amount of data which provides a dependable analysis also grows exponentially. This phenomenon is termed curse of dimensionality in the problem domain of dynamic optimization by Bellman [3]. A common method for problems related to high-dimensional datasets is to forecast the data to smaller number of features that can preserve information as much as possible. Also, there is an increasing difficulty in justifying the results statistically due to sparsity of meaningful data, because of significant growth in dataset dimensionality.

In intelligent building, besides the different communication protocol deployed, the data types utilized are also dissimilar among subsystems. Eventually, such condition would transform the IB into a data-intensive entity as it engrosses exchange of many data types to carry out appropriate interoperation among subsystems. Figure 1 shows the classification of data type's hierarchy for common subsystems interoperation in intelligent building. Based on the hierarchy shown in Fig. 1, many of these data types are disparate and fixed with respect to subsystems protocol and utilized based on their configured application domains.

By default, each subsystem is associated with their respective data types. Such structured data types are powerful and at same time are most complicated.



**Fig. 1** Classification of data type's hierarchy in intelligent building interoperation

The aim of feature selection is to find most relevant features in a problem domain. It significantly improves computational speed and accuracy of prediction. However, identifying meaningful features from thousands of related features in intelligent building dataset which are dissimilar in nature remains a nontrivial task. This has prompted for the deployment of a feature selection algorithm (FSA) that provides two possible solutions: first, to provide an efficient scheme that best optimizes the features for subsystem decisions, and second, to tackle feature subset selection bias problem.

There have been numerous studies proposed to evaluate the feature subset selection to obtain significant features in each data [4–15], and to solve the feature subset selection bias problem [16–18]. To date, none of this FSAs has been able to identify the best combination of significant features that works efficiently as well as solving the challenge caused by the feature subset selection bias.

In this paper, a new framework, multiple feature evaluation system (MFES), is proposed that uses a hybrid mechanism to tackle the problems of feature subset selection and subset bias problem of smart home data [19]. The mechanism uses the effectiveness of filters method with the accuracy of wrappers method. This paper is organized as follows: In Sect. 2, it addresses related work on feature selection algorithms. Section 3 describes the proposed framework and its operation. Section 4 explains performance evaluation and we conclude in Sect. 5.

## 2 Related Works

Several studies were proposed to address the feature subset selection and subset bias problems. In this section, it explains the solutions provided relating to these mentioned problems in machine learning paradigm.

Backstrom and Caruana conducted a work on cascade correlation using an internal wrapper feature selection technique. In this method, the features are selected while the hidden units are added to the growing correlation network architecture [20].

Liu et al. [21] conducted a work on wrapper method that utilizes the SVM model for demand forecasting. Firstly, a genetic algorithm based on wrappers is deployed to analyse the data. Then, the SVM regression model is then built by applying the selected data.

Deisy et al. [22] conducted research on filter method by using the analysis of symmetrical uncertainty with information gain. In their work, they calculated the difference between the features and the entropy of the whole class, which identifies the features that have less information. In addition, it highlighted that some methods of feature selection are based on features' discrimination ability.

Chen and Lin [23] adopted the *f*-score technique for performing feature selection. In this research, the Support Vector Machine (SVM) was utilized for measuring the performance of feature sets. The *f*-score analyses each feature's decimation ability. Inferable from the SVM additionally attempts to find a separation hyper-plane to

divide the different portions of the classes' data apart. The *f*-score helps the model in removing the features that have low decimation ability.

Another important work was conducted on feature selection bias in regression. In their research, they focused on the ability to make inference in the built model with feature selection bias [16, 24]. The highlighted work can make relationships between the output and the selected features stronger. It can negatively affect the prediction performance of the model, and that is because the model overfits the data in the presence of the selection bias and it may not generalize well. As a result of this findings, it has motivated the work. It is well known that classification was adopted for making qualitative predictions, while regression analysis was adopted for making quantitative predictions. However, results obtained from classification cannot be directly generalized to that of the regression.

In the work presented by Jensen and Neville [17], they used the context relational learning for feature selection bias. This is to indicate that some features may have an artificial linkage with the class, which causes their selection because of the relational nature of the data. Moreover, their work has been the first in terms of studying feature subset selection bias in classification. They used both independently and identical samples in parallel.

Surendra and Huan [18] conducted an experiment using classification learning for the feature subset selection bias. In this method, a multi-class synthetic dataset was used for the experiment. Different feature selection techniques like IG, One-R, Chi-Squared, Relief-*f*, are applied to evaluate effects of the feature subset selection bias in classification task. In this experiment, it was observed that, an increase in number of instances frequently decreases the selection bias and, bigger attribute variance usually leads to bigger selection bias. The result shows that selection bias has less effect on classification performance than it does on regression performance.

Wong et al. [25] conducted a work on feature selection methods in machine learning. They proposed a novel immune clonal genetic algorithm to solve the problem of feature selection. In their experiment, they combined immune clonal and genetic algorithm to perform the experiment and predict the result. The algorithm performed better when compared to the genetic algorithm and the Support vector machine. However, the model eliminated the local search step for optimizing the features and, it did not cater for the feature bias problem.

Another work was conducted by Huang et al. [26]. In their work, they proposed a hybrid genetic algorithm for feature selection using the wrapper method based on mutual information. The method uses learning machine as fitness function and searches best subset features in the space of all feature subset in the domain. It uses a heuristic algorithm to improve the local search for feature selection. The results show some improvement in the hybridization of the algorithm. However, it was noted that framework inherits the weakness such as the long run time and computational complexity.

In this research, a Multiple Feature Evaluation System (MFES) which introduces a hybrid feature selection method of filters and wrapper algorithms for finding best feature subset and tackles the feature subset bias problem in classification is proposed.

### 3 Proposed MFES Framework

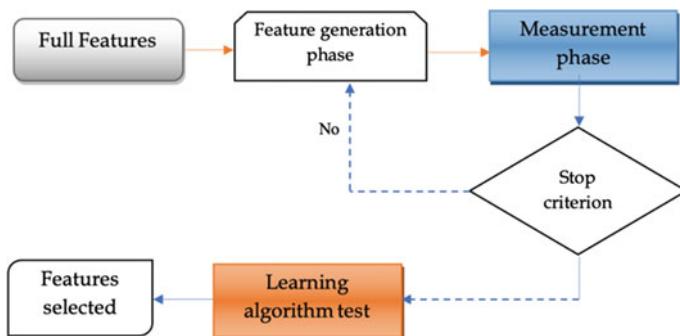
This section explains the design of the MFES. It further explains filter and wrapper methods and discusses the operations of the multiple algorithms used in the implementation of the system. As significant feature selection is crucial in obtaining a relatively strong accuracy result when analysing a large dataset, the combination of these two different methods will improve the process. The filters processes quickly, however, the results obtained are usually not acceptable. In addition, it takes longer processing time in wrapper method, but it results in high accuracy values [27]. The filters technique calculates the information from features; and because of that, the results of the feature selection highly depend on the measured information of the features. On the other hand, the wrapper technique uses the learning algorithm for making judgement; however, the classification result obtained is biased by the learning algorithm. Table 1 depicts the overall properties of the two techniques.

#### 3.1 Filter and Wrapper Methods

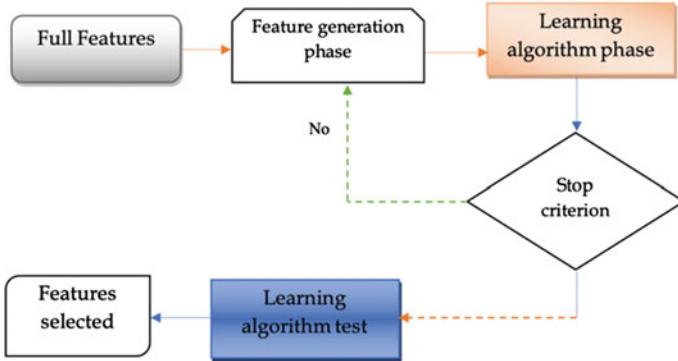
From information hypothesis perspective, the feature set information may well be determined by different statistical measures and because of that, it has been the centre for every filter type of feature selection method. As seen in Fig. 2, there are three main phases of filters which are feature generation phase, the measurement phase,

**Table 1** Properties of the filters and the wrappers

Entities	Filters	Wrappers
Classification accuracy	Depends on	High
Processing speed	Fast	Slow
Dependency on learning methods	No	Yes



**Fig. 2** The filters method (f-score and info-gain)



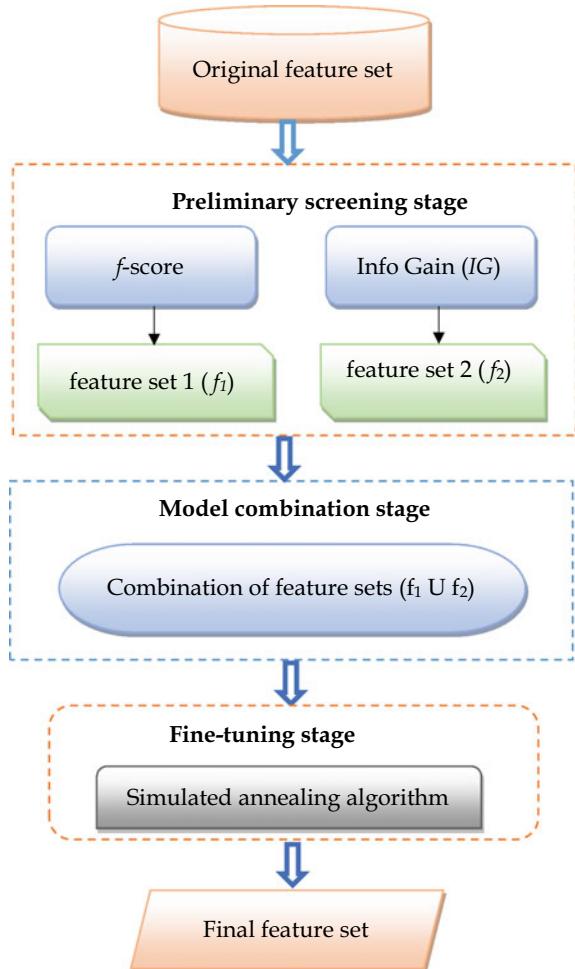
**Fig. 3** The wrappers method (simulated annealing algorithm)

and learning algorithm test phase. At the feature generation phase, a set of features are generated which are labelled as feature subsets. In the measurement phase, the current feature set information is measured. If the obtained results match the stop criterion, then the process is terminated, else, the steps will be performed repeatedly [27]. However, the stopping criterion is the threshold of the measurement results. Finally, at the learning algorithm test phase, an algorithm like Naïve Bayes (NB) or Support Vector Machines (SVMs) is utilized to test the significance of selected features. Hence, the final feature set contains the most significant features.

Figure 3 shows the procedure of the wrapper method. The process is like the filter method except a learning algorithm is used to replace the measurement phase. This has been the reason wrappers perform slower than the filters. However, due to the learning algorithm introduced at the second phase, the wrappers produce a more promising features selection results and solve the problem of feature subset selection bias. As the stopping criterion, when the number of features reaches some predefined threshold, the process stops automatically.

### 3.2 MFES (*Hybrid*) Framework

The overall MFES framework is illustrated in Fig. 4. It consists of mainly three stages: the preliminary screening stage where two filter methods (Information Gain and f-score) are chosen to remove/reduce most irrelevant or redundant features. These two resulted features are then combined at the second stage called Model combination stage. At the fine-tuning stage, a wrapper model with an experience of deep local search ability is then applied to eliminate the feature selection bias and improve the overall classification accuracy. The three stages are described in detail in the subsections.

**Fig. 4** MFES framework

### 3.2.1 Preliminary Screening Stage

At this stage, both techniques which are the f-score, and IG were selected to reduce or remove irrelevant and redundant features of the IB dataset. F-score has been recognized as a novel model widely used in filters to calculate the discriminative capability of each feature, which means that in the classification problems, the features that have higher f-score have better separation capabilities. It can be illustrated as:

$$F(i) \equiv \frac{\left( \bar{x}_i^{(+)} - \bar{x}_i \right)^2 + \left( \bar{x}_i^{(-)} - \bar{x}_i \right)^2}{\frac{1}{n_+ - 1} \sum_{k=1}^{n_+} \left( x_{k,i}^{(+)} - \bar{x}_i^{(+)} \right)^2 + \frac{1}{n_- - 1} \sum_{k=1}^{n_-} \left( x_{k,i}^{(-)} - \bar{x}_i^{(-)} \right)^2} \quad (1)$$

where  $\bar{x}_i^{(+)}, \bar{x}_i^{(-)}$  stands for the positive and negative averages of  $i$ th and  $x_i$  is the average of  $i$ th feature of the dataset;  $n_+$  and  $n_-$  signifies the positive and negative number of instances respectively; and  $x_{k,i}^{(+)}, x_{k,i}^{(-)}$  stands for  $i$ th feature of both  $k$ th positive and negative instances [26].

Consider Eq. 1, the larger the  $F(i)$  value is, the stronger the discriminative capability of the feature becomes. Since the f-score only examines the discriminative capability of only individual feature, it is not capable of identifying the multiple features. However, the features which score low are disregarded even if they can be complementary to the top features. Therefore, information gain (IG) which is another type of filter method was deployed which has the capacity to choose candidate features with related information.

$$\text{Entropy}(N) = \sum_{i=1}^k P_i \log_k \left( \frac{1}{P_i} \right) = - \sum_{i=1}^k P_i \log_k P_i \quad (2)$$

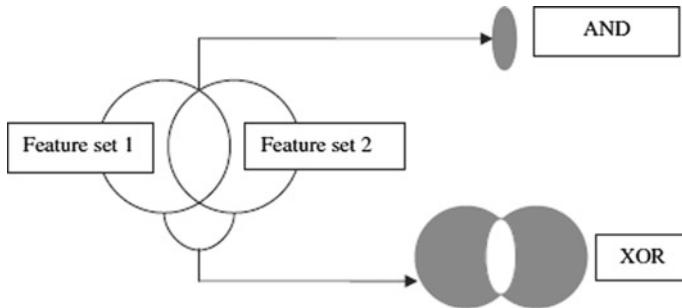
$$\text{Entropy}(D_j) = \sum_{i=1}^{|D_j|} \frac{D_{ji}}{N} \times \text{Entropy}(D_{ji}) \quad (3)$$

$$\text{IG}(D)_j = \text{Entropy}(N) - \text{Entropy}(D_j) \quad (4)$$

IG is concerned about the amount of information that can be provided by each feature. Equations 2–4 are used to determine the IG values. In Eq. 2,  $P_i$  stands for probability of class  $i$  that appears in the  $N$  points of data, and it calculates all the classes information.  $D_{ji}$  is that the  $j$ th feature contains  $i$  kinds of different values as described in Eq. 3. Finally, the IG if the  $j$ th feature is calculated by finding the difference of Eqs. 2 and 3.

### 3.2.2 Model Combination Stage

From the completion of the preliminary screening stage, feature sets F1 and F2 are selected when the two algorithms are run. The selected features obtained from the algorithms are classified as most class-related features [26]. Considering the obtained features as the final set of features can be more harmful to the classification task and thus, it is not a wise decision. It will not only affect the computational time, but the classification accuracy will also be affected significantly and the key idea in this stage is how the two feature subsets obtained can be effectively combined. The combination of the two features sets, i.e. the union ( $U$ ) of the IG and f-score, are separated into two distinct parts: exclusive-OR (XOR) and intersection (AND). Figure 5 illustrates the two parts of the selected feature set one (1) and two (2) respectively. From the full set, the intersection part of the feature sets (F1 and F2) is recommended by both algorithms (IG and f-score), and it can be conserved in our final feature set. However, for the exclusive-OR part of the feature sets (F1 and F2), some of the features might be included because they are valuable.



**Fig. 5** The intersection and exclusive-OR sets

Consequently, a fine-tuning stage which uses a wrapper method was implemented to further justify the significance of the selected features in both the intersection and exclusive-OR parts. However, the worst result coming from fine-tuning in respect to feature reduction is the union of feature set 1 and feature set 2.

### 3.2.3 Fine-Tuning Stage (Simulated Annealing Algorithm)

At this last stage, a wrapper type of feature selection algorithm called simulated annealing algorithm (SAA) is used, which has a strong local search capability of selecting a feature set and can eliminate the feature bias problem while improving the classification accuracy. Since, the initialization of big data research in organizations, significant feature selection has become necessary for analysing meaningful information. Thus, simulated annealing algorithm with the capability of deep searching has become a focal point for selection of significant features for big data analysis.

As discussed earlier, the wrapper method is not suitable when applied to wide range of features. Due to the reduction of feature performed at the preliminary screening stage, the wrappers can now be utilized with less computational effort. The simulated annealing solves the optimization problem by randomly manipulating the solution and then, increases the ratio of greedy improvement slowly until it reaches a point where no further improvements are found [28, 29].

To implement the simulated annealing technique, three parameters must be specified. First, the annealing schedule, consisting of the initial and final temperature,  $T_0$  and  $T_{\text{final}}$ , and the annealing constant  $\Delta T$ . This annealing schedule together governs how the search will proceed over time and when the search will stop. The second is the function, used for the evaluation of potential solutions (feature subsets). In this work, we assumed that higher evaluation scores are more significant. The neighbour function is the final parameter, it takes current solution and temperature as the initial input and returns new “nearby” solution. The temperature governs the size of the neighbourhood. At high temperature, the neighbourhood becomes large and allows the algorithm to explore more and at low temperature, the neighbourhood becomes small, thereby forcing the algorithm to explore locally. Algorithm

1 describes the procedure for the simulated annealing for feature subset selection. The final features are obtained from this algorithm which is passed to any learning algorithm for evaluating the performance.

**Algorithm 1:** The Simulated annealing algorithm for feature subset selection

**Given:**

Examples  $\mathbf{X} = \langle x_1, y_1 \rangle, \dots, \langle x_m, y_m \rangle$   
Annealing schedule,  $T_0$ ,  $T_{final}$ , and  $\Delta T$  with  $0 < \Delta T < 1$   
Feature set evaluation function  $Eval(\cdot, \cdot)$   
Feature set neighbor function  $Neighbor(\cdot, \cdot)$

**Algorithm:**

```

 $S_{best} \leftarrow$  random feature subset
while  $T_i > T_{final}$  do
     $S_i \leftarrow Neighbor(S_{best}, T_i)$ 
     $\Delta E \leftarrow Eval(S_{best}, \mathbf{X}) - Eval(S_i, \mathbf{X})$ 
    if  $\Delta E < 0$  then //if the new feature subset returns better
         $S_{best} \leftarrow S_i$ 
    else //if the new feature subset returns worse
         $S_{best} \leftarrow S_i$  with probability  $\exp(\frac{-\Delta E}{T_i})$ 
     $T_{i+1} \leftarrow \Delta T \times T_i$ 
return ( $S_{best}$ )

```

Algorithm 1 optimizes the feature subset by iteratively improving the initial randomly generated solution. For every iteration, it generates a neighbouring solution and computes the difference between the candidate solutions and the current in terms of quality. It retains the new solution if it is better. Otherwise, it retains the new solution with a probability that is dependent on the quality difference,  $\Delta E$ , and the temperature. The temperature is then reduced for the next iteration.

### 3.3 Training Model and Dataset

#### 3.3.1 Training Model

In the fine-tuning stage, a more elaborate and critical machine learning algorithm was adopted to select the best set of features for classification. However, the significance of the selected feature set can only be tested using a classification algorithm in terms of some performance metrics. In our research paper, the multilayer perceptron (MLP) algorithm [30] which is a type of feedforward artificial neural network (ANN) to test the performance of the hybrid framework. The model is utilized because it

**Table 2** Intelligent building dataset

Date	Time	Device/Action	Status	Device/Sensor ID
2014-02-27	12:43:27	Door	On/Off	5
2014-02-27	12:43:27	Energy management (EM)	On/Off	17
2014-02-27	12:43:27	Fire alarm	On/Off	6
2014-02-27	12:43:27	Public address	On/Off	24
2014-02-27	12:43:27	CCTV	On/Off	7

adopts the property of a supervised learning technique for its training which is called backpropagation [31] and has the capability of distinguishing data that are not linearly separable.

### 3.3.2 Intelligent Building Datasets

We utilized the IB data [19] for testing the MFES framework. The dataset is however collected in seconds, and for each day in the following format 00:00:00 to 23:59:59. It has several types of sensors used. In general, the database consists of about 28,000 attributes observed in the building environment. To our work, to emphasize the performance of the framework, we focused on five sensors and measure the accuracy. Table 2 shows the dataset type and the sequential arrangement.

### 3.3.3 Experimental Setup

The experimental setup for the MFES framework was done with the following tools and environment.

- Windows 7 32-bit operating system that runs on an Intel machine of Corei7-3610QM.
- System specification of 20 GB Hard disc space, 4 GB RAM and 2.30 GHz processor.
- Weka 3.8.0 tool and Notepad++ are deployed to run the feature selection algorithms.

## 4 Performance Evaluations

This section presents performance metric adopted to evaluate the proposed system. In addition, results and discussions are also presented.

## 4.1 Performance Metric

**Accuracy:** The Accuracy of prediction is used to characterize the classification algorithm's performance, derived as follows:

$$X = \frac{Y}{N} * 100$$

where,  $Y$  stands for number of correct prediction and then  $N$  stands for total number of samples.

## 4.2 Results and Discussions

In the first step of the result, we used the two filter algorithms outlined in the preliminary screening stage to reduce the number of features. Table 3 depicts the results obtained from the preliminary screening process. In this method, a greedy process is used to resolve the threshold setting. The resulting accuracies of the two feature sets obtained from the two algorithms (f-score and IG) when run on MLP are 85.17% and 86.33% respectively. Also, the accuracy of the original feature set when run on the MLP resulted in 78.88%. Furthermore, the number of original features was however reduced to 18,554 from 28,668 for f-score and 19,004 for IG respectively. Nevertheless, with the reduction of features, the wrapper algorithm can perform much faster with little complexity.

**Table 3** Preliminary screening procedure results

Algorithm	Threshold setting	No. of removed features	Retained (final) features	Performance metric (accuracy @tenfold cross-validation) (%)
Unused	–	–	28,668	78.88
F-score	0.001	10,114	18,554	85.17
IG	0.01	9664	19,004	86.33

**Table 4** Combination of model results

Association (relationship)	No. of features	Performance metric (accuracy @tenfold cross-validation) (%)
The full feature set	28,668	78.88
Feature (f-score U IG)	15,790	85.1
Feature (f-score $\cap$ IG)	10,528	86.33
Feature (f-score XOR IG)	5262	–

**Table 5** Comparison of proposed method with existing method

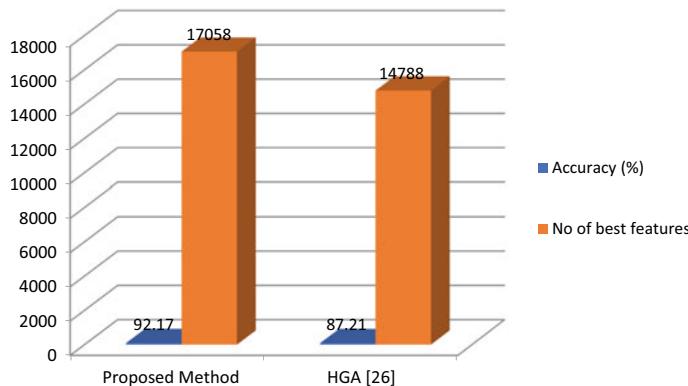
Method	Accuracy (%)	No of best features
HGA [26]	87.21	14,788
Proposed Method	92.17	17,058

Table 4 describes the results obtained from the possible combinations of feature sets as discussed in Sect. 3.2.2. We have the sum of 15,790 features collected from the union set (f-score U IG), 10,528 features obtained from the intersection set (f-score  $\cap$  IG), and they are used in the fine-tuning stage (for running simulated annealing algorithm). The process determines which of the 5262 features obtained from the exclusive-OR set (f-score XOR IG) should be included. The accuracy results of both algorithms on the different parts (f-score U IG) and (f-score  $\cap$  IG) are depicted. After taking the simulated annealing algorithm (SAA), a total of 6530 features are obtained and added successfully to the starting features obtained in the intersection part, resulting in a set of 17,058 features in total and an accuracy of 92.17%. From the experiment, most features obtained from the preliminary screening stage are kept (f-score U IG). This means the original feature set of 28,668 is already very compact for this problem and most of the irrelevant features are removed by the preliminary screening stage. Furthermore, as obtained from the result, the accuracy performance has improved when the test is performed on the simulated annealing algorithm and the percentage (%) and number of features have reduced by about 45–48% (28,668  $\rightarrow$  17,058). Consequently, that will certainly accelerate the subsequent process in this problem domain.

Finally, to further elaborate on the proposed method, an existing method [26] obtained from literature is used to compare the results. To test, same dataset was deployed for testing the performance of the framework. Table 5 compares the existing feature selection method with our proposed method. The result shows that our proposed method outperformed the existing method when applied to this domain (Fig. 6).

## 5 Conclusion

A MFES framework was developed and tested in this research work. The idea is to find the most significant feature subset for the classification task. A 3-stage procedure was designed to have the best feature set in this problem domain. The preliminary screening is used to remove irrelevant and redundant feature. Feature bias is tackled at the fine-tuning stage and further examines the combined feature set. The results obtained showed that the hybridization of filters and wrappers yields to a better result in this problem domain.



**Fig. 6** Comparison of features and accuracy results

## References

1. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157–1182 (2003)
2. Liu, H., Yu, L.: Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans. Knowl. Data Eng.* **17**(4), 491–502 (2005)
3. Bellman, R.E., Dreyfus, S.E.: *Applied Dynamic Programming*. Princeton University Press (2015)
4. Glass, H., Cooper, L.: Sequential search: A method for solving constrained optimization problems. *J. ACM (JACM)* **12**(1), 71–82 (1965)
5. Hall, M.A.: Correlation-Based Feature Selection for Machine Learning (1999)
6. Guyon, I., et al.: Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**(1–3), 389–422 (2002)
7. Ooi, C., Tan, P.: Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics* **19**(1), 37–44 (2003)
8. Hruschka, E.R., et al.: Feature selection by Bayesian networks. In: Conference of the Canadian Society for Computational Studies of Intelligence. Springer (2004)
9. Jiang, H., et al.: Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics* **5**(1), 81 (2004)
10. Jirapech-Umpai, T., Aitken, S.: Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes. *BMC Bioinformatics* **6**(1), 148 (2005)
11. Díaz-Uriarte, R., De Andres, S.A.: Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7**(1), 3 (2006)
12. Jafari, P., Azuaje, F.: An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. *BMC Med. Inform. Decis. Mak.* **6**(1), 27 (2006)
13. Ma, S., et al.: Supervised group Lasso with applications to microarray data analysis. *BMC Bioinformatics* **8**(1), 60 (2007)
14. Rau, A., et al.: An empirical Bayesian method for estimating biological networks from temporal microarray data. *Stat. Appl. Genet. Mol. Biol.* **9**(1) (2010)
15. Yang, P., et al.: A multi-filter enhanced genetic ensemble system for gene selection and sample classification of microarray data. *BMC Bioinformatics* **11**(1), S5 (2010)
16. Zhang, P.: Inference after variable selection in linear regression models. *Biometrika* **79**(4), 741–746 (1992)

17. Jensen, D., Neville, J.: Linkage and autocorrelation cause feature selection bias in relational learning. ICML (2002)
18. Singhi, S.K., Liu, H.: Feature subset selection bias for classification learning. In: Proceedings of the 23rd International Conference on Machine Learning (2006)
19. Alemdar, H., et al.: ARAS human activity datasets in multiple homes with multiple residents. In: 2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops. IEEE (2013)
20. Backstrom, L., Caruana, R.: C2FS: an algorithm for feature selection in cascade neural networks. In: The 2006 IEEE International Joint Conference on Neural Network Proceedings. IEEE (2006)
21. Liu, Y., et al.: Wrapper feature selection optimized SVM model for demand forecasting. In: 2008 The 9th International Conference for Young Computer Scientists. IEEE (2008)
22. Deisy, C., et al.: Efficient dimensionality reduction approaches for feature selection. In: International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007). IEEE (2007)
23. Chen, Y.-W., Lin, C.J.: Combining SVMs with various feature selection strategies. In: Feature Extraction, 315–324. Springer (2006)
24. Chatfield, C.: Model uncertainty, data mining and statistical inference. *J. R. Stat. Soc. A. Stat. Soc.* **158**(3), 419–444 (1995)
25. Wong, S.J. et al.: Evaluating the system intelligence of the intelligent building systems: Part 2: construction and validation of analytical models. *Autom. Constr.* **17**(3), 303–321 (2008)
26. Huang, J., et al.: A hybrid genetic algorithm for feature selection wrapper based on mutual information. *Pattern Recogn. Lett.* **28**(13), 1825–1844 (2007)
27. Kirkpatrick, S., et al.: Optimization by simulated annealing. *Science* **220**(4598), 671–680 (1983)
28. Černý, V.: Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm. *J. Optim. Theory Appl.* **45**(1), 41–51 (1985)
29. Hastie, T., et al.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media (2009)
30. George, F.H.: Models of Thinking. Psychology Press (2015)
31. Cybenko, G.: Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst.* **2**(4), 303–314 (1989)

# QOS Management Protocol for Mobile Ad Hoc Networks Using Mobile Agents



Mallikarjun B. C. and H. S. Phalanetra

**Abstract** Mobile ad hoc networks (MANETs) are self-organized networks of dynamic nodes connected by multihop paths without central administrator and have been incorporated with many applications and routing protocols. Moreover, in a real MANET environment it is very difficult to distribute the network resources for any type of running applications on respective nodes and to maintain firmness of those resources with all other nodes. The various issues such as resource scarcity, node mobility, and quality of service (QoS) for applications running with different requirements over MANETs need to be addressed. The QoS guarantee for applications in MANET is crucial in providing services without interruption to the network users. In recent trends, negotiation and renegotiation of resources in such networks solve the problems associated with allocation of available resources to meet the QoS requirements. In our proposed protocol, QoS negotiation and renegotiation techniques (algorithms) using mobile agents have been incorporated for QoS management in a MANET. The protocol simulation is carried out with JADE framework which is a platform for mobile agents creation, migration, and their interaction with other agents. MANETs with different densities of nodes have been tested with the designed protocol with varying number of applications. We evaluated our proposed work with existing techniques, to show the improvement in terms of response time.

**Keywords** Mobile ad hoc networks (MANETS) · Quality of service (QoS) · Negotiation · Renegotiation

---

Siddaganga Institute of Technology, Tumkur.

---

Mallikarjun B. C. · H. S. Phalanetra

Department of E&TE, Siddaganga Institute of Technology, Tumkur, Karnataka, India  
e-mail: [mallikarjun\\_bc@sit.ac.in](mailto:mallikarjun_bc@sit.ac.in)

## 1 Introduction

The MANET is a decentralized autonomous network which has the mobile nodes connected by wireless links. It is infrastructureless network, which does not include a node for the administration of the network. The QoS management in a MANET refers to the QoS resource provisioning for the smooth running of the applications of a MANET. For the above purpose, the important QoS parameters like bandwidth, delay bounds, and security need to be set for application, in order to provide the information to the user uninterrupted. Traditional schemes with the results of QoS in Internet and wireless networks do not support for MANETs, due to their characteristics such as limited availability of resources, highly dynamic nodes, and lack of central control. These features add lots of difficulty in QoS provisioning. In MANETs, the QoS support is an open issue to be addressed [1]. With regard to this, several QoS management architectures were built for efficient QoS provision based on two aspects, i.e., resource reservation-based and traffic flow-based. The various works address the limitations to provide the QoS guarantee to the applications, such as in [2] the discussion of limitations of current technologies and issues is made in providing the QoS to VOIP applications. In [3], the author finds new needs of QoS in wireless sensor networks from various applications which are identified using models of data delivery and propose end-to-end QoS parameters. In [4], the seamless QoS support architecture which is modular gives the QoS support seamlessly over various wireless technologies used for accessing the network is discussed.

Quality of service plays an important role in improved communication in MANETs along with security attacks, and hence in [5] the QoS and security codesign in MANETs ensure QoS with security from attacks. The trust-based scheme is discussed in [6], which is basically intended for node's misbehavior and isolate them and offer required QoS for routing. The work in [7] discusses mainly on the bio-inspired techniques/algorithms to have the intelligent optimized solutions to achieve the required QoS in the wireless ad hoc network such as energy utilization, increasing the longevity, service quality. In [8], the concept of artificial intelligence techniques is incorporated to bring the required quality of service by obtaining the required throughput, the speed of data through MIMO antennas in mobile networks. The authors in [9] discussed an architecture in which bandwidth, delay, and jitter needs of the applications are met for MANET.

Many QoS frameworks, lack in addressing the issues associated with communication overhead, and response time for the communication. The QoS guarantee for applications running in MANET to provide the required service to the users of MANET uninterrupted is the major issue to be addressed. As the network is dynamic, mobile agent technology can be considered as the solution for addressing associated issues in order to fulfill the application QoS needs of the network. In [10], the QoS scheme for routing on demand has been discussed in finding various QoS paths and choosing the best path in the list for resource preservation. The author in [11] discussed the optimal routing protocol which ensures the optimal route and nodes for transferring the information with associated performance metrics such as storage space, residual energy, distance, trust. In [12], an ant colony optimization technique

(ACO) has been incorporated to choose the relay bus effectively to offer the required end-to-end delay as the quality of service in VANET. In [13], author discusses the estimation of reliable route for the task using mobile agent systems (MAS) depending on the wireless network condition. MAS includes a group of mobile agents for communication. The QoS requirements can be met with the help of mobile agents. The architecture of DAIDALOS II focused on MANET integration with infrastructure networks, use of scheme EDCA of IEEE 802.11 to give differentiated service. The demand of the user for lower delay and high bandwidth is increasing, and hence, there is a strong requirement of preferring the high priority traffic over low priority traffic [14], for guaranteed service.

The resources are diversified due to the fact that MANET is a distributed system, and the negotiation and management of those resources to meet the needs of QoS of applications running in the network is difficult, and hence, an agent-based technique might be the solution.

The authors in [15] proposed a QoS platform for monitoring and controlling the QoS processing tasks using the longest critical path method at a selected best node as a forwarding node among neighbors. In [16], the author discusses the concept of QoS monitoring during the routing process using mobile agents in MANET. In this paper, we propose the QoS management protocol, which uses static as well as mobile agents, for mobile ad hoc networks.

## ***1.1 Necessity for QoS Support in MANET***

The QoS is necessary for all the networks. In MANET, the QoS support is very much essential as the applications have dynamically varying resource requirements for their smooth operation. In the present-day scenario, the more number of services are consuming more resources leaving behind scarcity of resources. There must be a complex algorithm to consider the traffic with different priorities, like high priority to video streams or a real-time video and low priority to background load or best-effort traffic, for assuring the best service guarantees. Hence, QoS that supports with differentiation among different applications is essential, which offers the services to the user accordingly.

## ***1.2 Applications of Mobile Agents for QoS Management***

A mobile agent (MA) is a code, which performs the associated task and is able to autonomously migrate from one node to another. The mobile agents can also be defined as objects which have got certain behavior, state, and location information. In a broad sense, these agent's tasks are application-defined and vary from online banking, shopping to control of the device.

MAS have some important feature or properties that differentiate from the standard code. The mandatory properties associated with mobile agents are being autonomous, being decision making, mobile, communicative, mutual, and learning capability.

- Autonomy: The mobile agents have control on the actions and their internal state, and no user intervention.
- Decision making: This property is of reactive type and proactive type. In reactive decision making, the user intervention is involved. However, taking decisions proactively is without user intervention, an example of BDI model [17].
- Temporal Continuity: The agents are running continuously with time.
- Goal Based : The main purpose of the agent is to meet the defined goal.

The other properties are,

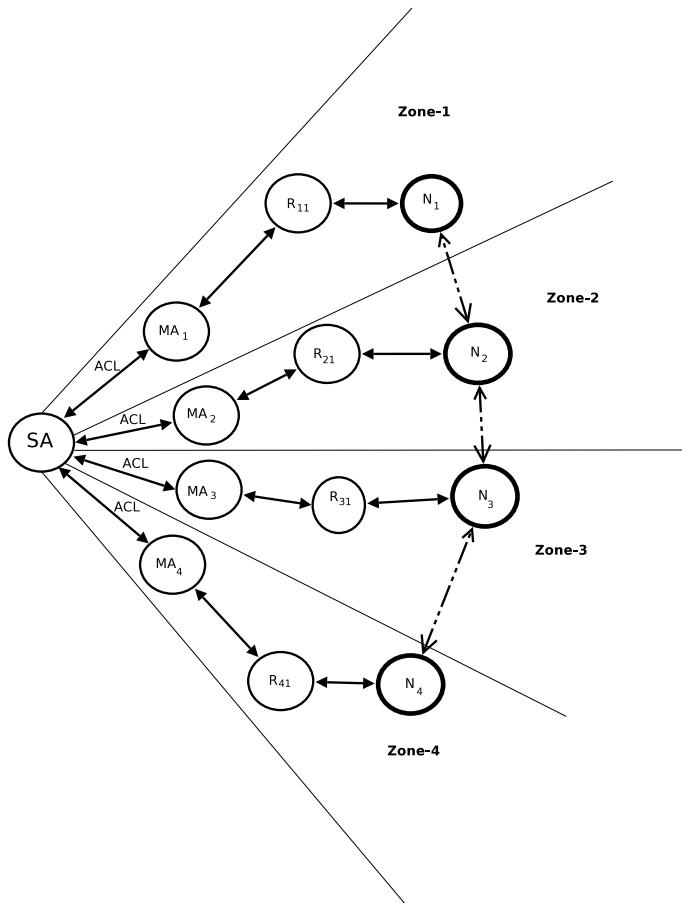
- Mobility: The agents have capability to move around the network easily.
- Communicative: The agents make use of agents communication language (ACL) to communicate with other agents.
- Mutual: The agent has an ability to solve the given task in co-operation with other agents.
- Learning capability: The agent has learning capability from environment, and it makes use of user preference and does the reasoning to a certain degree for intelligent decisions to improve the system efficiency.

## 2 QoS Management Protocol

The previous work discussed in [18] represents an agent-based node monitoring protocol designed for MANETs. This protocol makes use of a static agent (SA) and required number of mobile agents (MAS) one for each zone, to periodically monitor the health conditions of the nodes. The zones along with the respective sectors are formed using coordinates method discussed in [18].

To illustrate the functioning of the protocol, we consider an application running at nodes with  $N_1$  as the source and  $N_4$  as the destination as shown in Fig. 1.  $N_1$  is in zone 1,  $N_2$  is in zone 2 and  $N_3, N_4$  are in zones 3 and 4, respectively. At a given time  $t_1$ , let us assume the application needs 10% of the bandwidth more than it has been allocated to the nodes to run the application. This information has been passed on to SA successfully by  $MA_1, MA_2, MA_3$ , and  $MA_4$ , respectively, then SA calculates and reduces the bandwidth for the nodes  $R_{11}$  to  $R_{41}$  of the zones 1 through 4 and informs MAS to reallocate the bandwidth to  $R_{11}$  to  $R_{41}$ . MA reallocates without loosing the generality of application; i.e.,  $MA_1$  reallocates the bandwidth to the node  $N_1$  of zone 1. Similarly, it is followed by  $MA_2, MA_3, MA_4$  to distribute the resources to  $N_2, N_3$ , and  $N_4$  of zones 2, 3, and 4, respectively.

The basic principle of the proposed QoS management protocol is to dynamically distribute the required resources to the applications based on their priority and to



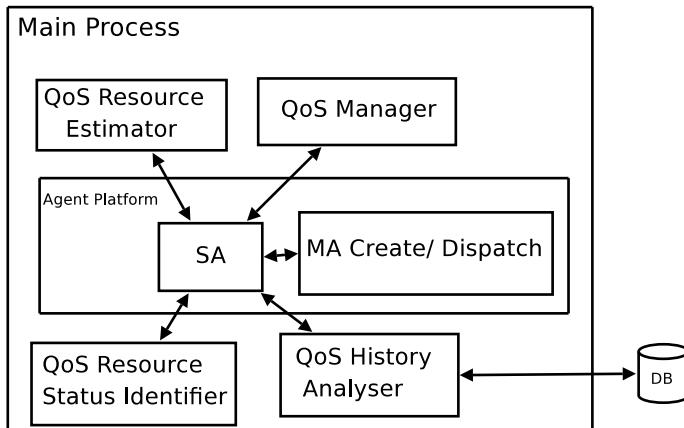
**Fig. 1** QoS management in a mobile ad hoc network

meet the variable requirements of the applications with the help of negotiation and renegotiation techniques.

The functioning of the protocol is based on two processes running on different nodes.

- (i) The main process runs at the central node, which controls the activities.
- (ii) and the other one which is a sub-process runs at mobile agent which is at the mobile node in a zone of the MANET.

The protocol functioning is given in subsections, by describing its architecture in detail.



**Fig. 2** Main process architecture of the QoS management protocol

## 2.1 Protocol Architecture

The protocol is divided into two parts QoS management part and resource management part of the protocol. The main process architecture of the QoS management protocol as shown in Fig. 2 gives its functionality by giving the description of each components, which basically focuses on QoS negotiation and renegotiation processes in coordination with SA and MA.

### 2.1.1 Static Agent (SA)

SA runs, at the RRN along with the main system, which creates MAs and sends them to specified zones of MANET, which plays a role for allocating the QoS resources to all applications of the nodes of the zone.

The purpose of SA is to interact with various system modules at a node where it stays and generates MAs based on the requirement.

SAs Functionality at the central node are:

- (i) It communicates with MA in order to get the nodes condition (QoS Monitoring process).
- (ii) It gathers the information regarding the utilization of resources of nodes of all zones with the help of MA.
- (iii) It gets the resource scarcity information using QoS resource status identifier module.
- (iv) It also predicts the resource scarcity with the help of past history based on information provided by QoS history analyzer.
- (v) It communicates with QoS manager for QoS management with control actions.

As a part of protocol functionality,

- Based on resource utilization history, node migration, computation cost of applications, SA does the assignment of priority to the nodes of MANET with QoS.
- SA dispatches the respective MAs with negotiation/renegotiation functionality in coordination with QoS manager, to all zones of the network periodically to satisfy the varying QoS needs of the running applications at the node.
- SA interprets the negotiation results of the MA with QoS manager and sends an information for fixing up the reserved resources if negotiation succeeds.
- SA predicts the QoS requirement if MA fails to bring the information using QoS history.

### 2.1.2 QoS Resource Status Identifier

With the monitored information, this module identifies the status of QoS resources of the nodes running different applications, with the help of QoS history analyser and informs back to SA.

Algorithm 1 represents the functionality of QoS resource status identifier.

---

#### Algorithm 1 Working of QoS Resource Status Identifier

---

```

1: Begin
2: Define
   Zones  $Z = \{Z_i \mid i \in \{1, 2, \dots, n\}\}$ ,
   Sectors  $C = \{C_q \mid q \in \{1, 2, \dots, m\}\}$ ,
   Nodes  $X = \{X_s \mid s \in \{1, 2, \dots, r\}\}$ .
3: for zone  $i = 1$  to  $n$  do
4:   create and deploy  $MA_i$  from SA to zone  $i$ .
5: end for
6: for zone  $i=1$  to  $n$  do
7:   for sector  $q=1$  to  $m$  do
8:     for node  $s=1$  to  $r$  do
9:        $MA_i$  collects status information and computes resource parameters at a node.
10:      end for
11:       $MA_i$  sends collected and computed information of nodes of zone  $p$ , sector  $q$  to SA.
12:    end for
13:    SA collects zone  $i$  information after time  $T_i$ .
14:    SA does interpretation with the help of other modules.
15:  end for
16: SA provides informatin to Resource status identifier.
17: if Computed Resource Utilization Value  $\leq$  Threshold Value then
18:   Increase periodicity of MA visiting particular zone.
19: end if
20: End

```

---

### 2.1.3 QoS Resource Estimation and Allocation Module

It gives the required QoS resource requirement information of the application to SA depending on the task that needs to be done. Depending on the priority computed, it estimates the resources required by finding the utility of the applications of nodes and allocating the resources with QoS with the help of MA.

### 2.1.4 QoS Manager

The QoS manager is the module defined with many functionality such as QoS mapping process, providing QoS levels and SLA, defining QoS policies, QoS controller, importantly negotiation and renegotiation. QoS manager mainly supports QoS mapping; i.e., it converts the QoS parameters to system level from user level. It instantiates the MAs (creates) with the help of SA and passes the negotiation and renegotiation functions to MAs through SA for negotiation of QoS resources at the destination based on the application requirement.

*Negotiation:* The dispatched MA interacts with local agent on the node with respect to the QoS profile of the node. After completion of negotiation process, it returns to the central node (where SA resides) with negotiation results and confirms the QoS requirements of the application of node. The local agent of the node is the one, which interacts with MA to provide QoS guaranteed service and systematize the available resources in an efficient manner. Negotiation of resources is done for getting required level of QoS guarantee for the application.

Algorithm 2 represents the mechanism for negotiation.

*Renegotiation:* The local agent at the node invokes the renegotiation process of MA to inform about the released resources through process interface, as MA visits the node periodically. MA carries the information regarding released resource information, then with SA interaction it renegotiates with local agent and finally does the redistribution of resources based on application resource utility information. The renegotiation of resources is carried out when QoS degrades.

Algorithm 3 represents the mechanism for re-negotiation.

### 2.1.5 QoS History Analyzer

Previous behavior of the application and its resource consumption history is analyzed with the module to predict behavior of nodes.

**Mobile Agent (MA)** The sub-process, which is termed as resource management part of the protocol, runs at MA (see Figure 3.). It includes sub-processes as, resource information collector, resource allocator, and process interface.

---

**Algorithm 2** Resource Negotiation Algorithm

---

```

1: Begin
2:  $R_s = Fetch\_resources(Profile\_QoS, Profile\_QoS.grade(s))$ 
3: if ( $R_{Suff}$ ) then
4:    $R_{Alloc} = R_s$ 
5:   if ( $func(R_{Assigned}) < R_s$ ) then
6:     Degradation confirmed
7:   else
8:      $R_{Alloc} = R_s$ 
9:     if ( $Degradation\_Confirmed$ ) then
10:       if ( $Profile\_QoS.grade \neq Null$ ) then
11:          $R_s = Fetch\_resources(Profile\_QoS, Profile\_QoS.grade)$ 
12:          $R_{Alloc} = R_s$ 
13:       else
14:          $R_{Alloc} = 0;$ 
15:          $R_{Consumed} = R_{Consumed} + R_{Alloc}$ 
16:       end if
17:     end if
18:   end if
19: end if
20: End

```

---



---

**Algorithm 3** Resource Renegotiation Algorithm

---

```

1: Begin
2: if ( $Resources\_ReleasedR_r$ ) then
3:    $R_{TOTAL} = R_{avail} + R_r$ 
4: end if
5: if ( $R_{Suff}$ ) then
6:    $Realloc\_list(Initial\_degraded\_list)$ 
7:    $Sort\_descending(Realloc\_list)$ 
8: end if
9: if ( $RRU_i \leq R_{min\_avail}$ ) then
10:    $R_{min\_avail} = R_{min\_avail} - RRU_i$ 
11:    $R_{consumed} = R_{consumed} + RRU_i$ 
12: else
13:   return
14: end if
15: End

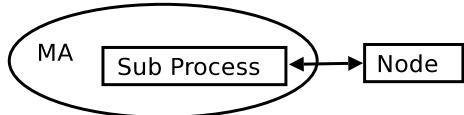
```

---

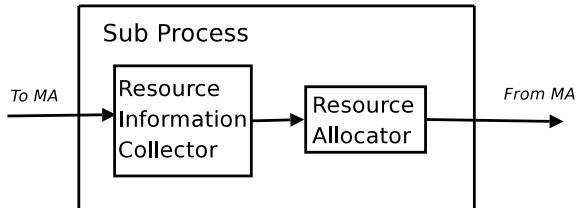
## The functions of MAs

- (i) Based on the QoS requirement information of application at a node, it interacts with the node for negotiation.
- (ii) Collects the negotiation results and informs back to SA.
- (iii) Allocates the required resources with QoS to the application if negotiation succeeds.
- (iv) Predicts the node migration and reports to SA.
- (v) Initiates a renegotiation process, as and when running application ends and releases resources.

**Fig. 3** Sub-process running at MA of the protocol



**Fig. 4** Architecture of the sub-process of the QoS management protocol



- (vi) The negotiation/renegotiation decisions are taken by main process, but the important function of the MA is for resource distribution (allocation) and resource redistribution in coordination with SA.

Figure 4 represents the sub-process of protocol running at MA of the node; i.e., MAs functionality, which is referred to as resource management part of the protocol, includes resource information collector, resource allocator, and process interface. It is assumed that each node is running with QoS agency, which includes local agent, QoS profile. The QoS profile includes users QoS requirements, and many grades of service quality are specified in the profile. Each grade contains the corresponding resource requirement defined with the help of set of system-level parameters like bandwidth, delay, and delay jitter.

Initially, MA interacts with local agent of the node through process interface as each node is installed with agent platform, and negotiation/renegotiation process takes place and the negotiation results are passed to SA. Later, resource allocation process running at MA does the allocation and reallocation of resources as per requirement.

### 3 Simulation Results

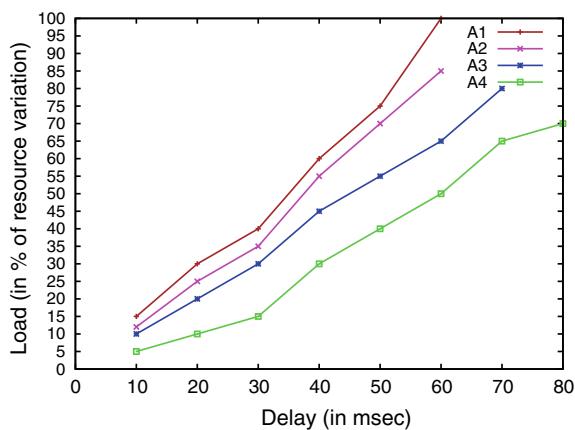
The proposed protocol is simulated using JADE framework with well-defined network environment and values as shown in Table 1.

MAs at the nodes of the zones will gather information about resource status and location periodically.

The JADE agent framework generates the mobile agents for QoS negotiation/renegotiation at the nodes of the network, and the negotiated information is returned to static agent which resides at RRN to take the final decision. To discuss the scheme, four applications, i.e.,  $A_1$ ,  $A_2$ ,  $A_3$ , and  $A_4$  running at different nodes with different sampling rates of 20, 16, 12, and 8KHz, respectively, are considered. The results are generated, with average value over simulation trials. The load resource

**Table 1** Parameters of the network model

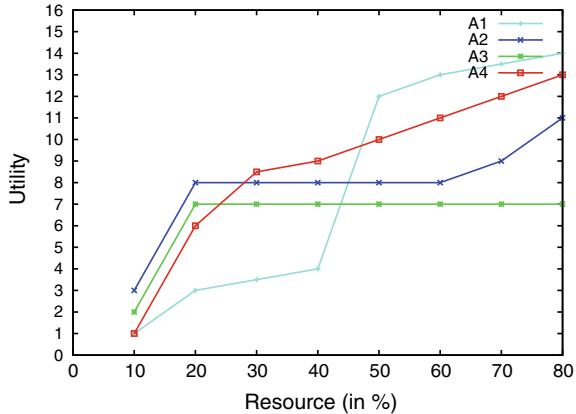
Parameter	Value
Simulation Area	100 × 100
No. of Nodes	100
Transmission Range	50–200
SST value	20
Speed	0–40 m/s
Agents	4–16
Agent Size	4 Kbytes
Hops	2 to 16
Applications	50–100
Power	0–99

**Fig. 5** Load (in %) with delay

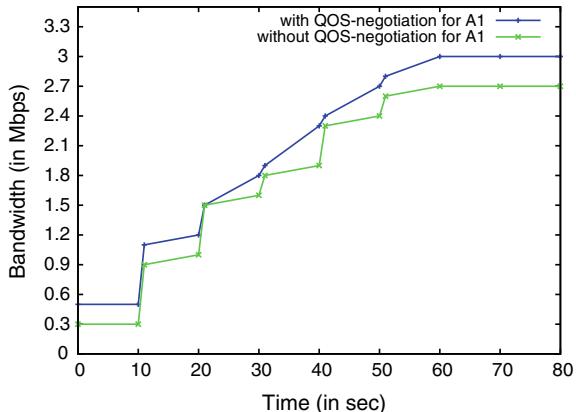
consumption with end-to-end delay for the applications with QoS dimension into consideration is shown in Fig. 5.

We note that the Fig. 6. represents utility of resources with number of applications with varying sampling rates. The improvement in the resource utilization using agents is shown in Fig. 7. The Fig. 7 shows the proposed algorithm performance over the existing general admission control algorithm, where the network resource utility of the algorithm is increased by 10% over the general admission control algorithm. As an example, considered in the simulation scenario in network the nodes will have the bandwidth as the resource between 4 and 13 Kbps for audio applications. After the allocation of bandwidth to the application, minimum required resource allocated to application varies from 2 to 7 Kbps and to a maximum of 8–13 Kbps. The negotiation policy which is introduced allows for the adaptation of applications, definitely going to have improved performance with higher resource utilization.

**Fig. 6** Applications resource utility



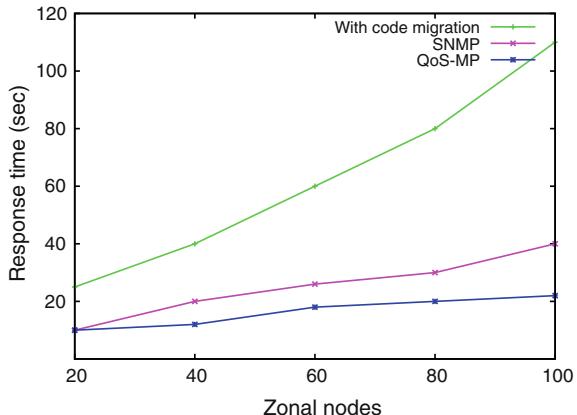
**Fig. 7** Bandwidth utilization over time



Validation of the proposed protocol performance is carried out with the response time of MAs in order to provide the QoS management. The proposed work shows the considerable response time reduction. The QoS resource bandwidth allocation for the nodes running the MANET is considered, and the response time for the same is measured. In wired networks with their devices which supports the agent platform (QoS agency), the SNMP implementation is carried out accessing the stubs and without code migration gave response time and network overhead justification.

In our proposed work, we have shown QoS management of resources of each zone by a mobile agent periodically and clusterwise. Figure 8 shows results for response time of our proposed work QoS-MP with relation to other techniques discussed. In the simulation, the response time of QMP-MA is better compared to other techniques discussed.

**Fig. 8** Response time with number of zonal nodes



## 4 Conclusion

A novel QoS management technique is discussed for MANETS. The proposed protocol makes use of mobile agents and static agents to provide the required QoS, depending upon resource utilization information of the applications running at the node. MA makes sure of QoS provision of network resources on demand. The proposed work ensures the QoS provision based on application and their resource utilization history, and the associated response time of mobile agent to complete the task, i.e., QoS management in MANET with lower network latency.

## References

1. Li, W., Guizani, M., Kazakos, D.: Quality of service in mobile ad hoc networks. *EURASIP J. Wirel. Commun. Netw.* **2006**, 1–3 (2006)
2. Chen, X., et al.: Survey on QoS management of VoIP. In: 2003 International Conference on Computer Networks and Mobile Computing, 2003. ICCNMC 2003. IEEE (2003)
3. Chen, D., Varshney, P.K.: QoS support in wireless sensor networks: a survey. In: International Conference on Wireless Networks (2004)
4. Carneiro, G., et al.: The DAIDALOS architecture for QoS over heterogeneous wireless networks. In: Proceedings of the 14th IST Mobile and Wireless Communications Summit, vol. 22 (2005)
5. Yu, F.R., Tang, H., Bu, S., et al.: Security and quality of service (QoS) co-design in cooperative mobile ad hoc networks. *J. Wirel. Comput. Netw.* **2013**, 188 (2013)
6. Subramanian, S., Ramachandran, B.: Trust based scheme for qos assurance in mobile ad-hoc networks. arXiv preprint [arXiv:1202.1664.2012](https://arxiv.org/abs/1202.1664)
7. Shakya, S.: Intelligent and adaptive multi-objective optimization in WANET using bio inspired algorithms. *J. Soft Comput. Paradigm.* **2**, 13–23 (2020). <https://doi.org/10.36548/jscp.1.002.2020>
8. Bashir, A.: Artificial intelligence based LTE MIMO antenna for 5th generation mobile networks. *J. Artif. Intell.* **2**(3), 155–162 (2020)

9. Calafate, C.T., et al.: QoS support in MANETs: a modular architecture based on the IEEE 802.11 e technology. *IEEE Trans. Circ. Syst. Video Technol.* **678–692** (2009)
10. Manvi, S.S., Venkataram, P.: Mobile agent based approach for QoS routing. *IET Commun.* **430–439** (2007)
11. Haoxiang, W., Smys, S.: Soft computing strategies for optimized route selection in wireless sensor network. *J. Soft Comput. Paradigm (JSCP)* **2**(1), 1–12 (2020)
12. Dhaya, R., Kanthavel, R.: Bus-based VANET using ACO multipath routing algorithm. *J. Trends Comput. Sci. Smart Technol. (TCSST)* **3**(01), 40–48 (2021)
13. Chowdhury, C., et al.: Reliability estimate of mobile agent system for QoS MANET applications. In: *Reliability and Maintainability Symposium (RAMS)*, IEEE (2011)
14. Natkaniec, M., et al.: Supporting QoS in integrated ad-hoc networks. *Wirel. Pers. Commun.* **56**(2), 183–206 (2011)
15. Rath, M., et al.: Monitoring of QoS in MANET based real time applications. *Smart Innovation, Systems and Technologies*, vol. 84. Springer, Berlin (2018)
16. Rath, M., et al.: QTm: a QoS task monitoring system for mobile ad hoc networks. In: *Recent Findings in Intelligent Computing Techniques. Advances in Intelligent Systems and Computing*, vol. 707. Springer, Berlin (2019)
17. Boukhtouta, A., et al.: Coordination strategies and techniques in distributed intelligent systems-applications. In: *Advances in Practical Multi-Agent Systems*, pp. 73–93. Springer, Berlin (2011)
18. Channappagoudar, M.B., Venkataram, P.: Mobile agent based node monitoring protocol for MANETS. In: *National Conference on Communications (NCC)*, vol. 1, issue no. 5, pp. 15–17 (2013)

# Design of IoT Platform for Monitoring and Control of Variables of Industrial Processes



Hernando González, Azarquiel Diaz, Luis Jaimes, and Carlos Meza

**Abstract** Emerson DeltaV is a distributed control system (DCS) that works in a decentralized way and offers hardware and software for large applications and advanced control in the industrial sector. Currently, Emerson is targeting each system to specific industries, such as the oil, chemical, food, beverage, and pharmaceutical industries. This paper presents the design and implementation of an IoT monitoring system. The data acquisition and transmission system use the microprocessors, ATmega2560-R3, ESP32s, and MAX485; the data are transmitted from the DeltaV control system to a server arranged in Heroku. The server is programmed to manage the storage and consultation of the data, which is stored in a MongoDB database. The variables of the processes controlled by DeltaV can be observed and manipulated from a Web interface.

**Keywords** Industry 4.0 · Distributed control system (DCS) · IoT platform · Heroku · MongoDB

## 1 Introduction

The distributed control system (DCS) is a modern system that can control different processes in parallel, with a central controller that works as the brain, remote terminal units (RTUs), among other control elements that allow information to be taken, processed and act regardless of whether the sensors and actuators are not close to

---

H. González (✉) · A. Diaz · L. Jaimes · C. Meza  
Universidad Autónoma de Bucaramanga, Bucaramanga, Colombia  
e-mail: [hgonzalez7@unab.edu.co](mailto:hgonzalez7@unab.edu.co)

A. Diaz  
e-mail: [adiaz106@unab.edu.co](mailto:adiaz106@unab.edu.co)

L. Jaimes  
e-mail: [ljaimes9@unab.edu.co](mailto:ljaimes9@unab.edu.co)

C. Meza  
e-mail: [cmeza823@unab.edu.co](mailto:cmeza823@unab.edu.co)

the controller [1]. This type of control system is used in power generation plants, sewage treatment systems, mining processes, environmental control systems, and in petrochemical and mining industries [2].

Current technological developments such as automated buildings, smart homes, telemedicine, surveillance systems, and the automotive industry are due to the development of the Internet of Things (IoT). These developments allowed the realization of new interconnected devices, establishing communication protocols and software development to ensure communication is established with the least amount of data loss. The IoT platform fosters the development of different sectors of the process industry, such as oil processing, pharmaceutical industry, agricultural industry, and renewable energy sector, among others. These technological developments allow a change in product quality and productivity increase of 1–2%, generating significant benefits through manufacturing time reduction, impact on the environment and waste reduction. All IoT platforms provide tools that integrate a wide set of functionalities through numerous APIs; this is due to the great diversity of applications and electronic devices that are currently being realized. This has become a challenge for software developers and platform administrators because they must adapt the available applications to customer needs: cost, number of connected devices, programming languages, and available graphical interfaces. Nowadays, IoT platform developers and administrators are not only concerned with ensuring that applications work according to customer needs but also with guaranteeing information security and network privacy among different users [3–6].

The control devices used in automation processes are as follows: the programmable logic controller (PLC) for small processes and distributed control systems (DCSs) for larger processes, for example, the petrochemical industry. Emerson Automation Solutions has been developed and supported a prototype system in the field for design validation and long-term performance evaluation. Currently, papers related to IoT platform have been developed for the DeltaV DCS, where the solution proposed consists of using field devices that have HART and WirelessHART networks together with the DCS DeltaV; these devices transmit the sensor reading and also its status and connect to the Microsoft Azure cloud through the "IoT edge gateways" service, which enables gateways for devices to store and monitor information [7]. A second example of an IoT platform for industrial processes is presented in [8], in which three slave modules and an NXPlpc1768 Master are used; each slave device consists of an ATmega328 microcontroller and the instrumentation to measure the relevant process variables. The microcontroller transmits the information provided by the instrumentation to the CAN controller MCP2515, which in turn connects to the Master NXPlpc1768 to transmit the data to the cloud. From a Web page, it is possible to read the value of the process variables and allows users to make decisions according to the information recorded by each sensor.

## 2 IoT Communication System

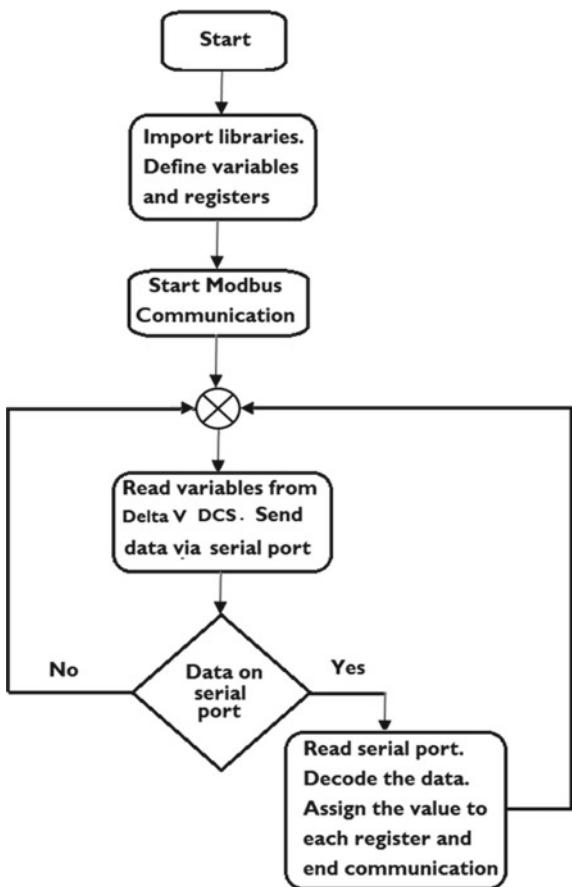
The system consists mainly of a boiler, which is responsible for generating steam for the heat exchangers and the water flow control system. In full, the number of variables to be monitored is 20, of which two allow to manipulate the set-point of the water flow and the temperature of the hot water at the outlet of the heat exchanger; the other variables are control signals for solenoid valves and to monitor of process signals like temperature, flow, and pressure. The system is installed in the Pilot Plant laboratory of Universidad Autónoma de Bucaramanga. The proposed system consists of the ATmega2560-R3, ESP32s, and MAX485; the latter allows communication with the DCS DeltaV using its serial communication card and configuring the Modbus RTU protocol. Figure 1 shows the connection that is made between the different units: the Max485 receives data through Modbus RTU and send them to ATmega2560-R3 microcontroller; the pins A and B of device are connected to the serial card of the DeltaV, and the communication is half duplex; the Rx and Tx pins of device are connected to the serial port to the ATmega2560-R3. Through another serial port of ATmega2560-R3 sends the data to the ESP32s, this microprocessor has the ability to connect to a Wi-Fi network, being connected to the Internet makes HTTP requests to the web server to store the data. The communication protocol programming in DeltaV is done in control studio, which presents blocks with specific functions that are structured in a similar way of a flux diagram.

Figure 2 shows the program that executes the ATmega2560-R3; it imports the Modbus communication libraries and defines the registers that it will handle; in this case, there will be 17 read registers and 3 write registers. After defining the registers, the serial communication with the ESP32s card and the Modbus communication with DeltaV are configured. The loop cycle is executed every certain time, in this case two seconds; each period of time, DeltaV variables are read if there is communication from the ESP32s card through the serial port; the reading variables are sent, and the register is updated of the write variables if there is no response from the server; the ATmega2560-R3 remains in a loop waiting for communication. As shown in Fig. 3, the program that runs the ESP32s establishes the connection to the Internet and initializes the serial communication. In the loop cycle, the Internet connection is verified if there is no connection then a reconnection loop enters being connected to the Wi-Fi network, so it reads the serial port to send the data to the server; in case there is no serial communication, it indicates an error in connection with DeltaV.



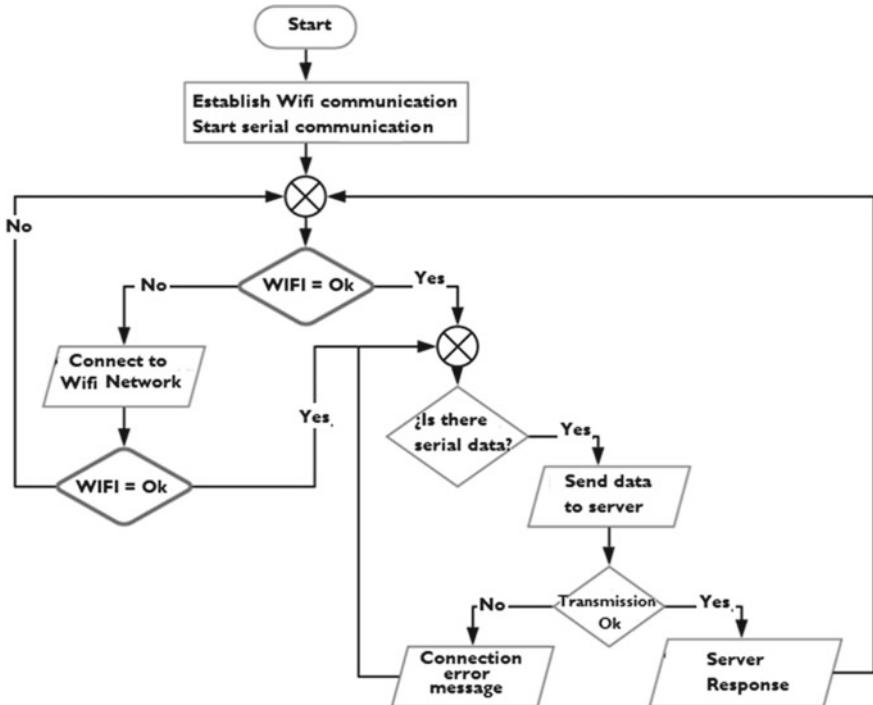
**Fig. 1** Microprocessor communication scheme

**Fig. 2** ATmega2560-R3 flow diagram



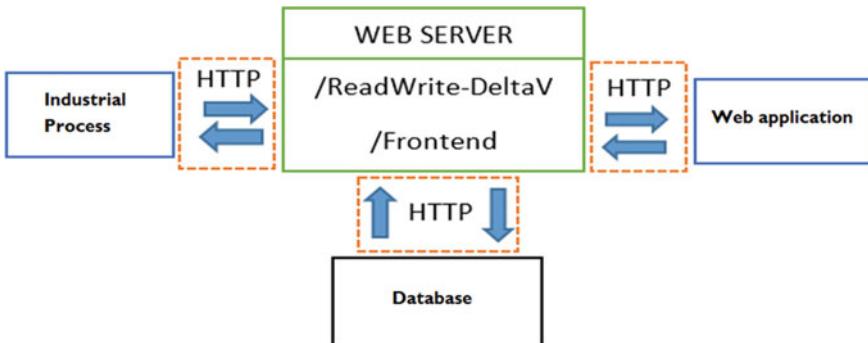
## 2.1 Server

For the development of this project, the platform used was Heroku although there are more server providers such as Amazon Web Services (AWSs), Microsoft Azure, Google App Engine, and AppScale. AWS is a collection of public cloud computing services that together form a cloud computing platform, offered over the Internet by Amazon.com. Microsoft Azure is a cloud computing service created by Microsoft to build, test, deploy, and manage applications and services using its data centers. Google App Engine is a Web hosting service provided by Google free of charge up to certain quotas. This service allows applications to run on Google's infrastructure. The fourth server, AppScale, is a software company that provides cloud infrastructure software and services to enterprises, government agencies, contractors, and third-party service providers. Heroku was developed since June 2007, which runs the application in virtual containers called Dynos. This server allows the programmer to scale the application by increasing the number of Dynos or by making a change in the



**Fig. 3** ESP32s flowchart

Dyno properties [9]. Heroku stands out, since it allows the implementation, execution and administration of Web applications, in different programming languages: Python, Java, Ruby, Clojure, Scala, Node.js, Go, and PHP [10]. To configure the server's microservices, the Python Flask library was used, which allows defining and executing functions through HTTP requests to a URL. In [11], an encryption system is presented to be implemented on a Heroku server, using the Advanced Encryption Standard protocol—AES-128. AES is one of the most widely used and secure encryption algorithms currently available. This method is considered to be very secure and efficient and is used to encrypt data of all types and is therefore often used in various protocols and transmission techniques. The algorithm uses 128 bytes; the encryption process is performed in four steps: byte substitution, row shifting, column shuffling, and finally, each byte is combined with a key using the XOR operation.



**Fig. 4** Web server structure

## 2.2 Server Structure

The Web server is the central hub; it attends to all the queries coming from DeltaV and the Web application. As shown in Fig. 4, the ESP32s card makes HTTP requests to the server. The data are entered in the ReadWrite-DeltaV function; this function takes the information and stores it in the database to make the connection between the server and MongoDB. The PyMongo Python library is used, which allows creating, deleting, and updating the information in the database from a Python application. The FronEnd function handles display requests in the Web and mobile interface, in addition to the users session. It should be noted that the database can only be managed by the server that is the end devices that transmit the variables, and the Web application do not have direct communication with the database, this to eliminate security problems.

## 2.3 Database

Since the role that the database will perform is real-time data storage and query, it requires a fast query response and handling a large volume of data; The SQL database was not used because its storage structure is not suitable for this type of work. Based on the comparison made by [12] in which the databases MongoDB, ArangoDB and CouchBase, were evaluated in two scenarios “Read Only” and “Heavy Read” for two data sources. As a result of this study, MongoDB gave the best results in the evaluation of the response time and obtained the least dispersed standard deviation. In [13], a comparison of the performance of two databases: MySQL and MongoDB are performed by querying the COMEX database; this database contains the information of Brazil’s foreign trade statistics and is consulted by several external agencies. The results of the article are analyzed under three criteria: the first is a single client with multiple write, update, and delete operations; the second case corresponds to a client

**Fig. 5** Data storage structure

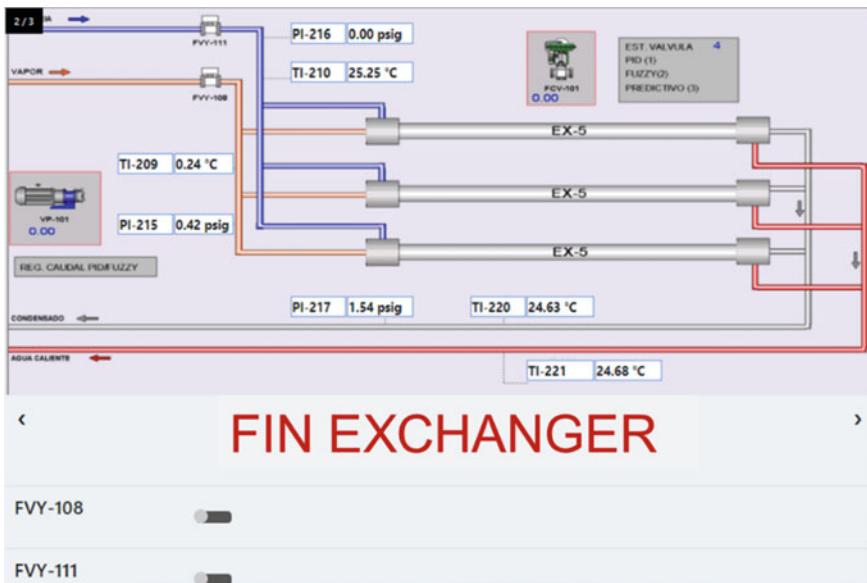
```
{
  "_id": {
    "$oid": "5fb1459e1983b0a397b0dc8d"
  },
  "varname": "var0",
  "varT0": [...],
  "varT1": [...],
  "varT2": [...],
  "varT3": [
    "29/9/2020/11/43/22#0#81#4077#17498",
    "29/9/2020/11/43/25#0#81#4076#17499",
    "29/9/2020/11/43/28#0#81#4076#17500",
    "29/9/2020/11/43/31#0#81#4075#17501",
    "29/9/2020/11/43/34#0#81#4074#17502",
    ...
  ]
}
```

with a single read operation; the last case corresponds to a multi-client with a single read operation. The results show that the MongoDB database took only 4% of the time required by the MySQL database; moreover, the MongoDB database presented a performance of over 92%, in simple operations, but in more complex operations both databases presented the same performance.

For this project, MongoDB was selected because it allows users to manage a free storage plan. The data stored in MongoDB are divided into documents, which have a JSON structure, and a storage size limit of 2 MB, so that the historical record of the variables is divided into several documents; as shown in Fig. 5, the file format is a JSON, which has by default a variable “\_id” to identify this document within the database; “varname” indicates the numbering of the document; in this case, it is “var0” because it is the first document created for this record. Since MongoDB also restricts the number of data within a vector, the document is made up of 16 vectors named “varT” of 1000 spaces, which yields a total of 16,000 data per document; this is equivalent to 1.81 MB so that meets the size limit [14]. The variables stored within each vector are concatenated into a single data type “string” separated by the symbol “#,” each time the record is updated, the date on which it was stored is added.

## 2.4 Web Interface

A Web application is composed of the backend which is in charge of carrying out the processes and where the logic is programmed; for that JavaScript was used linked to the services provided on the server, which are developed in Python, and the frontend that corresponds to the graphic part of the site. For its development, HTML files are



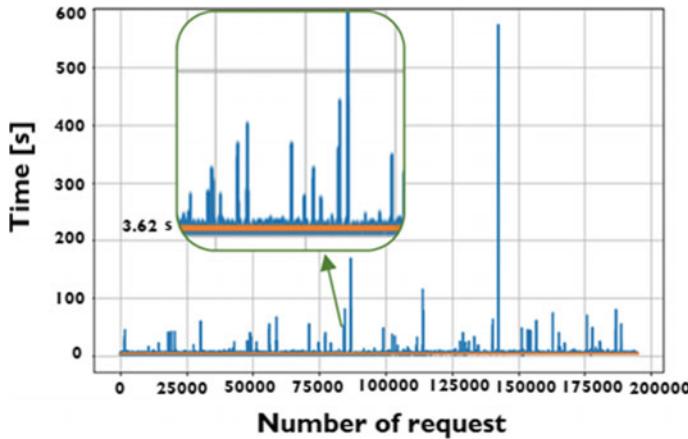
**Fig. 6** Web interface of heat exchanger (fin exchanger)

used to define the structure and add functionalities and CSS that adds the style of the interface. In Fig. 6, the heat exchanger (fin exchanger) and the variables that are being monitored are shown, among which are temperature and pressure, and the controls that allow the electromagnetic valves to be activated or deactivated are observed in the lower part.

### 3 Results

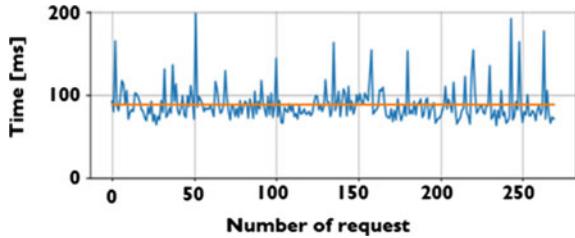
#### 3.1 Server Operation

To validate the connection between DeltaV DCS and the server, the data transmission system was left running for a week, making requests every two seconds to store the date, the hour, and process variables in the MongoDB database. Figure 7 shows the response time of the server to the request; the horizontal axis corresponds to a consecutive number associated with the request to store the data on the server. There are large time deltas; the largest is equivalent to about 10 min, and the other peaks are around two minutes; this is due to failures in the Internet connection. The average request time can be seen represented in the graph as the orange line and is 3.64 s.



**Fig. 7** Time delta in data log

**Fig. 8** Server response time

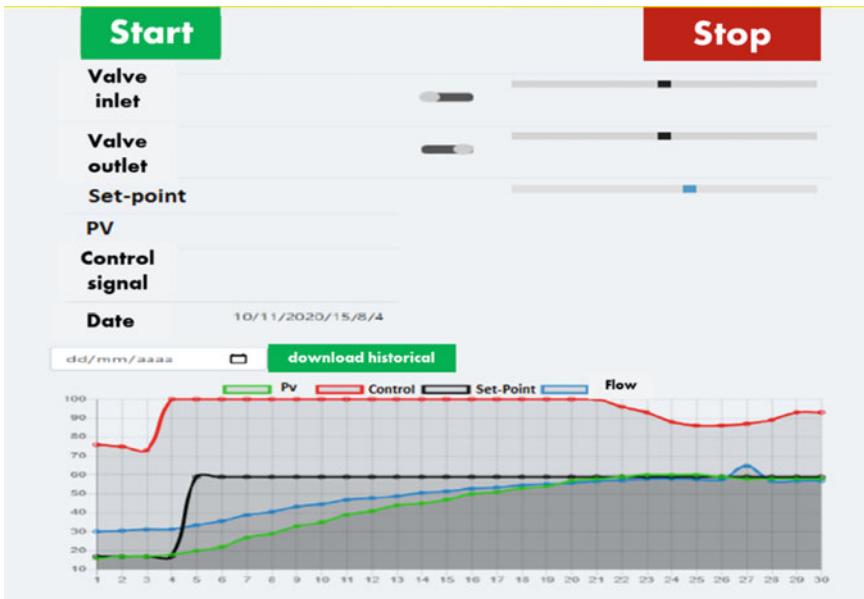


### 3.2 Server Response Time

The time takes a user to read a data of the MongoDB database is analyzed, which cannot be greater than 500 ms. When the time is greater than this limit, the request returns an internal error of server; to validate that the requests made to the server are less than this time; a script was made that simulates 270 consecutive requests to the server. Figure 8 shows the execution time, and the dispersion between the data of each request is low and is not close to the limit even though the test was carried out with a historical record of one week.

### 3.3 Human Machine Interface

Figure 9 shows the control panel; the user can manipulate the set-point, also activate and deactivate the valves located in the process. The option “download historical” allows download the process variables in an Excel file with a sampling time of one minute. In order to ensure the security of the data in the cloud, three cryptographic



**Fig. 9** Human machine interface

methods were combined: AES-256, secure hash algorithm (SHA-512), and information dispersal algorithm (IDA). The original data were encrypted using the AES-256 algorithm, and then, the encrypted file is split into several files using the SHA-512 and IDA algorithms. This topic has been analyzed by several researchers and corresponds to a second stage of the project to develop and implement new encryption algorithms for both the MongoDB database and the Heroku Web server [15, 16].

Another topic of study is the use of artificial intelligence techniques for the processing of data from the different industrial processes located in the laboratory that allow detecting faults. An example of this technology is presented in [17], in which data mining techniques are used to detect faults or defects in solar power generation systems, executing a preventive action that allowed an increase in the efficiency of the solar station. In [18], an ant colony optimization algorithm was programmed to optimize an IoT wireless network, obtaining lower energy consumption and reduced transmission delays between nodes.

## 4 Conclusions

The Heroku Web server offers a free development plan; however, the use time per month and the work capacity are limited, when deploying an application with several functions; it makes the server slow, so it is recommended to expand the capacity

of work and segment functions into microservices. The style of the Web interface was based on a template; for this reason, some of the added characteristics do not share the style and tend to lose the esthetics; in the future, it is recommended to develop your own template in order to add the necessary characteristics that the application requires. Since the database allows to create its own structure, different ways of storing the historical records of each process could be evaluated; this allowed to improve the query times to the database and reduce the workload of the server. Although the ESP32s microprocessor can be used independently, it was connected to the ATmega2560-R3 since making requests to a server consumes resources and adding Modbus communication using communication peripherals such as the MAX485, make communication collapse.

## References

1. Luz, D., Forero, L.: Estudio y Diseño de una Plataforma de Entrenamiento de Alto Nivel en Control Electrónico a partir de Sistemas de Control Distribuido (DCS). Universidad Distrital Francisco José De Caldas, Bogotá, Colombia (2015)
2. Quispe, O.G.: Diseño e implementación de un módulo de pruebas para la simulación de operaciones del sistema de control de procesos (PCS) y del sistema instrumentado de seguridad (SIS) DeltaV Emerson para el área de proyectos y servicios de la empresa SEIN S.A. Escuela Politécnica Nacional, Quito. Ecuador (2016)
3. Celik, Z.B., McDaniel, P., Tan, G.: Soteria: automated IoT safety and security analysis. In: USENIX Annual Technical Conference, USENIX ATC, Boston, MA, 2018. <https://www.usenix.org/system/files/conference/atc18/atc18-celik.pdf>
4. Celik, Z.B., Fernandes, E., Pauley, E., Tan, G., McDaniel, P.: Program analysis of commodity IoT applications for security and privacy: challenges and opportunities. ACM Comput. Surv. (2019)
5. Celik, Z.B., Tan, G., McDaniel, P.: IoT Guard: dynamic enforcement of security and safety policy in commodity IoT. In: Network and Distributed System Security Symposium, NDSS, San Diego, CA, February 2019
6. Babun, L., Denney, K., Celik, Z.B., McDaniel, P., Selcuk, A.: A survey on IoT platforms: communication, security, and privacy perspectives. Comput. Netw. **192**. ISSN 1389-1286 (2021). <https://doi.org/10.1016/j.comnet.2021.108040>
7. Wang, G., Nixon, M., Boudreaux, M.: Toward cloud-assisted industrial IoT platform for large-scale continuous condition monitoring. Proc. IEEE **107**(6), 1193–1205 (2019). <https://doi.org/10.1109/JPROC.2019.2914021>
8. Niveditha, A.T., Nivetha, M., Priyadarshini, K., Punithavathy, K.: IoT based distributed control system using CAN. In: Second International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2018, pp. 967–971 (2018). <https://doi.org/10.1109/ICCMC.2018.8488003>
9. Danielsson, P., Postema, T., Munir, H.: Heroku-based innovative platform for web-based deployment in product development at axis. IEEE Access **9**, 10805–10819 (2021). Electronic ISSN: 2169-3536. <https://doi.org/10.1109/ACCESS.2021.3050255>
10. Heroku: The Heroku platform (2021). Available: <https://www.heroku.com/platform>
11. Lee, B.H., Kusuma, E., Farid, M.: Data security in cloud computing using AES under HEROKU cloud. In: 27th Wireless and Optical Communication Conference (WOCC), 2018. Electronic ISBN:978-1-5386-4959-6. <https://doi.org/10.1109/WOCC.2018.8372705>
12. Treviño, M., Viquez, L., Quirós, R., Esquivel, G.: Una comparación de rendimiento entre MongoDB, ArangoDB y CouchBase para la operación lectura sobre bases de datos geográficas,

- 2018 IEEE 38th Central America and Panama Convention (CONCAPAN XXXVIII), San Salvador (2018). <https://doi.org/10.1109/CONCAPAN.2018.8596387>
- 13. Gomes, A., Lopes, V., Ribeiro, E., Lima, J., Costa, W., Garcia, L., Holanda, M.: An empirical performance comparison between MySQL and MongoDB on analytical queries in the COMEX database. In: 16th Iberian Conference on Information Systems and Technologies (CISTI), 2021. Electronic ISBN:978-989-54659-1-0. <https://doi.org/10.23919/CISTI52073.2021.9476623>
  - 14. Patil, M., Hanni, A., Tejeshwar, C.H., Patil, P.: A qualitative analysis of the performance of MongoDB vs MySQL database based on insertion and retrieval operations using a web/android application to explore load balancing—sharding in MongoDB and its advantages. In: International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC). February 2017. ISBN:978-1-5090-3244-0. <https://doi.org/10.1109/I-SMAC.2017.8058365>
  - 15. Kumar, J., Garg, V.: Security analysis of unstructured data in NOSQL MongoDB database. In: International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN), 2017. ISBN: 978-1-5386-0628-5. <https://doi.org/10.1109/IC3TSN.2017.8284495>
  - 16. Lee, B., Kusuma, E., Farid, M.: Data security in cloud computing using AES under HEROKU cloud. In: 27th Wireless and Optical Communication Conference (WOCC). June 2018. ISBN: 978-1-5386-4960-2. <https://doi.org/10.1109/WOCC.2018.8372705>
  - 17. Shakya, S.: A self monitoring and analyzing system for solar power station using IoT and data mining algorithms. *J. Soft Comput. Paradigm* **3**(2), 96–109 (2021)
  - 18. Chen, J.I.Z., Lai, K.-L.: Machine learning based energy management at Internet of Things network nodes. *J.: J. Trends Comput. Sci. Smart Technol.* **(3)**, 127–133 (2020)

# A Study on Identification of Plant Diseases Using Image Processing



Disha Sushant Wankhede, Amit Gamot, Kashish Motwani,  
Shaunak Kayande, Vidhi Agrawal, and Chetan Chinchulkar

**Abstract** Agriculture is considered to be one of the major sectors of a country's economy due to the rapid increase in the population size and increased demand for food. Agricultural development is, therefore, a precondition for a nation's prosperity. Several factors can affect agricultural development such as environmental conditions, natural disasters, and plant infection. The plants can be infected by the diseases due to bacteria or fungus during any possible stage, like the main field or nursery stage. There exists a term, "Plant Pathology" which means studying the diseases in plants by monitoring the patterns. As computerization is witnessing great success in many agricultural fields, therefore, to make this very time-consuming process much easier and efficient, image classification can be used. This paper primarily focuses on signaling existing cases of plant disease using image classification and identifying which machine learning model will be the most efficient for a particular given problem, in our case, plant diseases.

**Keywords** Image processing · Machine learning · Plant disease · SVM · ANN · Random forest classifier · KNN · Extreme learning machine

## 1 Introduction

At any certain time, on any certain day, there are only about three weeks enough food to feed all the people on earth. Thus, we know that since the beginning of time, agriculture has played an important role in producing food for human utilization. In a country, the nature of agriculture is reliant on product (crops/plants) quality and quantity. Elements such as weeds, pests, and diseases can be responsible for crop production loss.

---

D. S. Wankhede (✉) · A. Gamot · K. Motwani · S. Kayande · V. Agrawal · C. Chinchulkar  
Computer Engineering Department, VIIT, Pune, India  
e-mail: [disha.wankhede@viit.ac.in](mailto:disha.wankhede@viit.ac.in)

All plants can be infected by diseases. Plant diseases have an impact on society along with world history such as the great Irish potato famine in 1845. A delay in the detection of plant disease can also contribute to the cause of a decline in economic growth.

In the agriculture field, there are enormous advancements with the help of technology. Farmers face difficulties because the plowing fields are enormous and have a significant number of plants; therefore, it becomes very difficult for the naked eye to properly detect and categorize every plant. And doing so is very crucial as even singly diseased plants can spread the disease. Furthermore, the majority of farmers lack proper awareness of these diseases and how to treat them.

Modern technologies, such as ML and deep learning algorithms, have been made to accelerate the rate of recognition and accuracy. Several studies using classic algorithm approaches such as random forest, ANN, SVM, k-means method, and others have been conducted in the subject area of ML for the recognition of plant disease.

The image processing methods are to be employed for the identification of plant disease. In many cases, symptoms of diseases are seen on the fruit stem and also on the leaves. Accuracy is the main factor that will determine the automated plant disease detection and classification model's success.

## 2 Literature Review

Classification and disease detection using machine learning techniques are analyzed and compared in depth below.

### 2.1 *Support Vector Machine*

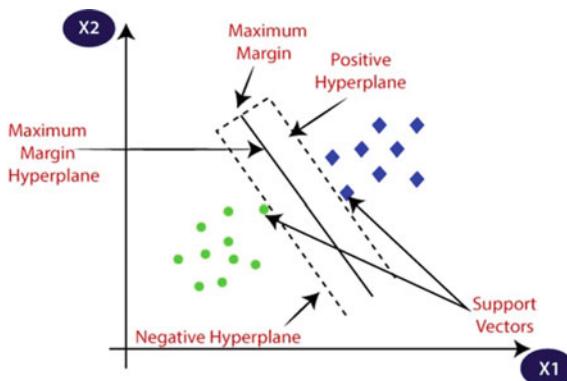
A machine learning (ML) procedure for categorization and regression that analyzes statistics and for categorizing data into two groups that use supervised learning is known as the support vector machine (SVM) which is also called a support vector network (SVN). It is used for a variety of tasks, including text classification, image classification, and handwriting recognition.

The SVM algorithm is preferred over the other algorithms because it can deal with classification problems using an SVM classifier and regression problems using an SVM regressor. The SVM classifier is the backbone of the support vector machine concept and, in general, is the most appropriate algorithm to solve problems regarding classification.

The basic purpose of the procedure is to discover a decision boundary or the optimal line for splitting n-dimensional space into separate classes so that a new data point can be allocated to the appropriate category rapidly. The decision boundary is referred to as a hyperplane.

**Fig. 1** [1] Two different categories that are classified by using a hyperplane.

Source JavatPoint Available: <https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm>



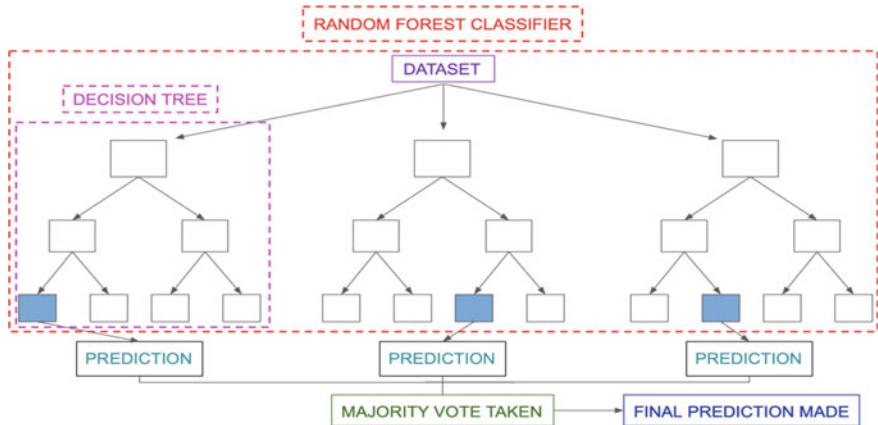
SVM is used to select the hyperplane's extreme vectors that aid in its formation. It is termed the support vector machine because of these extreme examples, which are known as support vectors (Fig. 1).

## 2.2 Random Forest

A supervised ML algorithm known as random forest can be used to solve problems like classification and regression. It combines many classifiers (decision trees) to provide solutions to complex problems. The “forest” produced by the random forest algorithm is trained by bagging or bootstrap aggregating. It anticipates by taking the mean of the output from various trees. As the quantity of trees increases, the precision of the outcome also increases.

Paper [2] represents the methodology for plant leaf disease detection, and classification using machine learning-integrated digital image processing techniques, it concludes that the random forest classifier outperforms compared with support vector machine, k-nearest neighbor, artificial neural network. However, with the greater number of classes and images, SVM gives significantly good accuracy compared with artificial neural network and k-nearest neighbor algorithms.

When compared to artificial neural network and k-nearest neighbor algorithms, SVM provides much better accuracy with a larger number of classes and images. A random forest annihilates the restrictions of the decision tree algorithm. It creates predictions without needing many configurations in packages and even handle large datasets efficiently. In [3], the author mentioned some authorized datasets related to plant soil (Fig. 2).



**Fig. 2** [4] Structure of random forest classifier. Source Miro medium, Available: [https://miro.medium.com/max/5752/1\\*5dq\\_1hnqkboZTcKFfwbO9A.png](https://miro.medium.com/max/5752/1*5dq_1hnqkboZTcKFfwbO9A.png)

### 2.3 Artificial Neural Network

An artificial neural network is a computer model based on the biological neural networks that make up the structure of the human brain (ANN). It is made up of multiple processing components that accept inputs and outputs based on their predetermined activation functions.

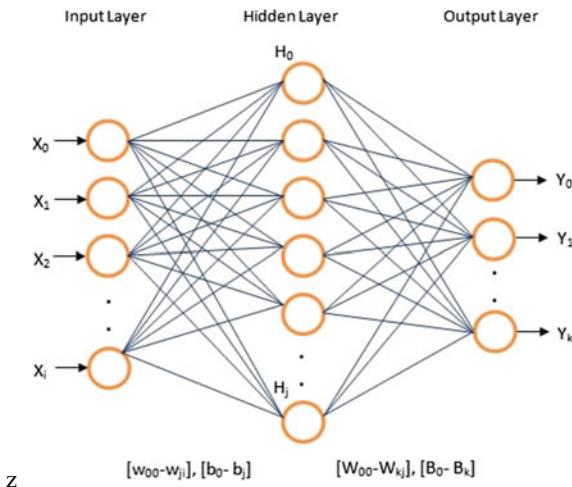
The activation function is a collection of transfer functions that are used to produce the desired result. Artificial neural networks, like the human brain, have nodes that are interconnected to one another in various layers of the network. An ANN typically consists of a large number of parallel processors arranged in tiers.

The first tier gets raw input information in a pattern like formation and picture in a vector form from an external source, similar to how optic nerves in human vision process information. The notations  $x(n)$  for every  $n$  number of inputs are then used to arithmetically assign these inputs.

After that each input is multiplied by the weights that are compatible with it. Within the computing unit, all of the weighted inputs are added together. If the weighted sum is zero, bias is applied to make the output non-zero.

In the same manner that neurons further away from the optic nerve receive signals from those closest to it, each subsequent tier receives the output of the layer before it rather than the raw input.

Artificial neural networks are noted for their flexibility, which means that they alter as they learn from their initial training, and successive runs supply more information about the world (Fig. 3).



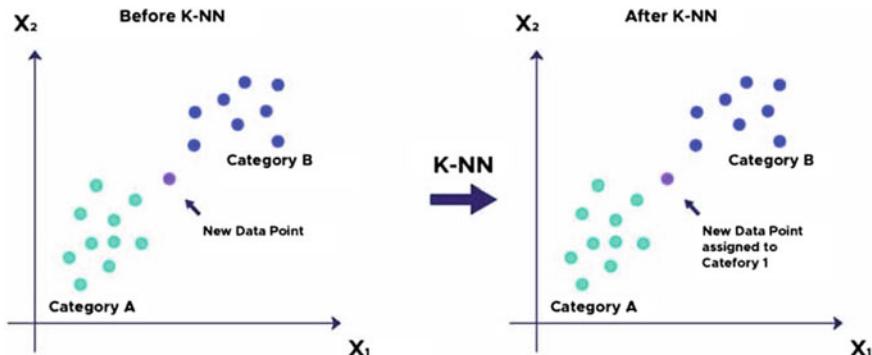
**Fig. 3** [5] ANN gets the input information in a pattern-like formation and picture in a vector form from an external source. The notations  $x(n)$  are then used to mathematically allocate these inputs for each  $n$  number of inputs. *Source* Analytics Vidhya, Available: <https://www.analyticsvidhya.com/blog/2014/10/ann-work-simplified/>

## 2.4 K-Nearest Neighbor

One of the simplest algorithms in supervised learning is k-nearest neighbor. It takes new data/points and compares them with the existing data points of different categories of data and puts new data into a category similar to that data. The KNN algorithm saves the information and categorizes fresh data points based on their similarity to other categories. When new data appear, the KNN method can be used to classify it into current or new categories. KNN can be used for both categorization and regression, but the most common application is classification. In paper [2], KNN is considered as a non-parametric method, and for distribution of the data, it does not make any underlying assumptions. It doesn't learn from the training set; it is known as a lazy learner algorithm; it saves the data and uses it to classify the data later. According to paper [6], the result of the KNN classifier is a class membership value that it belongs to. In KNN, the data are stored in categories, and when new data come, it compares that data to existing categories and puts that data into the category which is closest to it (Fig. 4).

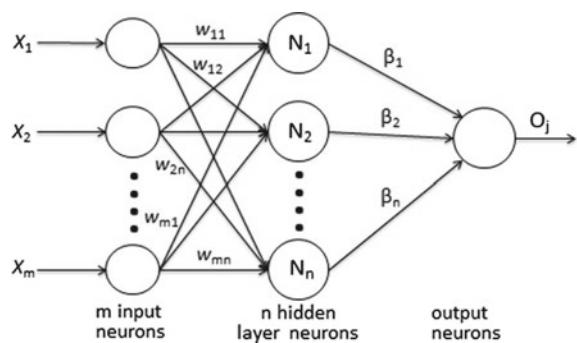
## 2.5 Extreme Learning Machine

The ELM, or extreme learning machine, has developed as a revolutionary single hidden layer feed-forward neural network approach in recent years. Because of its



**Fig. 4** [7] K-nearest neighbor classifier. *Source* yadsmic, Available: <https://www.yadsmic.com/post/k-nearest-neighbors-in-machine-learning>

**Fig. 5** [8] Structure of extreme learning machine. *Source* Research Gate, Available: [https://www.researchgate.net/figure/The-structure-of-extreme-learning-machine\\_fig3\\_265608741](https://www.researchgate.net/figure/The-structure-of-extreme-learning-machine_fig3_265608741)



rapid and efficient learning speed, quick convergence, strong generalization ability, and simple implementation, ELM is employed in batch learning, sequential learning, and incremental learning. Learning algorithms, biases, and hidden node weights that are randomly allocated and do not want to be tweaked, as well as output weights, are all calculated using this method. ELM, on the other hand, is incapable of managing enormous amounts of data with a huge number of dimensions. In the ELM, a higher number of nodes are required than standard algorithms (Fig. 5; Table 1).

### 3 Methodology

According to paper [24], the detection of plant disease involves five major steps, viz., image acquisition, image preprocessing, image segmentation, feature extraction and classification, disease identification. In image processing, acquisition of images is done through a digital camera or scanner; image preprocessing involves image enhancement, image segmentation where the affected and healthy areas are

**Table 1** Details of SVM, random forest, ANN, KNN, and ELM for classification and detection of the plant diseases

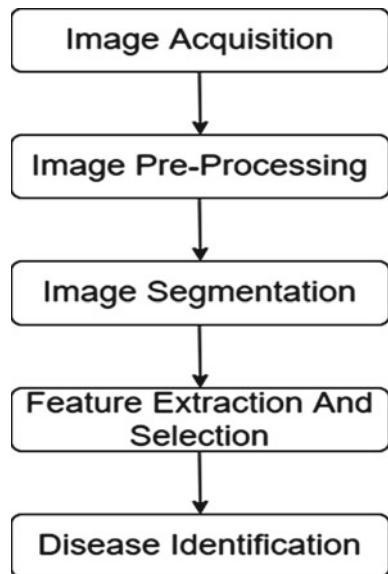
Authors and year	Classification methods	Plant	Disease	Result
Monzurul et al. [9] (2017)	SVM	Potato	Common diseases	95%
Pooja et al. [10] (2017)	SVM	Plants	Common diseases	92.4%
Yin Min Oo et al. [11] (2018)	SVM	Plants	Cercospora leaf spot, bacterial blight, rust leaf disease	Bacterial blight: 96% Powdery mildew: 99% Rust: 100% Average: 98.2%
Sumit et al. [12] (2018)	SVM	Wheat leaf	Common diseases	Good accuracy
Aditi et al. [13] (2021)	SVM	Potato plants leaves	Common potato diseases	95.99%
Appeltans et al. [14] (2021)	SVM	Plants	Leek rust and white tip disease	The overall accuracy of the disease model was 98.14% for rust and 96.74% for white tip disease
Vimal et al. [15] (2021)	SVM	Rice plant	Rice plant diseases	94.65%
Tan et al. [16] (2021)	SVM	Tomato plant	Common diseases	91.4%
Nilay et al. [2] (2020)	Random forest, SVM, KNN, ANN	Plants	Common disease	Random forest: 73.38%
Shima et al. [17] (2018)	Random forests	Papaya leaves	Abnormalities in plants in different environments like in the greenhouse and in the natural environment	Random forests: 70.14%
Kawcher et al. [18] (2019)	KNN, J48-decision tree, logistic regression and Naive Bayes	Rice leaf	Bacterial leaf blight, leaf smut and brown spot disease	Random forest (decision tree algorithm): 97%

(continued)

**Table 1** (continued)

Authors and year	Classification methods	Plant	Disease	Result
B. Luna-Benoso et al. [19] (2020)	KNN, ANN, SVM, Random Forest Naive Bayes	Tomato plants	Common disease	With a 90/10 ratio, random forest has highest accuracy of 0.907
Raj et al. [20] (2021)	SVM, Naïve Bayes, Random forest, Decision tree linear discriminant analysis (LDA), Principal component analysis (PCA), Logistic regression	Rice seed	Fungal blast	The random forest algorithm: 73.12%
Ramesh et al. [21] (2019)	KNN and ANN classification techniques	Rice crops	Blast disease and normal leaf diseases	99–100% accuracy can be achieved for blast disease and normal leaf diseases, respectively, using ANN-based classification mechanism
Prathusha et al. [22] (2019)	KNN and SVM classifier	Wheat, rice, tomato, cotton, maize, etc.	15 Diseases covered in this paper: wheat with a yellow tint rice leaf Aphelenchoides besseyi, tomato spotted leaf Aphelenchoides besseyi, Anthracnose, bacterial blight cotton reniform nematode, Laurel wilt, toxicogenic fungus in maize, powdery mildew in wheat, and other diseases	KNN classifier-95.6%
Paramasivam et al. (2021) [23]	ELM, SVM, linear and polynomial kernels	Plants	Disease detection	The ELM algorithm: 95%

**Fig. 6** Five major steps in the detection of plant disease



segmented; feature extraction defines the area of infection, and classification helps to detect the type of diseases (Fig. 6).

### 3.1 *Image Acquisition*

The ELM, or extreme learning machine, has developed as a revolutionary single hidden layer feed-forward neural network approach in recent years. A key element involved in image acquisition is the basic setup and long-term maintenance of the hardware used to capture the images. The equipment can be anything from a desktop scanner to a huge optical telescope. If the equipment is not correctly set up, then visual reminders can be produced that can obscure the image processing. Wrongly set up equipment also may provide low-quality images that cannot be restored even with large-scale processing. These fundamentals are important to particular areas, such as comparative image processing, which looks for definite characteristics between image sets.

The three basic sensor preparation measures that are used to transform the illumination energy into digital pictures are depicted in the given diagram below. The reasoning for it is simple: The combination of input energy and output energy converts the energy into a voltage-sensitive material that responds to the type of energy detected and electrical power. The sensor(s)' response is represented by the output voltage waveform, and each sensor's response is converted into a digital quantity by digesting it.

### ***3.2 Image Preprocessing***

Information preprocessing or information cleansing could be a crucial step for most developers as they pay a decent quantity of time in information preprocessing before building a model. Some examples of data preprocessing include outlier detection, missing value treatments, and taking away unwanted information.

Similarly, image preprocessing is the term for operations on pictures at the bottom level of abstraction. The results of image analysis might get a lot better by image preprocessing. The goal of preprocessing is to enhance image information by suppressing unnecessary distortions or enhancing image features that are valuable for the “plant disease classification model” process and analysis.

Different methods of image preprocessing enhancements and correction techniques are as follows: Filtering and noise reduction, threshold, edge enhancements, morphology, segmentation, and color space conversions are all examples of illumination, blur, and focus adjustments.

### ***3.3 Image Segmentation***

Image segmentation is the process of breaking down a digital image into subgroups in order to minimize the image’s complexity and make image analysis easier. It has made it easier to identify and categorize different plant diseases. In the majority of studies, image segmentation is completed to separate the leaves from the framework. As a result, the image’s useless information is removed, which improves accuracy.

Generally, the segmentation might be achieved with the assistance of various methodologies like conversion of RGB image to HIS model, k-means clustering. Further, methods like histogram equalization are often performed on images.

### ***3.4 Feature Extraction***

Feature extraction is one of the most important phases in picture classification since it provides features that may be used in the “plant disease classification model” to categorize the kind and level of infection.” Most research is carried out by the color, texture, shape, size, corners, edges, and morphology in plant disease classification. Texture refers to the distribution of colors, their roughness, and the hardness of the image. Sometimes few different methods are used to perform this extraction, as given below:

1. Extracting color features from the image and texture attributes from the GLCM of the picture. 13 qualities of the input photos will be extracted by the GLCM algorithm like IDM, contrast, skewness, correlation, kurtosis,

- energy, smoothness, homogeneity, variance, mean, RMS, standard deviation, and entropy.
2. MATLAB scripts were used to increase the image's size, improve contrast, and convert RGB to gray scale.
  3. For region description, HSV color and texture attributes are used. HSV color characteristics are critical for detecting the visual environment, recognizing objects, and extracting necessary data.
  4. In papers [25–27], the author had analyzed the images from the medical domain. In paper [28] the author focused on text data extracting features based on analysis.

### 3.5 Disease Classification

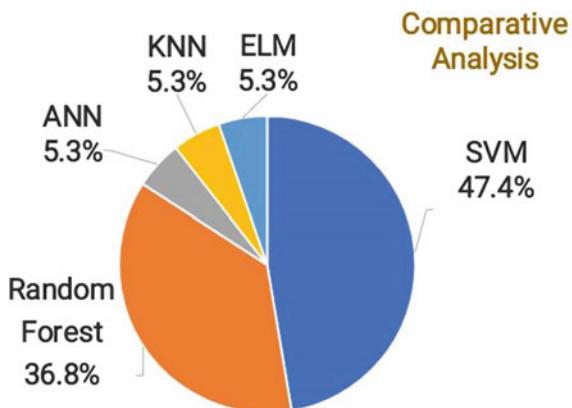
The method in which we categorize the given data into a class is called classification. It can be performed on both supervised learning as well as unsupervised learning. In the classification predictive modeling, we map the input variable to the output variable to check this data falls under which class. It is done by predicting the class of a given data.

## 4 Comparative Analysis

This work evaluates five machine learning algorithms (SVM, random forest, ELM, ANN, and KNN) based on their best performance for specific use cases (detection of the plant disease). It can be observed that the SVM method works well in the case of color and texture analysis/classification on plant leaves dataset, and as a result, it has the highest accuracy in the detection and classification of plant disease, followed by random forest. In paper [29], the author had also presented a survey based on soil nutrients pH, N, P, K values. As a part of analysis of large image or text data set some of the paper mentioned [30] on analysis of various attacks, cloud computing countermeasures, author has done survey on user awareness of cloud security requirements. As a part of secure data analysis [31] author has proposed a lightweight key mechanism for secure communication between cloud and IoT smart devices. Advanced Encryption Standard (AES), symmetric lighweigth cryptographic algorithm is used to provide security attributes.

The pie chart below was formulated after reviewing 19 papers that used different algorithms to detect the different plant diseases. As a result, nine papers, which is equivalent to 47.4%, showed SVM to be the most efficient algorithm for detecting plant diseases. Followed by random forest which had the most accuracy in seven papers, which is equivalent to 36.8% of all the papers. Other algorithms, such as KNN, ANN, and ELM, showed the most efficiency in only one paper each, which is equivalent to 5.3% (Fig. 7).

**Fig. 7** The pie chart to show the analytical data of the respective algorithms



## 5 Objectives

- Investigation of several machine learning techniques for image processing.
- Analyzing existing plant disease categorization cases using image processing and determining which machine learning model will be the most efficient for a certain use case.
- After going through various use cases with different approaches and with different plant species (like tomato, pomegranate, rice, wheat, and potato), we focus to get a generalized resultant algorithm that proves the best selection for plant leaves classification and efficient disease detection.

## 6 Conclusion

A survey of identification plant disease is presented in this paper by using image processing techniques to find efficient algorithms which give the most accurate results. We studied several different cases with various machine learning classifiers, and based on our study, in most of the cases, SVM has the highest accuracy. In some examples of SVM outperforming other classifiers include tomato plant disease detection, in which SVM outperforms ELM and decision trees, and rice plant disease, in which SVM outperformed KNN on colour features alone. Paper [32] analyzes the efficiency of the classification performed using support vector machine, k-nearest neighbor, and decision trees based on the extracted characteristics (shape and color), and the author concluded that SVM performs better compared to KNN and DT for the extracted features.

However, there were instances where different classifiers were more accurate, like, cotton leaf disease classification using a color that had ANN better accuracy than other classifiers like SVM, KNN, and random forest. Random forest is the best for recognizing fungal blast disease in rice seeds. A study on pomegranate plants

also showed that a hybrid framework of KNN and multi-class SVM had the highest accuracy surpassing SVM. However, it has been noticed that with more images and classes, SVM had drastically better results than other algorithms.

## 7 Future Work

Using machine learning methods, numerous research works are covered in this publication to automate the plant disease identification and classification system. Several widely accepted techniques for feature extraction, image acquisition, segmentation, preprocessing, and classification are presented in the survey.

The following points will assist researchers in improving the performance of the ML model shortly:

- (1) IOT enables aerial surveillance of farms and may significantly improve database quality, better data leads to better classification.
- (2) Some papers only consider the classification of one specific plant disease.
- (3) Developing hybrid algorithms employing cuckoo optimization, genetic algorithms, particle swarm, and ant colony optimization with ANN, KNN, and SVM would improve plant disease detection efficiency and allow us to create simple apps or online UI.
- (4) We can classify a large dataset of plant diseases into multiple classes and use deep learning models to improve accuracy.
- (5) We can widen feature extraction of the leaves to have better insights into plant health.

## References

1. Javatpoint: Javatpoint [Online]. Available: <https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm>
2. Nilay Ganatra, A.P.: A multiclass plant leaf disease detection using image processing and machine learning techniques. *Int. J. Emerg. Technol.* **11**(2), 1082–1086 (2020)
3. Wankhede, D., Selvarani, R.: Leaves: India's most famous basil plant leaves quality dataset. IEEE DataPort (2020). <https://doi.org/10.21227/a4f6-4413>
4. Miro medium: miro.medium.com [Online]. Available: [https://miro.medium.com/max/575/1\\*5dq\\_1hnqkboZTcKFfbwO9A.png](https://miro.medium.com/max/575/1*5dq_1hnqkboZTcKFfbwO9A.png)
5. Analytics Vidhya: Analytics Vidhya [Online]. Available: <https://www.analyticsvidhya.com/blog/2014/10/ann-work-simplified/>
6. Swain, Nayak, S.K. Barik, S.S.: A review on plant leaf diseases detection and classification based. *Mukt Shabd J.* **9**(6), 5195–5205 (2020)
7. Yadsmic: Yadsmic [Online]. Available: <https://www.yadsmic.com/post/k-nearest-neighbors-in-machine-learning>
8. Research Gate: Research Gate [Online]. Available: [https://www.researchgate.net/figure/The-structure-of-extreme-learning-machine\\_fig3\\_265608741](https://www.researchgate.net/figure/The-structure-of-extreme-learning-machine_fig3_265608741)

9. Islam, M., Dinh, A., Khan, W., Bhowmik, P.: Detection of potato diseases using image segmentation and multiclass support vector machine. In: IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), pp. 1–4, 04 2017
10. Pooja, V., Das, R., Venkatasubbaiah, K.: Identification of plant leaf diseases using image processing techniques. In: IEEE International Conference on Technological Innovations in ICT For Agriculture and Rural Development, pp. 130–133, 2017
11. Htun, N.C., Oo, Y.M.: Plant leaf disease detection and classification using image processing. *Int. J. Res. Eng.* **5**(9), 516–523 (2018)
12. Sumit Nema, A.D.: Wheat leaf disease detection using machine learning method—a review. *Int. J. Comput. Sci. Mob. Comput.* **7**(5), 124–129 (2018)
13. Aditi Singh, H.K.: Potato plant leaves disease detection and classification using machine. IOP Conf. Ser.: Mater. Sci. Eng. **1022**, 012121 (2021)
14. Simon, A., Pieters, J.G., Mouazen, A.M.: Detection of leek rust and white tip disease under field conditions using **13**, 1341 (2021)
15. Vimal, M.K.P., Shrivastava, K.: Rice plant disease classification using color features: a machine learning paradigm. Springer **103** (2021)
16. Tan Soo Xian, R.N.: Plant Diseases Classification Using Machine Learning, vol. 1962, pp. 15–16. IOP Publishing Ltd. (2021)
17. Ramesh, S., Hebbar, R., Niveditha, M., Pooja, R., Prasad Bhat N., Shashank N., Vinod, P. V.: Plant disease identification using machine learning. *Int. J. Innov. Res. Comput. Commun. Eng.*, pp. 41–45 (2018)
18. Ahmed, K., Shahidi, T.R., Irfanul Alam, S.M., Momen, S.: Rice leaf disease detection using machine learning techniques. In: International Conference on Sustainable Technologies for Industry 4.0 (STI), 24–25 December, Dhaka, pp. 1–5, 2019
19. Luna-Benoso, B., Martinez-Perales, J.C., Cortes-Galicia, J.: Tomato disease detection by means of pattern recognition. *Int. J. Comput. Optimization* **7**(1), 35–45 (2020)
20. Kumar, R., Baloch, G., Pankaj, Baseer, A., Bhatti, J.: Fungal blast disease detection in rice seed using machine learning. (*IJACSA*) *Int. J. Adv. Comput. Sci. Appl.* **12**(2) (2021)
21. Ramesh, D.V.S.: Application of machine learning in detection of blast disease in South Indian rice crops. *J. Phytol.* 2019, **11**(1), 31–37 (2019)
22. Prathusha, P., Srinivasa Murthy, K.E., Srinivas, K.: Plant disease detection using machine learning algorithms. In: International Conference on Computational and Bio Engineering, vol. 16, 2019
23. Alagumariappan, P., Dewan, N.J., Muthukrishnan, G.N., Bojji Raju, B.K., Bilal, R.A.A., Sankaran, V.: Intelligent plant disease identification system using machine learning. **02**, 49 (2020)
24. Shrivastava, G.: Review on emerging trends in detection of plant diseases using image processing with machine learning. *Int. J. Comput. Appl.* **174** (2021)
25. Wankhede, D., Selvarani, R.: Review on deep learning approach for brain tumor glioma analysis. In: International Conference on Convergence of Smart Technologies (IC2ST-2021). <https://doi.org/10.17762/itii.v9i1.144>
26. Singh, S., Bhavsar, M., Mahadeshwar, R., Rathod, S., Wankhede, D.: Predicting IDH1 mutation and 1P19Q CO-deletion status for brain tumor. *Int. J. Adv. Sci. Technol.* **29**(4s), 1196–1204 (2020)
27. Wankhede, D.S., Selvarani, R.: Dynamic architecture based deep learning approach for glioblastoma brain tumor survival prediction. *Neurosci. Inf.* **2**(4), 100062 (2022). ISSN 2772-5286. <https://doi.org/10.1016/j.neuri.2022.100062> (<https://www.sciencedirect.com/science/article/pii/S2772528622000243>)
28. Bhattacharjee, K., ShivaKarthik, S., Mehta, S., Kumar, A., Phatangare, S., Pawar, K., Ukarande, S., Wankhede, D., Verma, D.: Survey and gap analysis of word sense disambiguation approaches on unstructured texts. In: 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 323–327, 2020.
29. Wankhede, D.: Analysis and prediction of soil nutrients pH, N, P, K for crop using machine learning classifier: a review. In: International Conference on Mobile Computing and Sustainable Informatics. ICMCSI 2020. [https://doi.org/10.1007/978-3-030-49795-8\\_10](https://doi.org/10.1007/978-3-030-49795-8_10)

30. Prasad, G.S., Gaikwad, V.S.: A survey on user awareness of cloud security. International Journal of Engineering & Technology 7(2.32), 131–135. <https://doi.org/10.14419/ijet.v7i2.32.15386>
31. Gudapati, S.P., Gaikwad, V.: Light-weight key establishment mechanism for secure communication between IoT devices and cloud. In: Satapathy, S., Bhatia, V., Janakiramaiah, B., Chen, Y.W. (eds.) Intelligent System Design. Advances in Intelligent Systems and Computing, vol. 1171. Springer, Singapore (2021). [https://doi.org/10.1007/978-981-15-5400-1\\_55](https://doi.org/10.1007/978-981-15-5400-1_55)
32. Nandhini, N., Bhavani, R.: Feature extraction for diseased leaf image classification using machine learning. In: 2020 International Conference on Computer Communication and Informatics (ICCCI), pp. 1–4 (2020)

# A Robust and Accurate IoT-Based Fire Alarm System for Residential Use



Tanjil Hossain, Md. Ariful Islam, Alif Bin Rahman Khan,  
and Md. Sadekur Rahman

**Abstract** A fire alarm and smoke detection system detect fire, smoke, gas, temperature, and other factors in the surrounding area. The proposed system is designed to work with common flammable gases found in our environment, such as liquefied petroleum gas (LPG). This system has individual threshold values for each of those elements to determine whether or not any of those elements are raising an alarm. If gas, smoke, fire, and temperature are detected and exceed the threshold value, the system will check 5 times to see if the fire was detected or not. If a fire is detected, a message will be sent to the house's owner to take immediate action; however, if the owner does not respond, another message will be sent to the fire department after some other checking of those elements' current values after sending the message to the owner. If the environment is contaminated with smoke, LPG, or carbon monoxide (CO), the system will alert you to clear the area. The monitor continuously displays CO, smoke, and LPG values in parts per million (ppm) to show the current state of the surrounding environment.

**Keywords** Fire alarm and smoke detection system · Gas detection · Alarm · Warning system · Fire detection · Smoke detection · Gas detection

---

T. Hossain · Md. Ariful Islam (✉) · A. B. R. Khan · Md. Sadekur Rahman  
Department of Computer Engineering, Daffodil International University, Dhaka, Bangladesh  
e-mail: [ariful15-8393@diu.edu.bd](mailto:ariful15-8393@diu.edu.bd)

T. Hossain  
e-mail: [tanjil15-8468@diu.edu.bd](mailto:tanjil15-8468@diu.edu.bd)

A. B. R. Khan  
e-mail: [rahman15-7561@diu.edu.bd](mailto:rahman15-7561@diu.edu.bd)

Md. Sadekur Rahman  
e-mail: [sadekur.cse@daffodilvarsity.edu.bd](mailto:sadekur.cse@daffodilvarsity.edu.bd)

## 1 Introduction

In our daily lives, fire is the most important element. However, fire can sometimes be a curse in our lives, causing any unwanted situation through gas leakage, burning elements, inadvertent use of fire, and so on. Stove fires can also be the cause of such a disaster. According to WHO, fire kills 0.22% of people in Bangladesh [1]. The top three reasons for residential fire are cooking, heating equipment, and electronic malfunction [2]. In the air, there are several mixtures of gases such as nitrogen 78%, oxygen 21%, and carbon-di-oxide. Those are helpful to blaze fire [3]. The most important element for blazing fire is oxygen, heat, and fuel [4]. To avoid those fire accidents, we have developed such a system which may not stop fire but the information given by this system may reduce the losses by the fire. FASDS means fire alarm and smoke detection system which is developed with the help of IOT technology means Internet of things [5].

In this paper, an IOT-based system is introduced which is designed in such a manner so that the system can detect and measure flammable gases, smoke, temperature, light which are measured from the air by different sensible sensors and shows the results by providing information on the system's monitor. If the system can detect smoke, gases, fire, higher temperature, and higher light of the surrounding and cross the limits of the threshold for a certain period of time cycle, the system will send a message to the user. If the user/owner does not response for a given amount of time, a message will be sent to the nearest fire brigade with the address of the house to take action immediately. But if the condition for fire is false within the threshold time cycle, it will not send any alert message to anyone. The proposed system is focused on some facts:

- The system can detect smoke, fire, hazardous gas,
- The system is user friendly,
- The cost of the system must be optimal,
- The system must not generate any false alarms,
- Any occurrence related to fire, the system can notify the owner as well as fire brigade.

The content of the paper is organized as follows: In Sect. 2, related works are discussed for background analysis. In Sect. 3, the proposed method, requirement analysis, data collection, system architecture, and detection methods are discussed. In Sect. 4, the expected outcome is discussed. In Sect. 5, comparison analysis with related work is discussed. In the end in Sect. 6, the future work and the conclusion of the system are described.

## 2 Literature Review

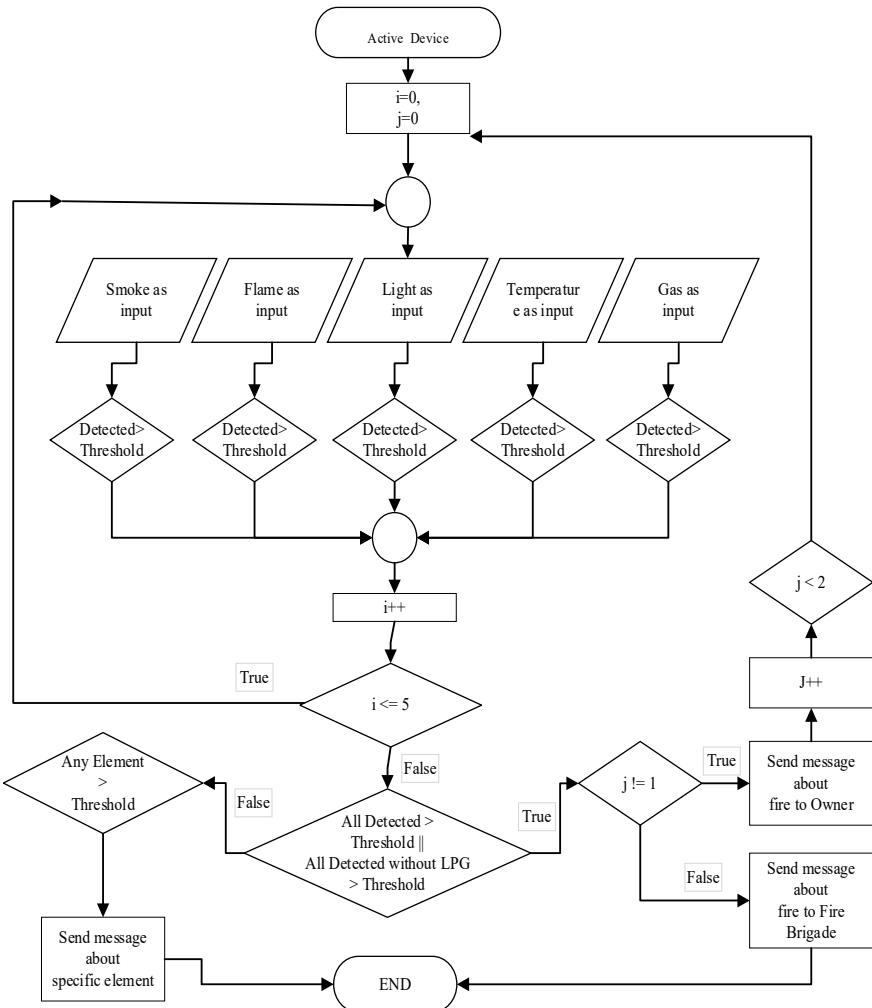
A lot of research work has been done before which are based on fire alarm system, fire detection, smoke detection, gas detection, and many more approaches have been taken but confined into either gas or smoke or fire or one single system. To detect fire correctly, only smoke or flame of fire are not capable to give right information about fire. Because sometimes smoke, flame can be produced from household activities or from our habits of smoking or other smoke able things that we use in our regular life. So, this can be the result of false alarm. From the previous publication, they developed kind of fire alarm system which has higher rate of false alarm. And for this most of the residence fire alarm devices are turned off to avoid this disturbance. The background analysis is as follows:

A system is developed by Ralevski et al. in [6], which is a gas leakage and fire detection system, using time series forecasting. The previous values were used to predict the future value that is based on moving average prediction scheme. Marman et al. in [7] introduced a system which can map flammable and nonflammable fires assembled by system to guide fire fighters and deactivate the local air conditioning system to control the fire. Gottuk et al. in [8] proposed a device that combines smoke and carbon monoxide sensor to reduce false alarm which takes output from sensors that measures the type of developing fire. An updated version of the system is developed by Jamadagni et al. in [9] that can detect both fire and gas at the same time, and if the system can detect anything wrong on air, a notification message is sent to the user. Sweeney et al. in [10] designed a fire alarm system connected to toxic gas monitoring system which can also detect toxic gases and response with a local alarm system and shut down the gas cylinder, and the location is sent to the command center. Jan et al. in [11] designed a device for a steel mill to detect the level of carbon monoxide, and if the level of gas is high, it simply turns off the exhaust fan automatically. Krajci et al. in [12] came up with a system which can detect hazardous gas by a controller that is a software and remote gas sensors, which can check the status of gas sensors. Neuburger et al. in [13] which can provide early warning of halogen gases by frequency measurement and a threshold is fixed to get this early warning. Buck et al. in [14] that can detect smoke and gas together by a predetermined threshold to compare the present smoke and gas levels, and if the level exceeds, an audio and visual notification are given by the system. Winner et al. in [15], the gas sensors were analog that receive analog data and convert it into digital data and send it to the controller and it tests the validity of data and identifies the fault and signals the cause of faults vocally. The controller computes the concentration of combustible gas and automatically finds sensor calibration constants from received data. Consadori et al. in [16] proposed a system that has a self-calibration routine which stored calibration data that is compared to the values measured at normal operation to accurately detect gas concentration levels and if when the sum of previously measured concentration levels exceeds, the predetermined hazardous level an alarm is blown. Fraiwan et al. in [17], a system is designed, and the device is wireless to detect gas for ensuring household safety that has two main modules. The detection and

transmission module sense the change of gas concentration and check if it exceeds a certain predetermined threshold. If any change is detected, it activates an alarm and sends a signal to the receiver. Ishii et al. in [18], a fire alarm is designed which can determine fire by measuring changes in temperature, smoke density, and gas concentration made of fire. King et al. in [19] developed an alarm system which can detect carbon monoxide (CO) and fire in a specific area. If the fire is detected, alarm is activated and emergency services are notified and the fans and vents are cut off as they can increase the power of fire. If the CO is detected only both alarm and notification process works, but in this situation, vents are opened and exhaust fans are activated. In [20], the project was done using GSM, GPS, accelerometer, Arduino, and vibration sensors to detect the vehicle accident detection model to reduce the proportion of human death toll, which occurs due to road accidents. The idea of the development is quite different from the proposed method. The project was created using a fusion of IoT and data mining methods. The idea describes an algorithm that would assist human employers in detecting the regularity of power generation as well as failure or defective regions in solar power systems [21]. The system focuses on fire detection and validation using IoT, WSN, and image processing techniques for detecting fire in city environments using smoke, light, humidity, temperature, and a cloud platform to store data. The system is intended for use in urban areas. As a result, it is unsuitable for residential areas and is more expensive. Allowing for faster fault correction and increased generating station efficiency [22]. All of those systems have different methods to detect fire. Some of them are useable for industries [11, 12, 19] and for kitchen [6, 12], etc. However, most of these systems are expensive, and they are meant to make decisions based on a single piece of data, such as smoke, LPG, fire flame, or CO. Those who have created fire detection systems such as analyzing both CO and fire [19], temperature changes, and smoke and gas [18] are expensive to maintain and give false alarms. Many different forms of activities are always taking place in every home in Bangladesh and around the world for residents. People are aware of fire, but we use it every day to light matches, smoke cigarettes, utilize mosquito coils, cook, and so on. As a result, a standard fire alarm will detect all of these types of fire uses and will sound an alarm. And we are not going to turn it off. As a result, the fundamental disadvantage of those systems is their inability to reduce false alarms and predict genuine fire in a house, a small classroom, or a tiny hospital chamber.

### 3 Methodology

There are a lot of ways to develop the proposed project. The architecture and different types of sensors are implemented to collect data about smoke, temperature, light, CO, etc., from surrounding environment for analyzing them in different manners. But the proposed method includes the simplest architecture to develop the system with low cost and less false alarm from other developed fire alarm system for residence.

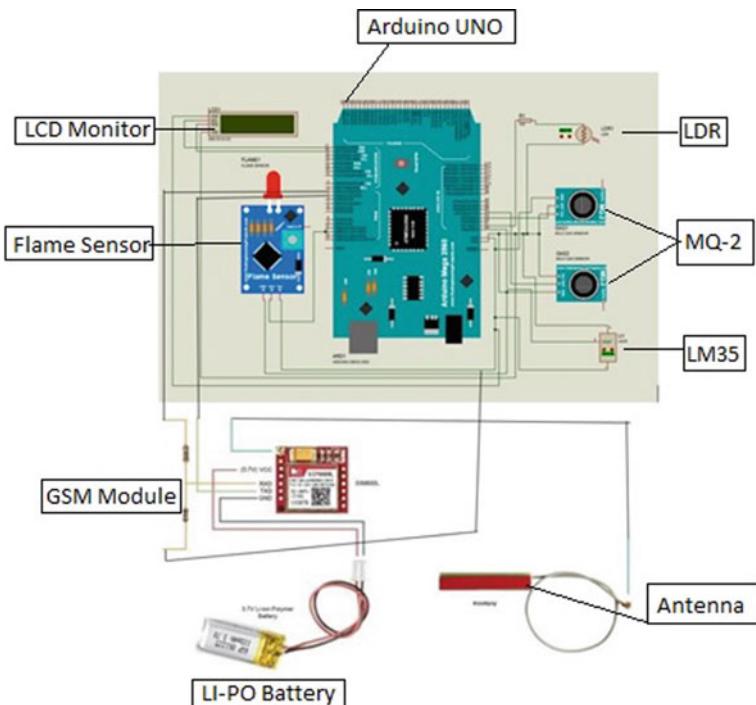


**Fig. 1** Dataflow diagram of the system

Figure 1 shows the dataflow diagram of the proposed system, and Fig. 2 shows the system architecture of the proposed system.

### 3.1 System Architecture

Figure 2 shows the circuit diagram of the project. Here, all the components are connected between them through wires. Different types of sensors are connected, an LCD screen and for sending message GSM module with antenna and battery for



**Fig. 2** Arduino board connected with LCD display, GSM module, and all sensors

power supply. The whole system is set on bread board. Here, Arduino is the main brain of the project. To figure our real fire, all those sensors are continuously sending data to the receiver, and Arduino will take action as it is coded based on the received data.

### 3.2 Proposed Method

Fire alarm system has been using from a long time ago for our safety, but most of the system remain shut down because of too much false alarm. The proposed system can prevent all kind of false alarm. The outcome from the research about the previous fire alarm system is that most of them are using smoke or flame or temperature to make a decision about fire. So, if it detects any trace of smoke or flame or temperature or other things that indicate fire that directly compare it with the threshold value set by the developer. If the presences of those things are higher than threshold value, it will alarm as fire. But in Bangladesh or other country, people normally use coil or chemical smoke to kill mosquito and also habits of smoking. Recently, in coronavirus pandemic situation, many developed countries are taking precautions like using

sanitizer smoke in local market, house, school, college, office, etc. So, if we use that kind of fire alarm system which takes decision basis on smoke, then the system will send continuously false alarm. Same as for flame detected fire alarm system. If a candle is lightened up or cigarette with lighter, it will detect flame and make noise about fire. This is a false alarm. But our proposed system will take a decision about fire when it sense smoke, flame, temperature all at the same time. If all the sensed values are more than threshold level and also to make it more prefects, we implement LDR to measure the changes of light. This system can also detect flame able gas. If only LPG is detected, then the system will measure the LPG level five times in a fix time of interval to ensure that the LPG is still leaking, and then, it will send a message to the owner's mobile about the LPG leaking. In those five times of LPG checking, the system will also check the temperature, flame, smoke, and light level as well, if all of them reached at the threshold level, then it is a sign of fire and it will send an alert message about fire to the owner or if nothing is changed only LPG level is crossed, then it will message about only LPG. Same thing will happen when it will detect smoke. It will measure the smoke level five times with checking of temperature, flame, smoke and light level. If nothing is changed but smoke level is not decreased by time, then we can say that it is not fire but the room air is containing too much level of smoke and CO. Because all types of smoke contain CO which much level of presence in air is dangerous for health. So, it will send an alert message about smoke and CO or if fire is detected, then the message will be a fire alert message. For both smoke and flammable gas, if fire is detected, then at first, it will send message to the owner after that it will check all the sensors value 5 times to ensure that the owner takes necessary steps to put off the fire. If it is not, then the system will send a fire alert message to the fire brigade with the location and the owner's phone number to contact with him/her immediately to ensure about the fire or they can take necessary steps to handle the situation. So, to ensure about real fire, all the element produced by fire must be present. Because real fire dose not contains only smoke or only CO or only temperature or flame. If fire created which is destroyable will have all those things at same time. If individual element is detected at extreme level, then it will only notify about it. This can prevent false alarm when people do other use of fire. The monitor of the system will continuously show the LPG, smoke, and CO level in ppm to see the current status of them in air. So, after get any notification about any of this harmful element of fire, it will be easy for the user of this device to check what its current status on the monitor.

### **3.3 Setup Analysis**

To develop the proposed system, the system requires some instruments. Some sensors to collect data from environment, instrument for sending alert message, etc. How the sensors and other elements are used in our development lifecycle is described below:

- MQ-2: MQ2 is a gas sensor which can detect methane, butane, LPG, and smoke. Two MQ2 sensors are used in this system. One is to detect smoke, CO, and another one is to detect LPG levels. Two same sensors are used to find the levels of smoke and gas individually to get more accurate output. But these two sensors are connected with each other. If one of the sensors became dead for some reason, then one sensor can do other's calculation. But for better performance with accurate data, this system integrated with two MQ2 sensors.
- LM35: LM35 is one of the temperature sensors that exist in the market. The sensor is used to find the reading of temperature around the environment. If the fire is detected, the temperature of the room must be high compare to the normal environment.
- LDR: LDR is a photo resistor which is used to find the value of light around the environment. If the fire is detected, the brightness or the light of the room must be increased.
- Flame Sensor: Flame sensor is kind of sensor that is used to measure the existence of fire around the environment. If any flame is detected, it will send a signal.
- LCD monitor with I2C module: LCD monitor is used to get the updated output continuously from the sensors. The levels of LPG, CO, smoke of the current environment are displayed in the monitor in ppm.
- Resistors: Resistors are implemented in different places to reduce and control the flow of current and signal levels. 1 and 10 k resistors are used by the requirement of sensors and other instruments in this system.
- GSM Module: GSM module is a kind of circuit to make communication between a mobile and a computing machine. This circuit is used to send message to the user in emergency situations.
- Arduino Mega R3 2560: Arduino mega is a board where the system is uploaded and run the whole circuit. Arduino can take any action based on the sensors value and logic that is implemented for every situation.
- Breadboard: Breadboard is a construction base to implement the whole circuit that is required for this system.
- 3.7v Li-Po Battery: Lithium-ion polymer battery is used for GSM module as the module required 3.7-4.4v power supply.
- Wire: All the sensors, modules, and other instruments are connected with wire to merge the whole system together.

### 3.4 Data Collection

In standard residence environment from Bangladesh perspective, the level of smoke, LPG, and carbon monoxide are 170, 0, and 5 ppm. The standard temperature of house is 23 °C. The standard reading of light of a house is 120–140. No fire can be detected in standard temperature. That study is all about the reading of standard environment till now. To justify the whole system, a test environment is created with smoke, CO, LPG, and real fire, and threshold value is assigned. For the proposed

system, the level of smoke, LPG, and carbon monoxide are 180, 50, and 1000 ppm. The threshold value of temperature of house is 28 °C. The reading of light of the house is 150. This threshold values are assigned after a short test by test elements to detect all of those elements that are active when fire or gases are detected in our environment. Those threshold values may vary for other cities or countries.

### ***3.5 Detection Methods***

The proposed system will take a decision about fire when it sense smoke, flame, temperature all the sensed value more than threshold level and also to make it more prefect we implement LDR to measure the changes of light. If only LPG is detected, then the system will measure the LPG level five times in a fixed time of interval to ensure that the LPG is still leaking and then, it will send a message to the owner's mobile about the LPG leaking as a warning message of LPG leakage and fire. In those five times of LPG checking, the system will also check the temperature, flame, smoke, and light level, if all of them reached at the threshold level, then it is a sign of fire. Same type of method is also applicable for detection of fire. If any of fire sign is detected at threshold level, then it will check it for five times to ensure it is really a fire. If none of them is identified at threshold level, then it will send a warning about the present element. And if, after all the measurement, the system detects any real fire, then it will send message to the user and after the ignorance or absent of user the system again checks the current situation for five times. If all the values are still at threshold level, then it will send message to the nearest fire brigade service with owner information.

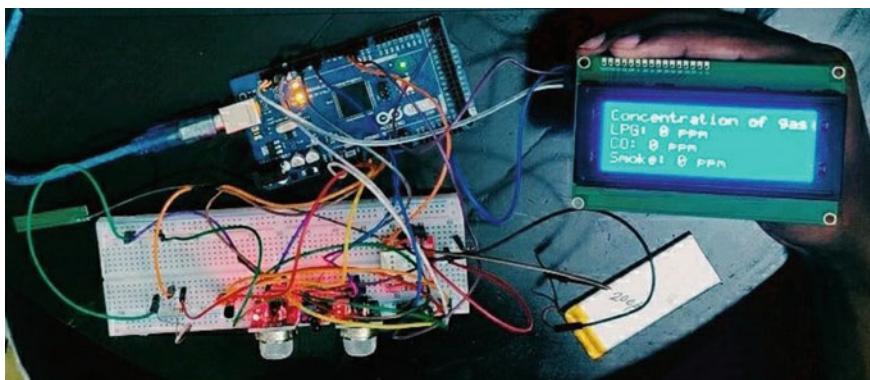
### ***3.6 Warning Methods***

If fire is detected, then at first, it will send message to the owner after that it will check all the sensors value 5 times to ensure that the owner takes necessary steps to put off the fire. If it is not, then the system will send a fire alert message to the fire brigade with the location and the owner's phone number to contact with him/her to ensure about the fire or they can take necessary steps to handle the situation. For individual harmful elements of fire like LPG, smoke or CO, it will notify to the owner. The monitor of the system will continuously show the LPG, smoke, and CO level in ppm to see the current status of them in air.

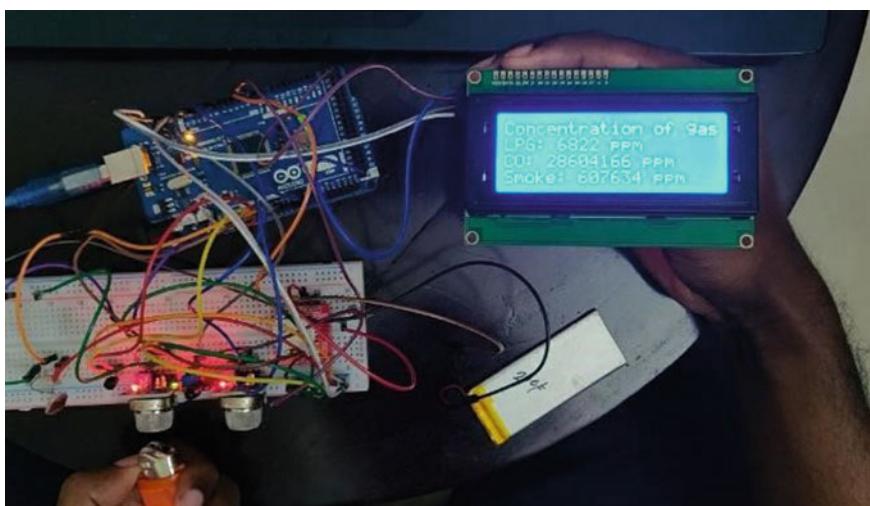
## 4 Result Analysis

Figure 3 is showing the whole system when it is on. Here, after switch on the device at first, it was showing LPG, CO, and smoke value are 0 ppm. To ignore the natural presence of those elements at home environment, this system has set the initial value as 0.

Figure 4 shows LPG, smoke, and CO level in ppm at the system's monitor when we press the lighter. Lighter contains flammable gas. The system is sensing LPG reading but other element of fire like flame and temperature is missing. The reading is much more than the threshold value.



**Fig. 3** Full setup of the system with Arduino, breadboard, sensors, and LCD monitor



**Fig. 4** Experimental setup and result

**Fig. 5** An alert message to the owner

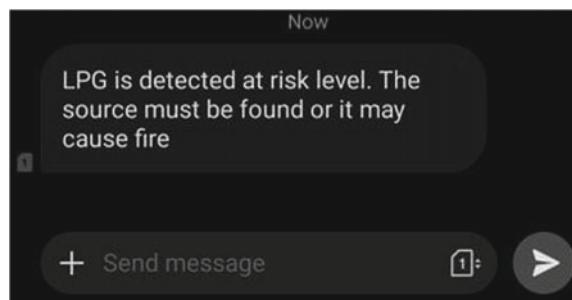


Figure 5 states that the LPG, smoke, and CO level were much more than the threshold value. So, it will send an alert message to the owner because the system detects the presence of LPG at high level. So, this is a clear case of gas leakage not a fire alert.

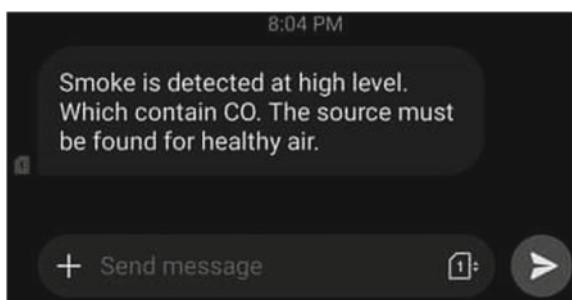
Here, in Fig. 6, the test includes smoke using coil or can be done with an element that can make only smoke. The recent smoke level is much higher than the regular smoke level in air. This had already crossed the threshold value for smoke.

In Fig. 7, we saw the previous system detects the presence of smoke at high level. This had already crossed the threshold value. So, the owner will receive an alert message about it. Because that much level of smoke contains much level of CO which makes our air enough polluted to breath. So, the alert message will also notify the user about CO as well as smoke.

**Fig. 6** Output on the monitor



**Fig. 7** Smoke detection message alert





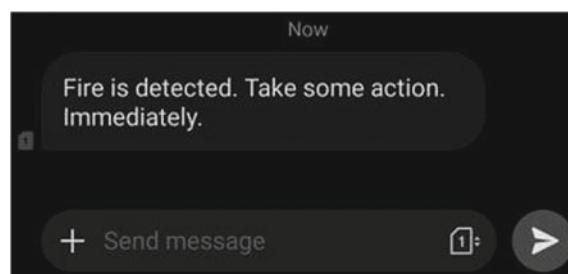
**Fig. 8** Indication of fire

In Fig. 8, the test was made with real fire. To make real fire, this system needs all the presence of light, smoke, CO, flame, and temperature. SO, to make this kind of environment, we set fire on leaves of tree. LDR, MQ2, temperature, flame all the sensors received value crossed the threshold value which is indicating that it is caused by fire.

After crossing all the sensors value at threshold level, it is definitely an indication of fire. So, the system will send a message immediately to the owner about fire like Fig. 9 is showing.

After sending message to the owner about real fire, the system will check 5 times to ensure that the user takes necessary steps to prevent the fire. If not, by the same way, it will send message to the fire brigade with all the necessary information about the owner. This system has been tested in a practical situation to ensure that it can behave as expected, and it is fully functional.

**Fig. 9** Detection of fire



**Table 1** Comparative analysis between the proposed work and other related works

Work done	Smoke	Fire	Gas	Light	Temperature	Cost
This work	Yes	Yes	Yes	Yes	Yes	Low
Ralevski et al.	No	Yes	Yes	No	No	Low
Marman et al.	Yes	Yes	No	No	No	Medium
Sweeney et al.	No	No	Yes	No	No	Medium
Gottuk et al.	Yes	No	Yes	No	No	Low

## 5 Comparative Discursion of Results

Comparative analysis is discussed in Table 1.

## 6 Conclusion and Future Work

The proposed technique introduces an unprecedented technique for detecting fire, smoke, and several toxic gases created from fire. The system can detect fire, smoke, and gas from different sensors that we used and measure the value to decide the situation of the surroundings. To evaluate the method, several tests have been done over the whole system by fire, smoke, and carbon monoxide gas to detect correctly and provide the actual value on the LCD monitor. To detect each of the elements, different sensors are used to measure values. The actual value is compared with the threshold value to check the status of the surroundings and provide us the visibility and the notifications accordingly. The test result shows the efficient outcome of our system. It is efficient for both performance and cost. Still, the system has some limitations.

More future works are yet to implement to develop the system more efficient to improve performance for detection. A mobile application has been developed for that device to remotely monitor the environment. So, for this, the device can be monitored remotely. Another feature is to add like releasing fire sprinkler basis on the fire situation. One more important feature must be added which is electricity power disconnection. If a real fire is detected, then the main electric station of a residence area will be disconnected. Addition of those features can help us to process and observe data more efficiently.

## References

1. Death rate in Bangladesh by fire: <https://www.worldlifeexpectancy.com/bangladesh-fires> [Last accessed 24 Sept 2021]

2. Reason of residential fire: <https://www.thezebra.com/resources/research/house-fire-statistics/> [Last accessed 24 Sept 2021]
3. Gases on the Air: <https://climate.nasa.gov/news/2491/10-interesting-things-about-air/> [last accessed 05 Oct 2021]
4. Elements of Fire: [https://www.sc.edu/ehs/training/Fire/01\\_triangle.htm](https://www.sc.edu/ehs/training/Fire/01_triangle.htm) [last accessed 5 Oct 2021]
5. Internet of things detail: <https://internetofthingsagenda.techtarget.com/definition/Internetof-Things-IoT> [Last accessed 24 Sept 2021]
6. Ralevski, M., Stojkoska, B.R.: IoT based system for detection of gas leakage and house fire in smart kitchen environments. In: 2019 27th Telecommunications Forum (TELFOR), 2019, November, pp. 1–4. IEEE.
7. Marman, D.H., Peltier, M.A., Wong, J.Y., Marman, D.H., Peltier, M.A., Wong, J.Y.: Fire and smoke detection and control system. U.S. Patent 5,945,924 (1999)
8. Gottuk, D.T., Peatross, M.J., Roby, R.J., Beyler, C.L.: Advanced fire detection using multi-signature alarm algorithms. *Fire Saf. J.* **37**(4), 381–394 (2002)
9. Jamadagni, S., Sankpal, P., Patil, S., Chougule, N., Gurav, S.: Gas Leakage and Fire Detection using Raspberry Pi. In: 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) 2019, March, pp. 495–497. IEEE
10. Sweeney, J.: Integration of toxic gas monitoring systems into building fire alarm systems at Harvard University. In: 2010 18th Biennial University/Government/Industry Micro/Nano Symposium 2010, June, pp. 1–6. IEEE
11. Jan, M.F., Habib, Q., Irfan, M., Murad, M., Yahya, K.M., Hassan, G.M.: Carbon monoxide detection and autonomous countermeasure system for a steel mill using wireless sensor and actuator network. In: 2010 6th International Conference on Emerging Technologies (ICET) 2010, October, pp. 405–409. IEEE
12. Krajci, J., Neodym Systems Inc.: Gas detection system and method. U.S. Patent 6,182,497 (2001)
13. Neuburger, G.G., Telcordia Technologies Inc, 1991. Early warning reactive gas detection system. U.S. Patent 5,065,140.
14. Buck, R.H., Cowan, K.G., Doughty, J.P., Marshall, S.E., BDC ELECTRONICS: Combined smoke and gas detection apparatus. U.S. Patent 4,688,021 (1987)
15. Winner, J.K., Bendix Corp.: Combustible gas detection system. U.S. Patent 4,464,653 (1984)
16. Consadori, F., Field, D.G., Banta, K.D., Atwood Industries Inc.: Method and system for gas detection. U.S. Patent 5,526,280 (1996)
17. Fraiwan, L., Lweesy, K., Bani-Salma, A., Mani, N.: A wireless home safety gas leakage detection system. In: 2011 1st Middle East Conference on Biomedical Engineering, 2011, February, pp. 11–14. IEEE
18. Ishii, H. and Ono, T., Hochiki Corp.: Fire alarm system, sensor and method. U.S. Patent 4,871,999 (1989)
19. King, S., Annex Security and Technical Services: Smart fire alarm and gas detection system. U.S. Patent 7,005,994 (2006)
20. Patil, P.J., Zalke, R.V., Tumasare, K.R., Shiwankar, B.A., Singh, S.R., Sakhare, S.: IoT protocol for accident spotting with medical facility. *J. Artif. Intell. 3*(02), 140–150 (2021)
21. Shakya, S.: A self monitoring and analyzing system for solar power station using IoT and data mining algorithms. *J. Soft Comput. Paradigm 3*(2), 96–109 (2021)
22. Sungheetha, A., Sharma, R.: Real time monitoring and fire detection using internet of things and cloud based drones. *J. Soft Comput. Paradigm (JSCP) 2*(03), 168–174 (2020)

# Kernel Feature Variant-Based Gaussian Process Regression for Prediction of Snail Rings



M. Shyamala Devi, N. Abhishek Rao, S. G. Kushal Kumar, G. Dheeraj, and K. Govinda

**Abstract** Estimating the snail age is a tedious process in which the snail shell has to eradicate from the cone where the rings are counted by using the microscope. The above methodology is a manual technique, which involves more time for complete age calculation. Snail age prediction remains a challenging task for the marine industry. Machine learning can be used for this process for the age prediction of snail using regression algorithms. With this overview, the snail age dataset is used for predicting the rings of the snail. The snail age dataset is preprocessed by solving the missing values and categorical variables. The dataset with all seven features is applied to Gaussian process regressor with various kernels like Dot Product, White Kernel, Pairwise Kernel, Rational Quadratic, ExpSine Squared, Matern, RBF, Constant Kernel, and Exponentiation for analyzing the performance of the regressor toward predicting the snail rings before and after feature scaling. The top four features extracted from the ensemble regressors like AdaBoost, gradient boost, extra trees, and random forest regressors are fitted with Gaussian process regressor for abovementioned kernels, and the performance is analyzed with EVS, MAE, MSE, and RScore. The scripting is written in Python and implemented with Spyder in Anaconda Navigator IDE, and the experimental results show that the Constant Kernel with Gaussian regressor applied to the random forest feature importance snail age dataset projects the RScore of 0.9687 when compared to other ensemble regressors.

**Keywords** Machine learning · Ensemble regressor · MAE · MSE · EVS · RScore

## 1 Introduction

By assessing RNA sequence gene expression patterns and DNA methylation patterns, genetic markers are investigated to predict the biological age of humans. In these

---

M. Shyamala Devi (✉) · N. Abhishek Rao · S. G. Kushal Kumar · G. Dheeraj · K. Govinda  
Computer Science & Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu, India  
e-mail: [shyamaladev@gmail.com](mailto:shyamaladev@gmail.com)

research findings, various machine learning algorithms are used to define appropriate biomarkers for age prediction. The RNA sequence gene expression dataset was created using human fibroblast cell lines. The skin elastic fibers cells protect the skin from age-related damage. These cells also contain epigenetic and heritable traits changes that are time of life. DNA methylation comes with age in each individual. This understanding is applied to the selection of useful biomarkers from a DNA methylation dataset. As genetic polymorphisms, the epigenetic sites with the highest similarity to age are chosen. In both of these studies, machine learning algorithms are used to evaluate different biological markers in order to create an age forecasting models. Concerns about imprecision or favoritism related to the current methods used to acquire abalone age and economic expansion data have prompted evaluation of multiple data collection techniques to implement effective management of abalone fishing industry. The paper is organized in such a way that literature review is discussed in Sect. 2 followed by the paper contributions in Sect. 3. The implementation setup and results are discussed in Sect. 4 followed by conclusion in Sect. 5.

## 2 Literature Review

Many research findings of molluscan age and economic expansion have used stable oxygen isotopes, but this technique has yet to be implemented to haliotids. The procedure is described here for the blacklip abalone *Haliotis rubra* in southeast Tasmania. The annual temperature cycle was reflected in the steady oxygen carbon isotopes of calcium carbonate sequential specimens taken at approximately 1–4 mm intervals along the direction of propagation of the shell [1]. We suggest a new classification scheme given its low parity-check codes, a powerful class of bitwise block codes. The key premise is to use the Boolean algebra characteristics of the codes to start generating code words for scripting matrix and describe two decryption methodologies that allow to identify and retrieve potential errors or refuses by dichotomize [2]. Transitions in organism's ranges are among the most commonly mentioned and globally pervasive impacts of climate change, with rates of motion especially high in the sea. The emergence of numerous range-expanding species can pose serious challenges to natural capital executives; some species threaten ecological system, while others present social and/or business potential. An immediate consequence of which species may be expanding their intervals can help managers plan investment plans in impact assessment, monitoring, or potential management interference [3].

The researchers use space station remote location which detected extreme sea surface temperature and the 1990s simulated marine climate to improve interactively down sampling ocean future climate estimates of sea surface temperature outliers in the Tasman Sea off southeastern Australia. This is accomplished using a Bayesian hierarchical model in which the parameters of an extreme value distribution are modeled using regression analysis on the key marine climatic parameters [4]. Globally, this paper investigates studies explaining range shift patterns in marine species

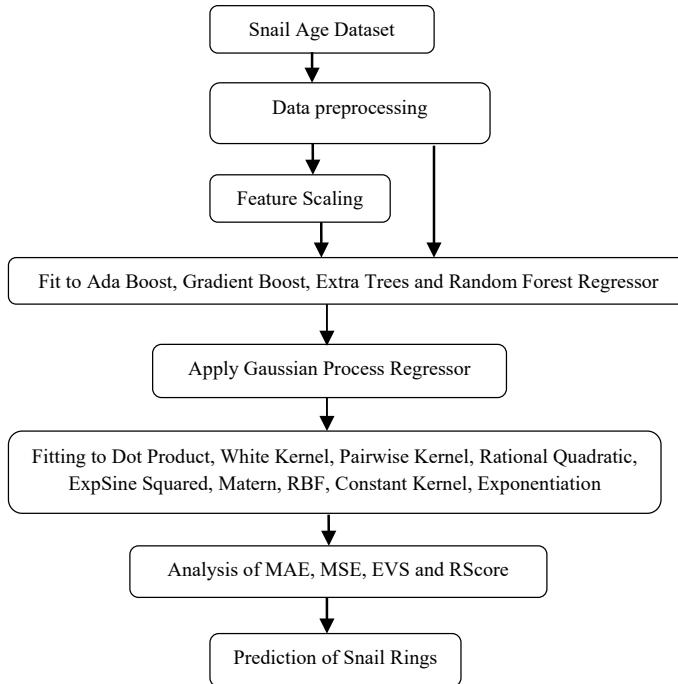
using rigorous selection process, including the isolation of individual studies as well as those who deduced range shifts through changing enormous amount. The length in kilometers decided to move per year for each organisms meeting these requirements was estimated, and life history attributes such as mobility, habitat, dissemination potential, and food web level were documented to see if they were connected to the shift patterns. We also looked into the possibility of a link between modifying ocean temperatures and scope shift patterns [5].

Existing seafood original source studies have limitations that can be overcome by using supplementary or self-contained nuclear methods. The purpose of this research is to determine the production process and geographical location of Asian freshwater fish utilizing carbon isotope analysis and X-ray fluorescence. Three statistical techniques were used to analyze the data: univariate and bivariate analysis, random forest, and LDA [6]. It is difficult to determine the life expectancy of aquatic creatures even though they live deep in the ocean, in the animal world, and not in a laboratory. To address this issue, varieties of implicit efforts have been made. This report outlines these methodologies and seeks to explain why certain non-mammalian marine life can live so long when tried to compare to land-based or aquatic mammalian species [7]. This study assessed the safety issues that arise both in the event of seismic events and in primary immune response, where quake danger assumption must be treated stochastically. This review presents two measurable techniques for predicting likely seismically damage to low and mid-rise built-up significant frameworks [8]. The multi-hazard system, which is divided into three parts: danger showcase, underpinning delicacy investigation, and damage probability computation, evaluates the injury dangers of an elevated surface revealed to unmanageable shaking and wind dangers differently and instantly [9]. This study explained the significant engineering problems especially in the post case of emergencies construction inspection procedures, and suggestions are provided based on extensive particular experience from past earthquakes [10].

### 3 Our Contributions

The overall architecture of the work is shown in Fig. 1. The following contributions are provided in this work.

- Firstly, the dataset is preprocessed to handle missing values, categorical value, and scaling.
- Secondly, the raw dataset is applied with the Gaussian process regressor for different kernels to Dot Product, White Kernel, Pairwise Kernel, Rational Quadratic, ExpSine Squared, Matern, RBF, Constant Kernel, and Exponentiation before and after feature scaling, and the performance is analyzed.
- Thirdly, dataset is fitted to AdaBoost regressor, gradient boost regressor, extra tree regressor, and random forest regressor to find the topmost four features from the dataset with and without the feature scaling.



**Fig. 1** Workflow methodology

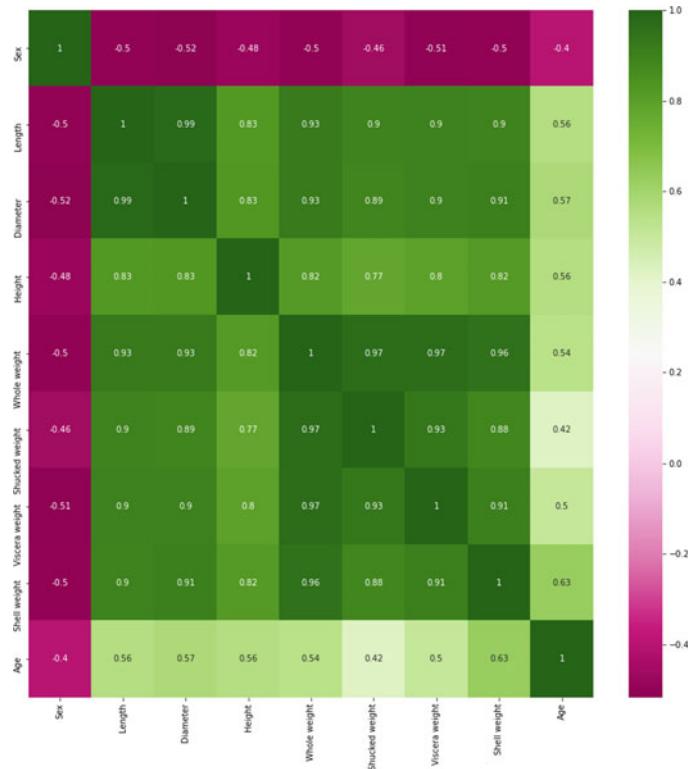
- Thirdly, the top features extracted from the above ensemble regressors are fitted with Gaussian process regressor for different kernels to Dot Product, White Kernel, Pairwise Kernel, Rational Quadratic, ExpSine Squared, Matern, RBF, Constant Kernel, and Exponentiation, and the performance is analyzed.
- The performance of snail rings prediction is done with EVS, MAE, MSE, and RScore.

## 4 Implementation Setup and Results

The snail age dataset from the KAGGLE is used for predicting snail rings, and the correlation is shown in Fig. 2 with dataset information which is shown in Fig. 3.

The raw dataset is applied with the Gaussian process regressor for different kernels to Dot Product, White Kernel, Pairwise Kernel, Rational Quadratic, ExpSine Squared, Matern, RBF, Constant Kernel, and Exponentiation before and after feature scaling, and the performance is analyzed and is given in Tables 1 and 2.

The top features extracted from the AdaBoost regressor are fitted with Gaussian process regressor for different kernels Dot Product, White Kernel, Pairwise Kernel,

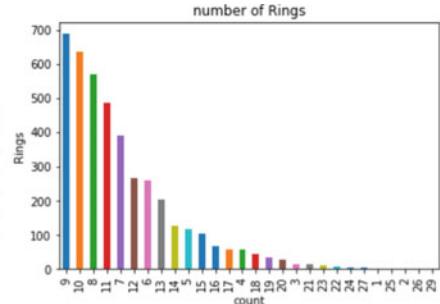


**Fig. 2** Correlation set of snail age dataset information

```

RangeIndex: 4177 entries, 0 to 4176
Data columns (total 9 columns):
 #   Column      Non-Null Count Dtype  
 --- 
 0   Sex          4177 non-null   int64   
 1   Length       4177 non-null   float64 
 2   Diameter     4177 non-null   float64 
 3   Height       4177 non-null   float64 
 4   Whole weight 4177 non-null   float64 
 5   Shucked weight 4177 non-null   float64 
 6   Viscera weight 4177 non-null   float64 
 7   Shell weight  4177 non-null   float64 
 8   Age          4177 non-null   int64  
dtypes: float64(7), int64(2)
memory usage: 293.8 KB

```



**Fig. 3** Snail age dataset information

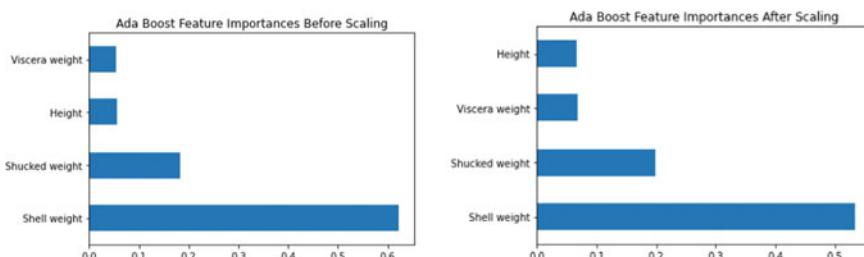
Rational Quadratic, ExpSine Squared, Matern, RBF, Constant Kernel, and Exponentiation before and after feature scaling, and the performance is analyzed and is shown in Fig. 4 and Tables 3 and 4.

**Table 1** Indices score of Gaussian process regressor with raw dataset before scaling

Gaussian kernels	EVS	MAE	MSE	RScore
Gaussian-Dot Product	0.8867	0.8777	0.8967	0.8867
Gaussian-White Kernel	0.8367	0.8267	0.8247	0.8357
Gaussian-Pairwise Kernel	0.8167	0.8117	0.8117	0.8147
Gaussian-Rational Quadratic	0.7332	0.7232	0.7332	0.7332
Gaussian-ExpSine Squared	0.7433	0.7432	0.7432	0.7533
Gaussian-Matern	0.7923	0.7943	0.7943	0.7935
Gaussian-RBF	0.7277	0.7337	0.7246	0.7567
Gaussian-Constant Kernel	0.8987	0.8987	0.8977	0.8987
Gaussian-Exponentiation	0.8332	0.8232	0.8332	0.8332

**Table 2** Indices score of Gaussian process regressor with raw dataset after scaling

Gaussian kernels	EVS	MAE	MSE	RScore
Gaussian-Dot Product	0.8967	0.8977	0.8997	0.8967
Gaussian-White Kernel	0.8667	0.8667	0.8647	0.8657
Gaussian-Pairwise Kernel	0.8767	0.8717	0.8717	0.8747
Gaussian-Rational Quadratic	0.8332	0.8232	0.8332	0.8332
Gaussian-ExpSine Squared	0.8433	0.8432	0.8432	0.8533
Gaussian-Matern	0.8923	0.8933	0.8993	0.8995
Gaussian-RBF	0.8267	0.8397	0.8356	0.8667
Gaussian-Constant Kernel	0.9387	0.9387	0.9377	0.9387
Gaussian-Exponentiation	0.8832	0.8832	0.8832	0.8832

**Fig. 4** Feature importance of AdaBoost regressor before and after scaling

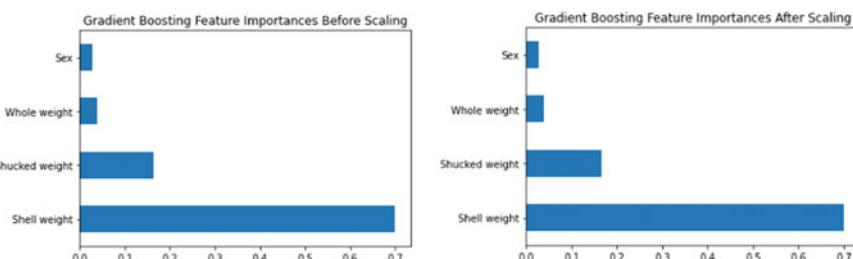
The top features from the gradient boost regressor are fitted with Gaussian regressor for different kernels to Dot Product, White Kernel, Pairwise Kernel, Rational Quadratic, ExpSine Squared, Matern, RBF, Constant Kernel, and Exponentiation before and after scaling, and the performance is analyzed and is shown in Fig. 5 and Tables 5 and 6.

**Table 3** Indices score of Gaussian process regressor with AdaBoost dataset before scaling

Gaussian kernels	EVS	MAE	MSE	RScore
Gaussian-Dot Product	0.8767	0.8777	0.8767	0.8767
Gaussian-White Kernel	0.8467	0.8267	0.8247	0.8257
Gaussian-Pairwise Kernel	0.8267	0.8217	0.8217	0.8247
Gaussian-Rational Quadratic	0.7432	0.7332	0.7332	0.7332
Gaussian-ExpSine Squared	0.7533	0.7432	0.7432	0.7433
Gaussian-Matern	0.7723	0.7643	0.7643	0.7635
Gaussian-RBF	0.7277	0.7337	0.7346	0.7367
Gaussian-Constant Kernel	0.8887	0.8887	0.8877	0.8887
Gaussian-Exponentiation	0.8232	0.8232	0.8232	0.8232

**Table 4** Indices score of Gaussian process regressor with AdaBoost dataset after scaling

Gaussian kernels	EVS	MAE	MSE	RScore
Gaussian-Dot Product	0.8867	0.8877	0.8897	0.8867
Gaussian-White Kernel	0.8567	0.8567	0.8547	0.8557
Gaussian-Pairwise Kernel	0.8967	0.8917	0.8917	0.8947
Gaussian-Rational Quadratic	0.8432	0.8432	0.8432	0.8432
Gaussian-ExpSine Squared	0.8533	0.8532	0.8532	0.8533
Gaussian-Matern	0.8823	0.8833	0.8993	0.8995
Gaussian-RBF	0.8367	0.8397	0.8356	0.8667
Gaussian-Constant Kernel	0.9487	0.9487	0.9377	0.9387
Gaussian-Exponentiation	0.8932	0.8932	0.8832	0.8832

**Fig. 5** Feature importance of gradient boost regressor before and after scaling

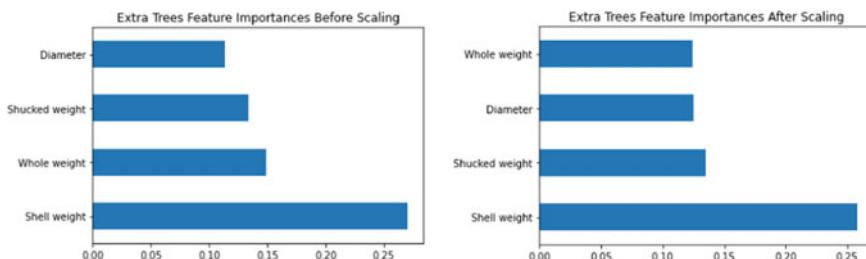
The top features from extra trees are fitted with Gaussian regressor for various kernels before and after scaling, and the performance is analyzed and is shown in Fig. 6 and Tables 7 and 8.

**Table 5** Indices score of Gaussian regressor with gradient boost dataset before scaling

Gaussian kernels	EVS	MAE	MSE	RScore
Gaussian-Dot Product	0.8267	0.8277	0.8267	0.8227
Gaussian-White Kernel	0.8367	0.8367	0.8347	0.8357
Gaussian-Pairwise Kernel	0.8167	0.8117	0.8117	0.8247
Gaussian-Rational Quadratic	0.7232	0.7232	0.7232	0.7222
Gaussian-ExpSine Squared	0.7333	0.7332	0.7332	0.7333
Gaussian-Matern	0.7223	0.7243	0.7243	0.7235
Gaussian-RBF	0.7177	0.7137	0.7146	0.7167
Gaussian-Constant Kernel	0.8287	0.8287	0.8277	0.8287
Gaussian-Exponentiation	0.8032	0.8032	0.8032	0.8132

**Table 6** Indices score of Gaussian regressor with gradient boost dataset after scaling

Gaussian Kernels	EVS	MAE	MSE	RScore
Gaussian-Dot Product	0.8967	0.8977	0.8997	0.8967
Gaussian-White Kernel	0.8867	0.8867	0.8847	0.8857
Gaussian-Pairwise Kernel	0.8967	0.8917	0.8917	0.8947
Gaussian-Rational Quadratic	0.8632	0.8632	0.8632	0.8632
Gaussian-ExpSine Squared	0.8733	0.8732	0.8732	0.8733
Gaussian-Matern	0.8923	0.8933	0.8993	0.8995
Gaussian-RBF	0.8867	0.8897	0.8856	0.8867
Gaussian-Constant Kernel	0.9587	0.9587	0.9577	0.9587
Gaussian-Exponentiation	0.8632	0.8632	0.8632	0.8632

**Fig. 6** Feature importance of extra trees regressor before and after scaling

The top features from the random forest are fitted with Gaussian regressor for various kernels before and after scaling, and the performance is analyzed and is shown in Fig. 7 and Tables 9 and 10.

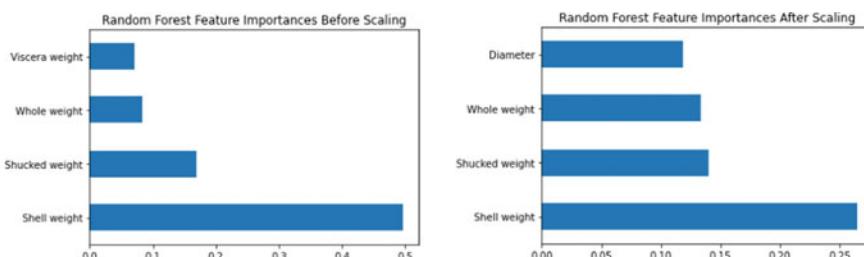
The feature importance probability values of the snail age dataset before and after feature scaling are given in Tables 11 and 12.

**Table 7** Indices score of Gaussian regressor with extra trees dataset before scaling

Gaussian kernels	EVS	MAE	MSE	RScore
Gaussian-Dot Product	0.8167	0.8277	0.8267	0.8227
Gaussian-White Kernel	0.8267	0.8367	0.8347	0.8357
Gaussian-Pairwise Kernel	0.8067	0.8117	0.8117	0.8247
Gaussian-Rational Quadratic	0.7632	0.7232	0.7232	0.7222
Gaussian-ExpSine Squared	0.7533	0.7332	0.7332	0.7333
Gaussian-Matern	0.7723	0.7243	0.7243	0.7235
Gaussian-RBF	0.7277	0.7137	0.7146	0.7167
Gaussian-Constant Kernel	0.8387	0.8287	0.8277	0.8287
Gaussian-Exponentiation	0.8432	0.8032	0.8032	0.8132

**Table 8** Indices score of Gaussian regressor with extra trees dataset after scaling

Gaussian kernels	EVS	MAE	MSE	RScore
Gaussian-Dot Product	0.8767	0.8777	0.8777	0.8767
Gaussian-White Kernel	0.8567	0.8567	0.8587	0.8557
Gaussian-Pairwise Kernel	0.8867	0.8817	0.8817	0.8847
Gaussian-Rational Quadratic	0.8732	0.8732	0.8732	0.8732
Gaussian-ExpSine Squared	0.8833	0.8832	0.8832	0.8833
Gaussian-Matern	0.8423	0.8433	0.8493	0.8495
Gaussian-RBF	0.8967	0.8997	0.8956	0.8967
Gaussian-Constant Kernel	0.9687	0.9687	0.9677	0.9687
Gaussian-Exponentiation	0.8532	0.8532	0.8532	0.8532

**Fig. 7** Feature importance of random forest regressor before and after scaling

## 5 Conclusion

This paper explores the performance of snail age dataset in predicting the rings of the snail by subjecting to the top features of the ensemble regressor like AdaBoost, gradient boost, extra trees and random forest regressors. The ensemble regressor

**Table 9** Indices score of Gaussian regressor with random forest dataset before scaling

Gaussian kernels	EVS	MAE	MSE	RScore
Gaussian-Dot Product	0.8167	0.8277	0.8267	0.8227
Gaussian-White Kernel	0.8267	0.8367	0.8347	0.8357
Gaussian-Pairwise Kernel	0.8067	0.8117	0.8117	0.8247
Gaussian-Rational Quadratic	0.7632	0.7232	0.7232	0.7222
Gaussian-ExpSine Squared	0.7533	0.7332	0.7332	0.7333
Gaussian-Matern	0.7723	0.7243	0.7243	0.7235
Gaussian-RBF	0.7277	0.7137	0.7146	0.7167
Gaussian-Constant Kernel	0.8387	0.8287	0.8277	0.8287
Gaussian-Exponentiation	0.8432	0.8032	0.8032	0.8132

**Table 10** Indices score of Gaussian regressor with random forest dataset after scaling

Gaussian Kernels	EVS	MAE	MSE	RScore
Gaussian-Dot Product	0.8767	0.8777	0.8777	0.8767
Gaussian-White Kernel	0.8567	0.8567	0.8587	0.8557
Gaussian-Pairwise Kernel	0.8867	0.8817	0.8817	0.8847
Gaussian-Rational Quadratic	0.8732	0.8732	0.8732	0.8732
Gaussian-ExpSine Squared	0.8833	0.8832	0.8832	0.8833
Gaussian-Matern	0.8423	0.8433	0.8493	0.8495
Gaussian-RBF	0.8967	0.8997	0.8956	0.8967
Gaussian-Constant Kernel	0.9687	0.9687	0.9677	0.9687
Gaussian-Exponentiation	0.8532	0.8532	0.8532	0.8532

**Table 11** Feature importance probability values of the dataset before scaling

Feature	AdaBoost	Gradient boost	Extra trees	Random forest
Sex	0.00448	0.02817	0.06498	0.02854
Length	0.02788	0.01177	0.09740	0.04875
Diameter	0.02429	0.02185	0.13148	0.05369
Height	0.06874	0.01832	0.07464	0.04789
Whole weight	0.05629	0.03893	0.13444	0.08720
Shucked weight	0.17421	0.16483	0.13530	0.16986
Viscera weight	0.09459	0.01510	0.09416	0.07065
Shell weight	0.54952	0.70103	0.26760	0.49342

**Table 12** Feature importance probability values of the dataset after scaling

Feature	AdaBoost	Gradient boost	Extra trees	Random forest
Sex	0.00992	0.02807	0.06889	0.08120
Length	0.02878	0.01296	0.12191	0.09900
Diameter	0.03219	0.02058	0.12486	0.11863
Height	0.06708	0.01849	0.07633	0.07436
Whole weight	0.06105	0.03874	0.12409	0.13342
Shucked weight	0.19921	0.16508	0.13515	0.13976
Viscera weight	0.06750	0.01537	0.09036	0.08852
Shell weight	0.53428	0.70070	0.25840	0.26510

reduced dataset is applied to the Gaussian process regressor with various kernels like Dot Product, White Kernel, Pairwise Kernel, Rational Quadratic, ExpSine Squared, Matern, RBF, Constant Kernel, and Exponentiation for analyzing the performance of the regressor toward predicting the snail rings before and after feature scaling. Experimental results show that the Constant Kernel with Gaussian regressor applied to the random forest feature importance snail age dataset projects the RScore of 0.9687 when compared to other ensemble regressors.

## References

1. Gurney, L.J., Mundy, C., Porteus, M.C.: Determining age and growth of abalone using stable oxygen isotopes: a tool for fisheries management. *Fish. Res.* **72**(2–3), 353–360 (2005)
2. Marrocco, C., Tortorella, T.: Exploiting coding theory for classification: an LDPC-based strategy for multiclass-to-binary decomposition. *Inf. Sci.* **357**, 88–107 (2016)
3. Robinson, L.M., Gledhill, D.C., Moltschanivskyj, N.A., Hobday, A.J., Frusher, S., Barrett, N., Stuart-Smith, J., Pecl, G.T.: Rapid assessment of an ocean warming hotspot reveals “high” confidence in potential species’ range extensions. *Glob. Environ. Chang.* **31**, 28–37 (2015)
4. Oliver, E.C.J., Wotherspoon, S.J., Chamberlain, M.A., Holbrook, N.J.: Projected Tasman Sea extremes in sea surface temperature through the twenty-first century. *J. Clim.* **27**, 1980–1998 (2014)
5. Przeslawski, R., Falkner, I., Ashcroft, M.B., Hutchings, P.: Using rigorous selection criteria to investigate marine range shifts. *Estuarine Coastal Shelf Sci.* **113**, 205–212 (2012)
6. Gopi, K., Mazumder, D., Sammut, J., Saintilan, N., Crawford, J., Gadd, P.: Isotopic and elemental profiling to trace the geographic origins of farmed and wild-caught Asian seabass. *Aquaculture* **502**, 56–62 (2019)
7. Le Bourg, B., Le Bourg, E.: Age Determination and Lifespan of Marine Animal Species, Encyclopedia of Biomedical Gerontology, pp. 26–36. Academic Press (2020)
8. Yucemen, M.S., Askan, A.: Estimation of earthquake damage probabilities for reinforced concrete buildings. Seismic Assessment and Rehabilitation of Existing Buildings, NATO Science Series, vol. 29, pp. 149–164. Springer (2003)

9. Zheng, X.-W., Li, H.-N., Yang, Y.-B., Li, G., Huo, L.-S., Liu, Y.: Damage risk assessment of a high-rise building against multihazard of earthquake and strong wind with recorded data. *Eng. Struct.* **200**, 1096971 (2019)
10. Nakano, Y., Maeda, M., Kuramoto, H., Murakami, M.: Guideline for post-earthquake damage evaluation and rehabilitation of RC buildings in Japan. In: 13th World Conference on Earthquake Engineering, vol. 1, no. 1, pp. 124 (2004)

# Mutual Information Score-Based Clustering for Evaluation of Image Dominant Color



**M. Shyamala Devi, N. K. Manikandan, D. Manivannan,  
Y. Lakshmi Akshitha, G. Chandana, K. Lasya Priya, and G. Vijayalakshmi**

**Abstract** Dominant color of an image is a subset of image classification that interacts with image processing and analysis. Finding the dominant color of an image is used for many creative gaming and medical applications. With the growing technological equipment and social media usage, people use Web camera for security purpose. Even police department tends to analyze the victim of many cases through analyzing the videos and images by finding the hair, dress color of an image. Computer vision is used for detecting the dominant color of any parts of an image. The colorful image is used for implementation for finding the top 8 dominant color of an image. The image is preprocessed by converting the BGR to RGB image format. The image is resized with respect to the width and height of the image and reshaped with three dimensions. The image is applied with several clustering methods to find the dominant colors of an image along with the percentage of its existence. The performance of clustering is analyzed with score, Rand index, adjusted Rand index, mutual information score (MIS), normalized MIS, adjusted MIS, homogeneity score, completeness score, and VMeasure. The scripting is written in Python and implemented with Spyder in Anaconda Navigator IDE, and the experimental results show that the mini-batch clustering outcomes by showing all the performance indices with 0.99 approximately close to 1 in finding the dominant colors of an image.

**Keywords** Machine learning · Clustering · Rand index · Mutual information

## 1 Introduction

Dominant color is a concise and effective signifier that uses representative colors to characterize the color features in a picture's exciting area. The dominant color

---

M. Shyamala Devi (✉) · N. K. Manikandan · D. Manivannan · Y. Lakshmi Akshitha ·

G. Chandana · K. Lasya Priya · G. Vijayalakshmi

Computer Science & Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu, India

e-mail: [shyamaladev@veltech.edu.in](mailto:shyamaladev@veltech.edu.in)

descriptor is appropriate for depicting local visual features and can be used for efficient retrieval in big image files. A normal picture or image is made up of a multitude of distinct colors. Nevertheless, certain proximate analysis of the hyperspectral side limits the number of colors and affects their esthetic, such as screen resolution, color depth, noise, the impacts of neighboring pixels, and so on. Furthermore, due to the restrictions of the visual perception (HVS), particularly in range of vision with poorer color vision deficiency and lower spatial acuity, we can only distinguish a tiny number of them. The deposition of an image into a majority of different relevant areas with almost the same characteristics is referred to as image segmentation. Image segmentation is a critical technology in image processing technique, and the exactness of edge detection has a direct impact on the effectiveness of subsequent tasks. Given its complexity and difficulty, the existing segmentation method has varying levels of success, but research on the topic still faces several challenges. Because the clustering analysis method divides large datasets into separate categories based on a specific standard, it has a wide range of applications in the field of image segmentation. The paper is organized in such a way that literature review is discussed in Sect. 2 followed by the paper contributions in Sect. 3. The implementation setup and results are discussed in Sect. 4 followed by mathematical modeling in Sect. 5 and concluded in Sect. 6.

## 2 Literature Review

The goal of trying to extract color elements from such an image is to create a color palette comprised of the image's dominant colors. In this article, we will establish a color infrastructure to connect the intrinsic color information of pixels. We obtain initial color themes by using an enhanced linear iterative clustering algorithm. In the subsequent steps, we can obtain the final sorted color themes result by learning from human-extract color themes [1]. Color arrangement is essential in art, design, and communication because it can affect the user's perspectives, feelings, and mental well-being. It is time-consuming to create a color theme from scratch when dealing with large batches of images. This paper describes a novel automated framework for obtaining color themes from fabric images. A foreground map is created to aid in the recognition of video analysis areas in the source images. Because the feature representation divides the image into video analysis foreground area and non-visual attention background area, we compute the dominant colors of these two areas and combine them based on some rules to form the original target color theme [2]. This paper proposes a methodology for automatically extracting dominant color based on a area growing algorithm. The proposed method and its steps are as follows: First, an image color space is converted from RGB to HSV. Second, area growing methodology is used to divide an image into color regions, and regions with areas that are less than 1% of the image are filtered out. Third, neighboring regions with similar color are merged, and if the area of the merged region compared to the image is greater than 5%, it is regarded as dominant color [3].

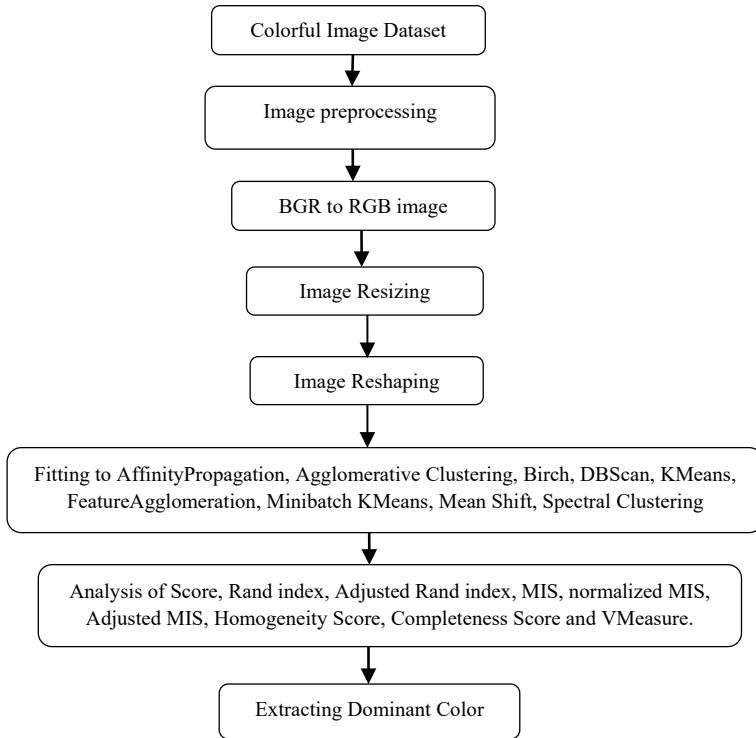
This study gives an overview of the video compression, also known as the multi-media information description interface. It focuses on visual knowledge description, which includes low-level visual descriptors and segment description schemes. The study also highlights some challenges in video image analysis that will need to overcome in future in order to enable effective MPEG-based applications [4]. A signifier that uses no further memory than a standard histogram equalization but outshines it significantly. The color structure descriptor is based on color structure histogram, which is a generalization of color histogram. The color structure histogram encrypts information on the location structure and probability of occurrence of colors in an image. Extraction of the CS histogram in the novel nonlinear HMMD color space and non-uniform quantization of histogram amplitude and frequency to represent bin amplitude statistics derived from consumer oriented image databases improves performance [5]. This paper describes a novel and efficient image searchable technique for extracting features from JPEG compressed images. The proposed technique can produce an efficient histogram from DCT coefficients, which are the critical elements of JPEG, using vector quantization methods and a reference manual generated using a K-means clustering algorithm [6]. Based on machine learning, a framework for developing a model for extracting image prominent colors is presented. The model is trained on life form themes of predominant colors and employs a slew of features centered on human graphic system parameters. We created a database of images with their relation to human themes of influential colors for the context of this research, which is available to the public and available to other investigators. The analysis of data from observers reveals a high level of interobserver agreement on prominent color categories as well as a wide range of prominent colors [7]. This paper describes a data-driven technique for improving an object's preferred color theme. We define our objective as a unified maximization that takes into a desired color, texture-color relationships, and automatic or user-specified color constraints. Color mood spaces and a generalization of a commutative relationship for two configurations make it possible to quantify the distinction between an image and a color theme [8].

Content-based image retrieval refers to a group of method, which aims to solve the issue of finding digital images in a massive database based on their video elements. Private photo and art catalogs, medical imaging, audiovisual news sources, and surveillance cameras are all applications where retrieval of similar images is critical. The primary goal of our research was to improve the efficiency of the query-by-example recognition system using texture features [9]. Art, biochemical, biostatistics, neuroimaging, perceptual neuroscience, color preferences, spectrophotometry, computer simulations, architecture, electrocardiography, language and comprehension, biomedical sciences, neuroscience, biological optics, psychophysics, and physiological optics are just a few of the disciplines that study color vision. This wide range of disciplines, combined with the elusive nature of the subjective experience of color, makes the study of color as demanding, as it is extremely interesting [10]. This paper investigated the role of visual features in categorizing the color of leaves. Our subjects classified images of fallen leaves and their tampered editions, which included randomized reconfigured pixels, leaves universally colored with their mean

color, leaves made by representing the original leaves chrominance allocation about their average, and simple patches colored with the initial leaves mean color schemes [11]. This paper comprehensively explores the color impression of geographically blended configurations of two colors from a barely perceptible distance and shows that it differs significantly from their physically averaged chromogenic color. All mixed color patterns exhibit significant luminance-enhanced perception, and the test of blended monochrome patterns exhibits significant spectrally augmented viewpoint [12].

### 3 Our Contributions

The overall architecture of the work is shown in Fig. 1. The following contributions are provided in this work.



**Fig. 1** Workflow methodology

- Firstly, the image is preprocessed by converting BGR to RGB image format.
- Secondly, the image is resized with respect to the width and height of the image and reshaped with three dimensions.
- Thirdly, the image is applied with several clustering methods like affinity propagation, agglomerative clustering, birch, DBSCAN, K-means, feature agglomeration, mini-batch K-means, mean shift, spectral clustering to find the dominant colors of an image along with the percentage of its existence.
- The performance of clustering algorithm is analyzed with score, Rand index, adjusted Rand index, mutual information score, normalized MIS, adjusted MIS, homogeneity score, completeness score, and VMeasure.

## 4 Implementation Setup and Results

The following image is used for data preprocessing and is shown in Fig. 2.

The image is preprocessed by converting the BGR to RGB image format. Then, the image is resized with respect to the width and height of the image and reshaped with three dimensions. Then, the image is applied with several clustering methods like affinity propagation, agglomerative clustering, birch, DBSCAN, K-means, feature agglomeration, mini-batch K-means, mean shift, spectral clustering to find the dominant colors of an image along with the percentage of its existence. The dominant colors extracted for the image are shown in Fig. 3.



**Fig. 2** Image used for finding dominant color



**Fig. 3** Dominant color for the image in Fig. 2

**Table 1** Clustering performance indices score with dominant colors

Classifier	Score	RI	ARI	MIS	NMIS
Affinity propagation	0.9067	0.9167	0.9067	0.9167	0.9167
Agglomerative	0.9367	0.9267	0.9247	0.9357	0.9357
Birch	0.9167	0.9117	0.9117	0.9147	0.9147
DBSCAN	0.9433	0.9432	0.9432	0.9533	0.9533
K-means	0.9423	0.9443	0.9443	0.9435	0.9435
Feature agglomeration	0.9477	0.9337	0.9446	0.9567	0.9567
Mini-batch K-means	0.9987	0.9987	0.9977	0.9987	0.9987
Mean shift	0.9667	0.9667	0.9667	0.9667	0.9667
Spectral	0.9332	0.9232	0.9332	0.9332	0.9332

**Table 2** Clustering performance indices completeness score with dominant colors

Classifier	AMIS	Homogeneity	Completeness	VMeasure
Affinity propagation	0.9117	0.9167	0.9167	0.9267
Agglomerative	0.9227	0.9227	0.9237	0.9327
Birch	0.9067	0.9117	0.9117	0.9147
DBSCAN	0.9433	0.9432	0.9432	0.9533
K-means	0.9323	0.9343	0.9333	0.9335
Feature agglomeration	0.9477	0.9337	0.9446	0.9567
Mini-batch K-means	0.9987	0.9987	0.9977	0.9987
Mean shift	0.9117	0.9117	0.9117	0.9117
Spectral	0.9273	0.9173	0.9273	0.9273

The performance of the clustering algorithms is done through the metrics like score, Rand index, adjusted Rand index, mutual information score, normalized MIS, adjusted MIS, homogeneity score, completeness score, and VMeasure and is shown in Tables 1, 2, and 3, Figs. 4, 5, and 6.

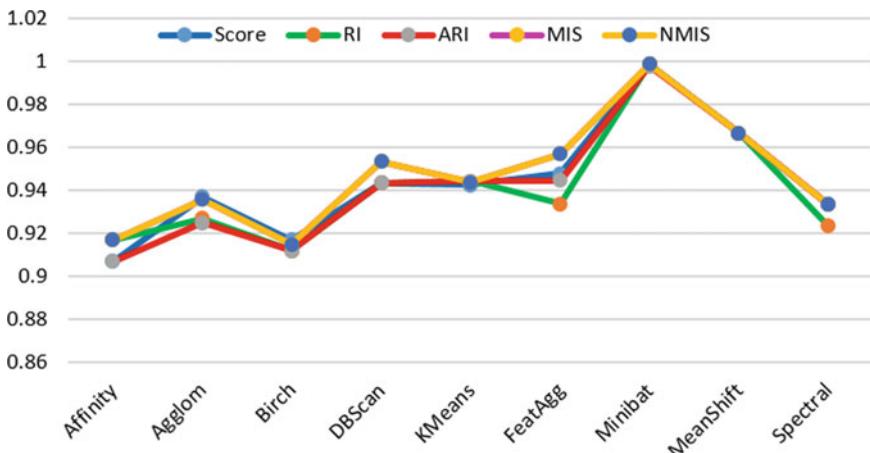
## 5 Mathematical Modeling and Analysis

The image dataset is represented as follows in Eq. (1),

$$\text{Image}_{1000} = \left[ \bigcup_{h=1}^n \left\{ \sum_{i=1}^{255} \sum_{j=1}^{255} I_{ijh} \right\} \right] \quad (1)$$

**Table 3** Percentage of dominant colors for the image

Clustering method								
AffinityPropagation	10	12	11	14	09	13	17	09
Agglomerative	26	13	16	14	10	09	06	14
Birch	11	14	11	18	12	09	13	09
DBScan	10	12	11	14	09	13	17	09
KMeans	17	16	12	13	08	12	16	08
FeatureAgglomeration	12	13	11	12	07	13	14	07
MinibatchKMeans	13	14	13	13	08	14	12	06
Mean Shift	11	15	12	15	09	11	14	08
Spectral	10	12	10	16	10	12	15	09

**Fig. 4** Performance score comparison of clustering

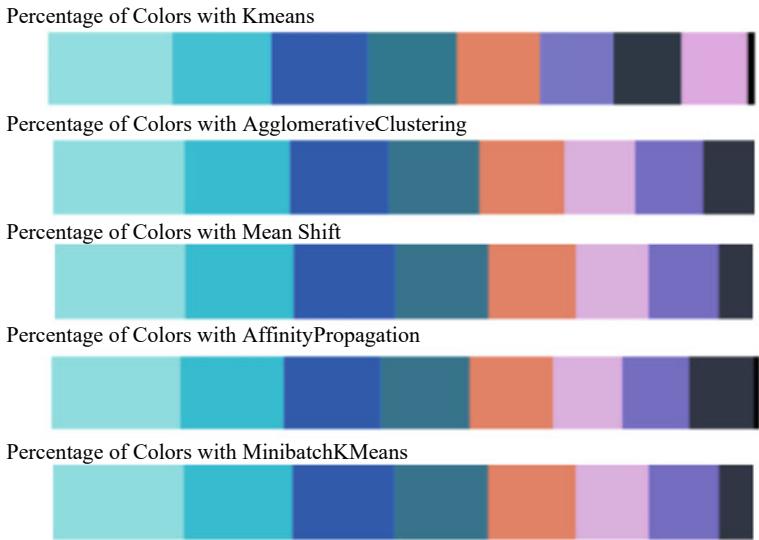
where “ $I_{ij}$ ” represents single image and its pixels in rows. The feature extraction of image is shown in Gaussian equation (2), where “ $r$ ” represents the free parameter denoting the variance of the Gaussian function.

$$\text{Feat}(i, j, r) = \frac{1}{\sqrt{2\pi r}} \exp \frac{(i^2 + j^2)}{2r^2} \quad (2)$$

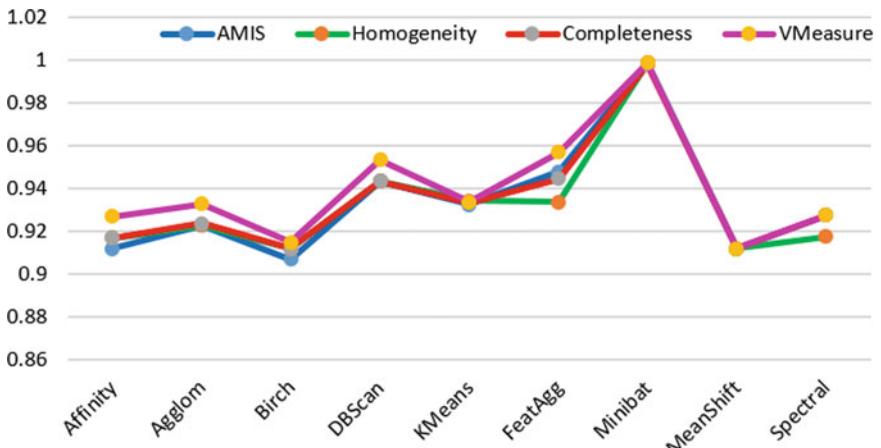
The filtering of the images in the dataset is denoted with the Gaussian orientation function as shown in Eq. (3).

$$\text{Ori}(i, j, r, \theta) = \text{Feat}_{ii} \cos^2(\theta) + 2\text{Feat}_{ij} \cos(\theta) \sin(\theta) + \text{Feat}_{jj} \sin^2(\theta) \quad (3)$$

where  $\text{Feat}_{ii}$ ,  $\text{Feat}_{ij}$ , and  $\text{Feat}_{jj}$  denote the second derivatives of Gaussian function and are shown in Eq. (4)–(6)



**Fig. 5** Percentage of dominant colors of an image



**Fig. 6** Performance completeness score comparison of clustering

$$\text{Feat}_{ii}(i, j, r) = \frac{(i^2 - r^2) \exp\left(\frac{-(i^2 + j^2)}{2r^2}\right)}{\sqrt{2\pi r}^5} \quad (4)$$

$$\text{Feat}_{jj}(i, j, r) = \frac{(j^2 - r^2) \exp\left(\frac{-(i^2 + j^2)}{2r^2}\right)}{\sqrt{2\pi r}^5} \quad (5)$$

$$\text{Feat}_{ij}(i, j, r) = \frac{ij \exp\left(\frac{-((i^2+j^2))}{2r^2}\right)}{\sqrt{2\pi r}^5} \quad (6)$$

The image resizing is done based on the aspect ratio of the image as in Eq. (7)

$$\text{Aspect Ratio} = \frac{\text{Image\_width}}{\text{Image\_height}} \quad (7)$$

The clustering is performed between the image clusters and the pair wise distance is in Eq. (8) and pooled clusters in Eq. (9),

$$PD_r = \sum_{i,j \in C_r} pd_{ij} \quad (8)$$

$$PW_k = \sum_{r=1}^k \frac{1}{2n_r} pd_{ij} \quad (9)$$

The clustering gaps between the image pixel points are shown in Eq. (10) and the expectation rate in Eq. (11)

$$\text{CGap}_n(k) = E_n^k(\log(PW_k)) - \log(PW_k) \quad (10)$$

$$\log(PW_k) = \frac{\log(PW_k)}{12} - \left(\frac{2}{p}\right) \log(k) + \text{constant} \quad (11)$$

The distance between all clusters in Eq. (12) and cluster mean in Eq. (13).

$$PD_r = \sum_{i=1}^{C_r} \sum_{j=1}^{C_r} \|pd_i - pd_j\|^2 \quad (12)$$

$$\frac{1}{2n_r} PD_r = \frac{1}{2n_r} \sum_{i=1}^{C_r} \sum_{j=1}^{C_r} \|pd_i - pd_j\|^2 \quad (13)$$

The performance of image clustering is done by Rand index in Eq. (14), where “x” is number of pairs in same cluster and “y” is number of pairs in different clusters.

$$\text{Rand index} = \frac{x + y}{C_r} \quad (14)$$

The mutual information score and normalized MIS are given in Eq. (15), (16).

$$\text{MIS}(i, j) = \sum_{i=1}^{C_r} \sum_{j=1}^{C_r} \frac{\|pd_i - pd_j\|^2}{Cn} \log \frac{Cn \|pd_i - pd_j\|^2}{\|pd_i\| \|pd_j\|} \quad (15)$$

$$\text{NMIS}(i, j) = \frac{\text{MIS}(i, j)}{\text{Mean}(\text{MIS}(i), \text{MIS}(j))} \quad (16)$$

The adjusted Rand index if found by Eq. (17) with homogeneity and completeness score in Eq. (18)–(21).

$$\text{AMIS}(i, j) = \frac{\text{MIS}(i, j) - \text{Entropy}(\text{MIS}(i, j))}{\text{Mean}(\text{MIS}(i), \text{MIS}(j)) - \text{Entropy}(\text{MIS}(i, j))} \quad (17)$$

$$\text{HomoScore} = 1 - \frac{H(\text{MIS}(i/j))}{H(\text{MIS}(i))} \quad (18)$$

$$\text{CompleteScore} = 1 - \frac{H(\text{MIS}(i, j))}{H(\text{MIS}(j))} \quad (19)$$

$$H(\text{MIS}(i/j)) = \sum_{i=1}^{C_r} \sum_{k=1}^{C_k} \frac{C_{r,k}}{C} \log \left( \frac{C_{r,k}}{C_k} \right) \quad (20)$$

$$H(\text{MIS}) = \sum_{i=1}^{C_r} \frac{C_{r,k}}{C_k} \log \left( \frac{C_{r,k}}{C_k} \right) \quad (21)$$

## 6 Conclusion

This paper explores the performance of clustering in finding the dominant colors of an image. It attempts to analyze the performance of clustering with respect to various performance indices like score, Rand index, adjusted Rand index, mutual information score, normalized MIS, adjusted MIS, homogeneity score, completeness score, and VMeasure. The analysis of dominant colors of an image is done through various clustering methods like affinity propagation, agglomerative clustering, birch, DBSCAN, K-means, feature agglomeration, mini-batch K-means, mean shift, spectral clustering. Experimental results show that the mini-batch clustering outcomes by showing all the performance indices with 0.99 approximately close to 1 in finding the dominant colors of an image.

## References

1. Zunlei, F., Wolong, Y., Chunli, F., Jie, L., Mingli, S.: Finding intrinsic color themes in images with human visual perception. *Neurocomputing* **273**, 395–402 (2018)
2. Liu, S., Jiang, Y., Luo, H.: Attention-aware color theme extraction for fabric images. *Text Res.* **J.** **88**(5), 552–565 (2018). <https://doi.org/10.1177/0040517516685278>
3. Li, A., Bao, X.: Extracting image dominant color features based on region growing. In: Proceedings of International Conference on Web Information Systems and Mining, pp. 120–123 (2010). <https://doi.org/10.1109/WISM.2010.116>
4. Philippe, S.: Overview of the MPEG-7 standard and of future challenges for visual information analysis. *EURASIP J. Appl. Signal Process.* **4**, 1–11 (2002)
5. Messing, D.S., Van Beek, P., Errico, J.H.: The MPEG-7 colour structure descriptor: image description using colour and local spatial information. *Image Process.* **1**, 670–673 (2001)
6. Pourstani, P., Nezamabadi, H., Askari Moghadam, R., Saeed, M.: Image indexing and retrieval in JPEG compressed domain based on vector quantization. *Math. Comput. Model.* **57**, 1005 (2013)
7. Weingerl, P., Hladnik, A., Javorsek, D.: Development of a machine learning model for extracting image prominent colors. *Color Res. Appl.* **45**, 409 (2020)
8. Wang, B., Yu, Y., Wong, T.T., Chen, C., Xu, Y.-Q.: Data-driven image color theme enhancement. *ACM Trans. Graph.* **29**(6), 1–10 (2010)
9. Hladnik, A., Poljicak, A.: Improving performance of content based image retrieval system with color features. *Acta Graphica* **27**, 7–12 (2017)
10. Conway, B.R., Eskew, R.T., Martin, P.R., Stockman, A.: A tour of contemporary color vision research. *Vision Res.* **151**, 2–6 (2018)
11. Zarko, M., Robert, E., Matteo, T., Gegenfurtner, K.R.: Categorizing natural color distributions. *Vision Res.* **151**, 18–30 (2018)
12. Choi, S.H., Kim, H., Shin, K., Kim, H., Song, J.: Perceived color impression for spatially mixed colors. *J. Disp. Technol.* **10**(4), 282–287 (2018)

# Hybrid Deep Learning-Based Music Recommendation System



M. Sunitha, T. Adilakshmi, and Mehar Unissa

**Abstract** The fundamental objective of musical recommendations is to propose songs that are appropriate to the tastes of the user. In this paper, we have developed, implemented, and analyzed music recommendation systems with variations of music recommendation algorithm using MLP neural network such as content-based model using MLP neural network, collaborative model using MLP neural network, and hybrid model using MLP neural network. We have also looked into the accuracy and precision of all the three algorithms with various activation functions.

**Keywords** Content-based filtering · Streaming history · Collaborative filtering · Hybrid recommendation · Deep learning · Music recommendations

## 1 Introduction

The digitization of music has a huge effect on the music industry. This transition was also brought about by the advent of the Internet since it used to be both the source of digital music and its distribution route. A tremendous amount of music was thereby made available. This is an issue in the current age of digital music distribution that restricts the scope of music. We already face the new music-based paradigm: listeners are already provided with instantly accessing unprecedentedly large digital music libraries. Most music recordings are online, with the quantity of digital music growing and counting tens of millions. Major Internet shops like the iTunes Store have up to 28 million music each month adding thousands of new music. This is not surprising as the music plays a vital role in everyday living, and more and more people use modern technologies to express and share their creativity in music. The music advisor system is a system that learns from the history of the

---

M. Sunitha (✉) · T. Adilakshmi · M. Unissa  
Vasavi College of Engineering, Hyderabad, India  
e-mail: [m.sunithareddy@staff.vce.ac.in](mailto:m.sunithareddy@staff.vce.ac.in)

T. Adilakshmi  
e-mail: [t\\_adilakshmi@staff.vce.ac.in](mailto:t_adilakshmi@staff.vce.ac.in)

users and recommends songs they'd want to hear probably in future. To construct an efficient recommending system, we have incorporated numerous algorithms.

While email has early embraced the benefits of the use of advisor systems, the music domain is impacted by offline radio stations where static playlists are transmitted to every listener based on track popularity and expert pre-selections. The balance has shifted with the emergence of music streaming systems, such Last.fm<sup>1</sup> or Spotify<sup>2</sup>, and users can now build their own personal radio stations. In turn, consumers are now obliged to construct their own playlists and discover new music less likely. An elegant complement that makes use both knowledge of the audience and of the user's previous listening experience is a musical advisory system. As the Internet is now widely used, one or more recommendation systems have been found among most of the computer, tablet, and smartphone users. For instance, you can envision a visit to a favorite online shop to explore a specific item. After you find this product and click on the direct link, the page can contain a section named, "customers who have also purchased this item." These objects are listed in relation to the product under review as potentially interesting items. A customized list of recommendations will be provided for registered users automatically when they log on to the Website. A recommendation system is the program that provides advice. Custom suggestions require a system to get some user knowledge. In other words, a user profile including the preferences of each user must establish and retain a recommendation system. These preferences of users can be expressly gained through the user's request to rate a certain item, or by user activity tracking.

## 2 Motivation

The strength of the Internet has enabled many users to integrate recommendation systems into their daily lives. If someone has used Facebook, LinkedIn, or even Netflix, they have a system that promotes new stuff according to different parameters. Amazon.com uses a system that recommends products based on the browsing/buying history of the user and products bought by other users with the same taste. The prominent Internet services Pandora and Last.fm choose music for users. These and other Web-based music applications generate revenues not previously available and help businesses grow to other markets. Music recommendation algorithms contribute to the fuel economy of digital songs, helping consumers to discover music. In 2012, digital revenues of record labels have been estimated at 9% higher than in 2011, according to the International Federation of Phonographic Industries (IFPI).

### 3 Literature Review

Isinkaye et al. (2015), recommendation algorithms open up additional options to collect tailored Internet information. It also helps to reduce the problem of over-loading information, which is a very common phenomena of information recovery systems and allows consumers access to products and services not readily accessible to users on the system. In this research, two classic recommendation methodologies were explored, and their merits and limitations were highlighted by various types for improving hybridization policies [1].

Zhang et al. (2018), the author has presented in this essay an exhaustive evaluation of the most remarkable work on high-level learning systems. Author has recommended the organization and grouping of current articles by proposing a range of inertial research prototypes. We also addressed how deep learning techniques could be used to advise on the benefits/disadvantages. Authors also discuss some of the most important open issues and promising extensions for future years. In the last few decades, the themes of hot study have been both deep learning and recommendation systems. Each year, numerous new techniques are being developed, and new models are emerging. Authors believe this survey will give readers with an overview of the important components of the project, clarify the most significant developments, and clarify future investigations [2].

Hornung et al. (2018), music really is an area that is subjective. That is why TRecS balances generally recognized measurements such as track and tag similarity with measurements such as time similarity, which are music specific. With our serendipity metric, new music can be discovered. We have found in our empirical study that the predictive quality improves with each user's number of recommendation lists. Since all ratings are directly returned to the predictive algorithm, the model works efficiently. We also wanted to find out, based on four groups, which orchestrate of the three sub-recommenders results best [3].

Zeshan Fayyaz et al. (2020) in this work, authors presented a complete RS survey that presents many sorts of RSs such as collaborative filtering, content, demo, utility, knowledge-based, and hybrid. The hybrid system is also given and characterized in a variety of combination tactics in weighted, blended, and switched. Their four major obstacles included cold starts, data scarcity, scalability and variety, and measuring measures utilized for the assessment of the success of a recommendation system. Furthermore, authors demonstrate how e-commerce and various fields including transport, e-health, agriculture, and the media have been implemented in recommendation systems [4].

Betru et al. (2017) in this paper, authors discussed standard recommendation and deep learning recommendation system methodologies. The deeper learning of recommender systems will next be investigated and criticized. By comparing the previous system recommendation methodologies, we may conclude that the hybrid approach would provide superior accuracy and performance. The research also suggests that a lot can still be done to achieve better results with content-based and collaborative filtering recommendations methodologies [5].

Schedl et al. (2019), like in many other academic fields, in music recommendation systems, deep learning (DL) is increasingly accepted (MRS). This field uses deep neural networks, primarily for the extraction of latent characteristics from audio or metadata of music items and the learning of sequential patterns of music items (paths or artists) from playlists or hearing sessions. Latent item factors are usually included in a content-based filtering or hybrid MRS, whereas music elements sequence models, e.g., automatic playlist continuation, are utilized for sequence music recommendations. This page describes specific characteristics of the RS research music field [6].

Smith et al. (2019), recommendation algorithms are prevalent in discovery and music distribution services but significantly less prevalent in software for music composition, such as EarSketch, an online learning environment that enables students to write musician code. The EarSketch interface includes a sound library, which students can access via a browser. The present sound browser implementation supports basic search and filtering functions, but no sound discovery mechanism such as a recommendation system. Therefore, users have typically chosen a restricted section of high-frequency sounds that lead to a decreased diversity of composition [8].

Kathavate et al. (2021), twenty artists are involved in the experiment. In future, authors will aim to make the recommendation stronger by adding a higher number of artists and languages to give even better playlists for people. In order to compare the findings and seek out better results, here also use the system with various learning models. Authors wanted to provide the customers a preference for what they want to hear when there are thousands of music out there and after a step closer, we were satisfied. An emotional detector system can be built for future use, which can propose the music through recognition of our face emotion [9].

Schedl et al. (2018), the authors have recognized numerous major obstacles to the field of music advisory systems research (MRS) throughout this trends and survey article. Among others, these are at the heart of modern MRS research. Authors have [1] discussed the problem of cold beginning items and users with their features in music, [2] discussed the challenge of automatic continuation of the playlist due to the recently emerged user's demand for recommended musical experiences rather than single tracks.

Schedl et al. (2018), the authors have recognized numerous major obstacles to the field of music advisory systems research (MRS) throughout this trends and survey article. Among others, these are at the heart of modern MRS research. Authors have [1] discussed the problem of cold beginning items and users with their features in music, [2] discussed the challenge of automatic continuation of the playlist due to the recently emerged user's demand for recommended musical experiences rather than single tracks [10].

Vall et al. (2019), music advisory systems have become a significant tool to enable consumers' interaction with ever-greater music archives, online music shops, and personal gadgets, for example. The autonomous continuation of music play listing allows the suggestion of music streams that adapt to certain (maybe short) listening

sessions is an important task in music recommender systems. Earlier investigations showed that collaborative filtering reveals underlying play listed co-occurrence patterns that are beneficial for predicting playlist continuations for collections of precise music playlists. However, a significant long-tailed distribution exhibits in most music collections [11].

Wang et al. (2020), this work proposes to recommend higher quality song sequences with a hybrid music recommendation system based on reinforcement learning (PHRR). WMF and CNN are taught to learn audio signals for the tune. Authors also provide a model-based reinforcement learning framework, which simulates listeners and model the problem of reinforcement learning as a decision-making process based on the preferences of listeners, for both the transitions of songs. Authors innovatively improve the simulation of the interaction process to update the model more data efficiently to capture modest changes in listeners' preferences sensitively. Real-world dataset experiments show that PHRR works better than other comparison algorithms in the music suggestion [12].

Paper title	Authors	Proposed method	Drawbacks
Location-based Orientation Context Dependent Recommender System for users	Joe and Raj [11]	This study introduces an RS that employs the smallest bounding box for every consideration, as opposed to other typical RS that rely solely on the geographic point to establish the location of an object	The cold state problem will occur when a person who has no knowledge of social network joins the recommendation systems. This is due to the fact that using the preference of a person that has no history would result in an empty user-matrix
Recommendation systems: algorithms, challenges, metrics, and business opportunities	Ebrahimian et al. [4]	In this work, authors presented a complete RS survey that presents many sorts of RSs such as collaborative filtering, content, demo, utility, knowledge-based, and hybrid. The hybrid system is also given and characterized in a variety of combination tactics in weighted, blended, switched	Their four major obstacles included cold starts, data scarcity, scalability and variety, and measuring measures utilized for the assessment of the success of a recommendation system

(continued)

(continued)

Paper title	Authors	Proposed method	Drawbacks
Comparison of machine learning algorithms to classify Web pages	Arthur et al. [5]	EcoRec is a perspective rating prediction scenario for users who have not rated purchased products and employs two or more recommender algorithms. It recommends products to clients and provides an enriched user-item matrix	Data that were previously unlabeled were used and then labeled. This is the only parameter that is specified explicitly. The measure of Web engagement and temporal activities is not taken into account
Chemical sensing with familiar devices. Angewandte Chemie International Edition	Raphaeli et al. [7]	Considered Web engagement measures when working on clickstream data. This strategy could assist online retailers in engaging customers on their mobile devices at any time and from any location	Only three types of Web interaction metrics are evaluated out of a total of eight
Recommender system based on pairwise association rules	Osadchiy et al. [8]	Using the implicit social graph to generate the pairwise association rule. It sort the things that take the least amount of time to get to the consumer	The timestamp is not regarded as significant. In terms of areas and locales, there is no recommendation for a product pair. The algorithms' costs have not been investigated

## 4 Methodology

### 4.1 Working

The algorithm makes use of an input vector that contains collaborative, content-based filtering, and streaming history information as well. This information is loaded into a deep neural network, which trained to recognize patterns in the user's history through a training process, eventually recommends songs it believes the user will appreciate.

## 4.2 Overview

The following is a broad description of our music recommendation system: A multi-layer perceptron model takes song metadata like genre, year, playlist, tracks, and the user's streaming history other audio features as input and outputs a list of songs that the user may or may not like. It returns a score ranging from 0 to 1, indicating the likelihood that the listener would enjoy the selected song based on the previous experiences. Our algorithm locates the  $k$  highest-scoring songs and suggests them to the user.

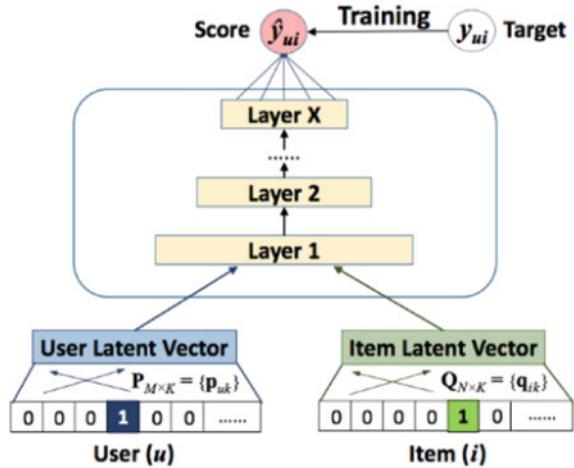
## 4.3 Content-Based Filtering

Song information is used in content-based filtering. It takes two types of metadata into account: genre and release year. This information comes from the Spotify Music released between 1922 and 2021 years. When given as input to the multi-layer perceptron neural network described below, the model can learn a content-based filtering strategy. One metadata category (genre) is unorderable, but the other (release year) can be ordered. Then, based on the numerical audio attributes of each genre, we utilized a basic K-means clustering technique to partition the genres in this dataset into ten groupings. We also use K-means from the dataset to cluster songs. The same genres will sound similar and come from similar eras, and the same may be said for the songs within those genres. We use this information to create a recommendation engine that takes the data points from the songs, a user has listened to and suggests tracks that match neighboring data points. We use "Spotipy," a Python client for the Spotify Web API that makes it simple for developers to interact with it.

## 4.4 Collaborative Filtering

In order to produce a prediction, collaborative filtering focuses on user-item associations. We create a sparse one-hot encoding matrix that depicts user-item associations (tracks). We can now construct the model using the sparse user-item encoding matrix. Because there is a common in-valid playlist-track combination, we create negative training examples (in-valid playlist-track pairings). We know which music is in the playlist, but not which ones are not. The embeddings (low-dimensional) for each playlist and item are generated using the playlist (user) and item vectors. These vectors are then fed into the model as input. Our model uses multi-layer perceptron (neural net) techniques similar to those presented in Fig. 1. We propose an instantiation of NCF using a multi-layer perceptron (MLP) to analyze the user-item interaction function.

**Fig. 1** Deep collaborative filtering model



#### 4.5 Multi-layer Perceptron (MLP)

As NCF uses two paths to model users and items, it is natural to concatenate their features. In multimodal deep learning research [9], this design has been widely used. However, simply concatenating vectors does not account for any interactions between user and item latent features, making it unsuitable to describe collaborative filtering. To solve this problem, we propose adding hidden layers to the concatenated vector and learning the interaction between user and object latent attributes using a conventional MLP. In this method, rather than using a fixed element-wise product to learn the interactions between  $p_u$  and  $q_i$ , we can give the model a lot more flexibility and non-linearity. Our NCF framework defines the MLP model as follows:

$$\begin{aligned}
 z_1 &= \phi_1(p_u, q_i) = \left[ \frac{p_u}{q_i} \right], \\
 \phi_2(z_1) &= a_2(w_2^T z_1 + b_2), \\
 \phi_L(z_{L-1}) &= a_L(W_L^T Z_{L-1} + b_L), \\
 \hat{y}_{ui} &= \sigma(h^T \phi_L(z_{L-1})) \tag{1}
 \end{aligned}$$

where  $w_x$ ,  $b_x$ , and  $a_x$  are the  $x$ -th layer's perceptron's weight matrix, bias vector, and activation function, respectively. Sigmoid, hyperbolic tangent (tanh), and rectifier (ReLU) are just a few of the activation functions available for MLP layers. We would like to look at each function individually: (1) The sigmoid function requires each neuron to be in the (0,1) range, which may reduce the quality of the model; and it is believed to suffer with saturation, in which neurons cease to acquire when their output is around 0 or 1. (2) Despite the fact that tanh is a better alternative and has been largely accepted [6], it only mitigates the drawbacks of sigmoid to a limited degree, since it may be considered as a rescaled variant of sigmoid ( $\tanh(x/2) =$

$2\sigma(x) - 1$ ). And (3) as a result, we choose ReLU, which cease to acquire when their output is around 0 or 1.2) Despite the fact that tanh is a better alternative and has been largely accepted [6], it only mitigates the drawbacks of sigmoid to a limited degree, since it may be considered as a rescaled variant of sigmoid ( $\tanh(x/2) = 2\sigma(x) - 1$ ). And (3) as a result, we choose ReLU, which may be more biologically plausible and has been shown to be non-saturated [9]; also, it favors sparse activations, making it well-suited to sparse data and reducing the likelihood of overfitting the model. Our findings suggest that ReLU outperforms tanh, which outperforms sigmoid.

## 4.6 Algorithm Overview

The main aim of this study is to construct a music recommendation application. The app lets users select the songs on the device and listen to them. If a certain song is listened to by a user, a log is established. We employ many ways to create recommendation engine to offer songs to consumers. The fundamental reason for this proposed system is to enhance the capacity of the standard system of recommendations. Traditional music advice systems depend on collaborative filtering or content-based filters for recommendations to be generated. This design is to create music recommendations that are user-friendly, without significant co-use and that are useful in terms of music similarity. A large number of co-utilized music in various styles might provide diversion in the recommendations. The several random steps in both models allow various recommendations to be created, although they do not ensure innovation, given the same combinations of inputs.

**Algorithm:** Multi-layer Perceptron for Hybrid Music Recommendation System

Input: Genre, year, playlist, tracks, streaming history, audio features.

Dataset: Spotify million playlist.

Output: recommendationList[];

Method:

Begin

- (1) Using the dataset, create a test file.
- (2) Using metadata as an input, create a metadata-related training file.
- (3) Send the multi-layer perceptron the training and testing files.
- (4) Comparing test data for similarities and classifying
- (5) If they are comparable,
- (6) class equals to 1
- (7) If not,
- (8) class equals to 0
- (9) Adding songs to the matched recommendationList that have class = 1.

end

## 5 Results

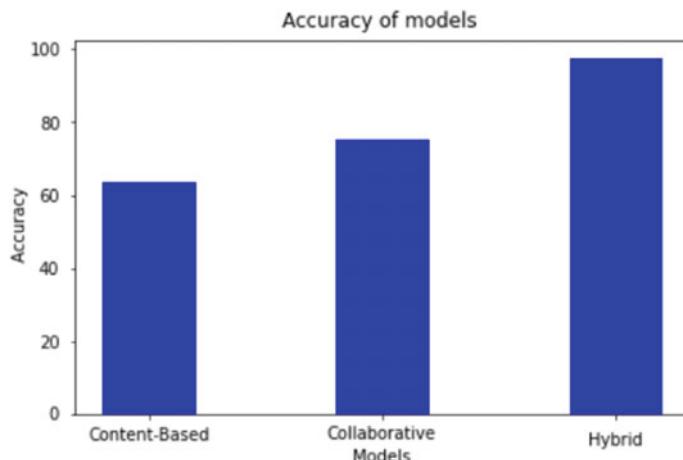
### A. Dataset

The performance of our model is evaluated using the publicly accessible Spotify playlist dataset. To deploy the data, we employ the layered approach shown in Fig. 2. This dataset is based on the selection of individuals in the #nowplaying dataset who broadcast their #nowplaying Twitter via Spotify. In principle, the database comprises subscribers, tracks, and the songs that appear in those playlists. “User id,” “artistname,” “trackname,” and “playlistname” are the necessary fields in the csv format containing the dataset, where user id is a hashed user’s Spotify user name.

### B. Performance evaluation

The model is implemented using the Python language. The MLP model is implemented using the TensorFlow framework. We create a sparse one-hot encoding matrix that links users (playlist) and item associations (tracks). Then, by using numeric audio properties of every category, we calculated the cosine similarity. Finally, we used a standard K-means clustering approach to divide the genres in this database into ten groups. The embeddings (low-dimensional) for every playlist and item are created using playlists (more commonly termed u for user) item I vector and numeric characteristics. The dot product is used to integrate the three embeddings in generalized matrix factorization (GMF) (this is the classic matrix factorization).

Table 1 shows what results would look like after performing hybrid recommendations. It gives the song along with the probability that the particular song will be liked by user. We compare various methods using MLP Network such as content-based



**Fig. 2** Accuracy comparison of the hybrid model with content and collaborative models

**Table 1** Results for sample users

User_id	Probability	Track_name	Track_artist
4962	0.999999	Runaway	Timeflies
3177	0.999999	Ocean away	Ellon
6228	0.999999	Daisy cutter	311
715	0.999995	Elder scroll	JAY Z
9901	0.999995	Little queen	Heart
10,950	0.999994	Pure	David Keller

**Table 2** Comparison of the proposed models

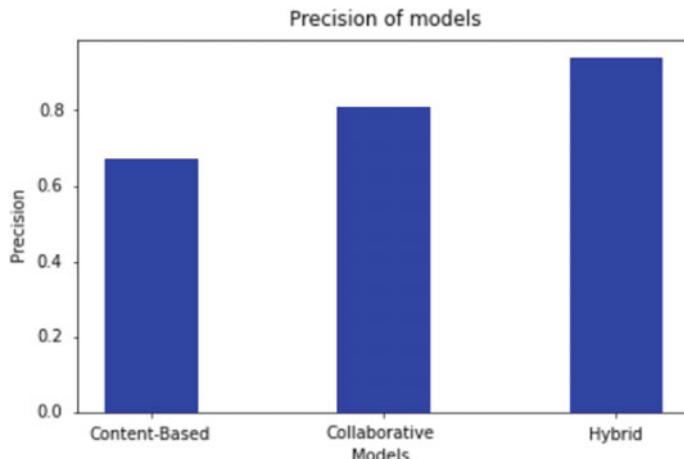
Measure/model	Content-based filtering using MLP	Collaborative filtering using MLP	Hybrid model using MLP
Accuracy (%)	63.5	75	97.5
Precision (%)	67	81	94

model using MLP neural network, collaborative model using MLP neural network, and hybrid model using MLP neural network.

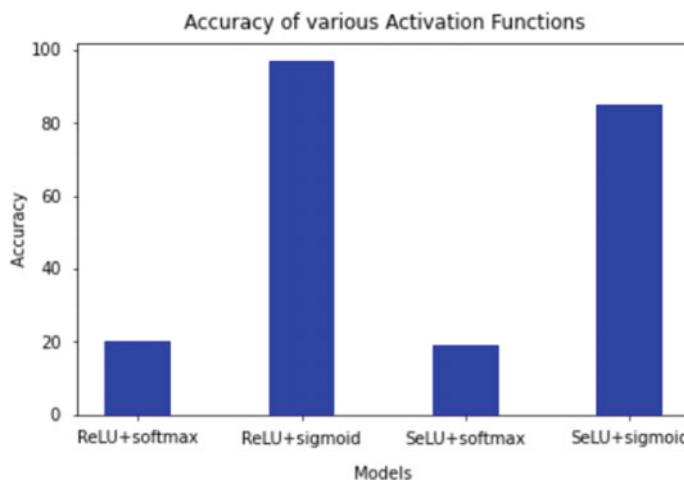
As we can see accuracy of hybrid model is more at 97% when compared to collaborative and content-based model. It tends to give more accurate results than the other two. Whereas content-based model is the lowest with 64% accuracy (Table 2).

From the graph, we can understand that the hybrid model has much more precision than the other two models. And content-based model has the lowest precision when compared to other models. We have also looked into the accuracy and precision of all the three algorithms with a set of activation function such as ReLU+ sigmoid, where ReLU is used as the hidden layer and sigmoid is used as the output layer similarly, we used ReLU+ Softmax, SeLU+ sigmoid, and SeLU+ softmax and have selected the one with the highest accuracy and precision. The music advice is a highly complicated subject since it is necessary to structure music so that the favorite songs are recommended to users that will not be defined. Practical tests by real users assessed the offered algorithms and framework satisfactorily (Fig. 3).

We have analyzed our model with different activation such as ReLU, sigmoid, softmax, and SeLU. We choose ReLU as the hidden layer because it carries out complex calculation faster and easy way; SeLU gives slope larger than one than one for the positive input. These two-activation functions are taken as an input/hidden layer. Whereas Softmax and sigmoid activation function are used as the output layer. Our observation is that ReLU+ sigmoid has higher accuracy and precision as in comparison with other pair of activation function. So, we take it into consideration for building our model (Figs. 4 and 5; Table 3).



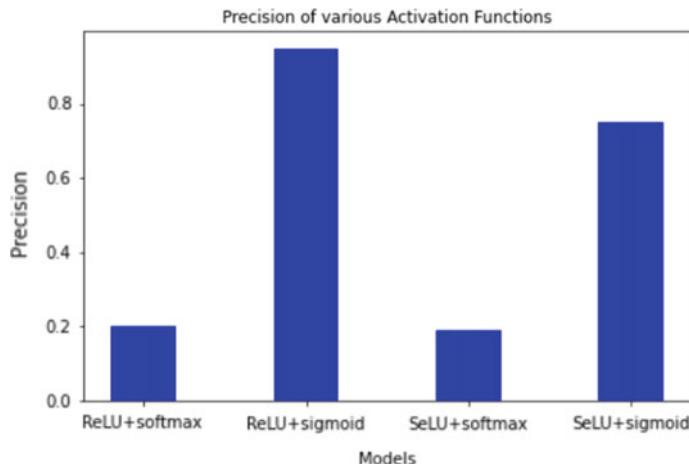
**Fig. 3** Precision comparison of the hybrid model with content and collaborative models



**Fig. 4** Accuracy comparison of proposed hybrid model with various activation functions

## 6 Conclusion

Recommendation algorithms are an underappreciated component of our daily lives, dictating what we listen to on Spotify and view on YouTube. The precision of these systems is still being improved via research. We discuss such a recommendation engine in our research, which takes both content-based and collaborative filtering into consideration as data to a deep neural network. Similarly, to the methods used by Spotify, Pandora, and other music streaming services, this app analyzes customer musical interests to make music predictions. We aimed to solve frequent issues in



**Fig. 5** Precision comparison of proposed hybrid model with various activation functions

**Table 3** Analysis with different activation functions

Hidden layer + output layer	ReLU + Softmax	ReLU + Sigmoid	SeLU + Softmax	SeLU + Sigmoid
Accuracy (%)	20.50	98	19.00	85
Precision	0.2	0.94	0.19	0.82

existing algorithms in the literature, such as the lack of genuine updates and many iterations, in the creation of our method. Input types that are varied as a conclusion, we have a system that makes high-precision recommendations and seems to be easily expandable. Amazon or Netflix is examples of such services.

## 7 Future Scope

We will continue to work on the model to improve prediction accuracy. An emotional detector system can be built for other applications that recommends the music by recognizing our face emotion. We wish to grasp better, also, the consequences and the serendipity function of the rating explanation feature. More human behavioral features will be incorporated into the model in future. For suggestion, we wish to assess the role of these features. As the system that stimulates users' mood becomes more and more relevant user context data available, more improvisational, optimal selection of songs may be produced to recommend.

## References

1. Isinkaye, F.O., Folajimi, Y.O., Ojokoh, B.A.: Recommendation systems: principles, methods and evaluation. *Egypt. Inf. J.* (2015)
2. Zhang, S., Yao, L., Sun, A., Tay, Y.: Deep learning based recommender system: a survey and new perspectives. *ACM Comput. Surv.* (2018)
3. Hornung, T., Ziegler, C.-N., Franz, S.: Evaluating Hybrid Music Recommender Systems. *IEEE* (2013)
4. Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A., Kashef, R.: Recommendation systems: algorithms, challenges, metrics, and business opportunities. *Appl. Sci.* (2020)
5. Betru, B.T., Onana, C.A., Batchakui, B.: Deep learning methods on recommender system: a survey of state-of-the-art. *Int. J. Comput. Appl.* (2017)
6. Schedl, M.: Deep learning in music recommendation systems. *Front. Appl. Math. Stat.* (2019)
7. Kathavate, S.: Music recommendation system using content and collaborative filtering methods. *IJERT* (2021)
8. Schedl, M., Zamani, H., Chen, C.-W., Deldjoo, Y., Elah, M.: Current challenges and visions in music recommender systems research. *Int. J. Multimedia Inf. Retr.* (2018)
9. Joe, A., Vijesh, C., Raj, J.S.: Location-based orientation context dependent recommender system for users. *J. Trends Comput. Sci. Smart Technol. (TCSST)* **3**(01), 14–23 (2021)
10. Chen, C.-W., Boom, C.D., Garcia-Gathright, J., Lamere, P., McInerney, J., Murali, V., Rawlinson, H., Reddy, S., Yon, R.: The million-playlist dataset. Spotify - RecSys Challenge 2018 (2018). [Online]. Available: [https://www.aicrowd.com/challenges/spotify-million-playlist-dataset-challenge/dataset\\_files](https://www.aicrowd.com/challenges/spotify-million-playlist-dataset-challenge/dataset_files)
11. Joe, V.C., Raj, J.S.: Location-based orientation context dependent recommender system for users. *J. Trends Comput. Sci. Smart Technol. (TCSST)* **3**(01), 14–23 (2021)
12. Sungheetha, A., Sharma, R.: Transcapsule model for sentiment classification. *J. Artif. Intell.* **2**(03), 163–169 (2020)

# CNN-Based Deep Learning Network for Human Activity Recognition During Physical Exercise from Accelerometer and Photoplethysmographic Sensors



Sakorn Mekruksavanich and Anuchit Jitpattanakul

**Abstract** Wearable device advances such as electrocardiography (ECG), surface electromyography (sEMG), electroencephalography (EEG), photoplethysmography (PPG), and inertial data sensors have resulted in a complicated, massive, and diverse biometric analysis of data. Consequently, PPG sensors have gained popularity in wearable appliances as the primary method for tracking heart rate daily. Unlike the ECG and sEMG signals which must be collected using adhesive metal electrodes across the skin, the PPG signal could be collected at peripheral body locations and requires a more superficial physical contact. In this study, we aim to employ a learning technique to the PPG signal for the recognition enhancement of human activity. A new deep learning model named SE-DeepConvNet is proposed for accurately classifying physical exercise patterns. To assess the recognition performance of deep learning networks, including our proposed network, we utilize a benchmark dataset called the PPG dataset. The SE-DeepConvNet surpasses CNN standard deep learning network regarding accuracy (99.32%) and F1-score (99.32%).

**Keywords** Wearable sensors · Photoplethysmography · Physical exercise · Deep learning · Human activity recognition

---

S. Mekruksavanich (✉)

Department of Computer Engineering, School of Information and Communication Technology,  
University of Phayao, 56000 Phayao, Thailand  
e-mail: [sakorn.me@up.ac.th](mailto:sakorn.me@up.ac.th)

A. Jitpattanakul

Department of Mathematics, Faculty of Applied Science, King Mongkut's University  
of Technology North Bangkok, 10800 Bangkok, Thailand  
e-mail: [anuchit.j@sci.kmutnb.ac.th](mailto:anuchit.j@sci.kmutnb.ac.th)

Intelligent and Nonlinear Dynamic Innovations Research Center, Science and Technology  
Research Institute, King Mongkut's University of Technology North Bangkok,  
10800 Bangkok, Thailand

## 1 Introduction

Physical activity is an integral component of many people's daily lives. Various studies have demonstrated the advantages of physical exercise on both mental and physical health [23]. While most individuals instinctively understand the benefits of physical activity, maintaining a consistent fitness routine may be challenging [14, 21]. Even though possessing a community could benefit in motivation, not all of us can or wants to join a club or take fitness courses. Fortunately, there are now many wearable gadgets with exercise monitoring and advising capabilities, such as fitness trackers and smartwatches. With the increasing usage of wearable connected devices, tracking physical activity using inertial sensors has become a prominent field of study to increase the overall quality of life and physical contentment. The activity monitoring issue has often been characterized in earlier research as a human activity recognition (HAR) challenge [16, 17, 20].

One of the most commonly used sensors in smartwatches and wristbands recently is photoplethysmography (PPG), along with electrocardiography (ECG), electroencephalography (EEG), and surface electromyography (sEMG). Plethysmogram signals recorded using an optical system are known as PPG signals. They could be used to monitor changes in blood flow within the microvascular bed of tissue [3]. Another measure of heart rate (HR) is to transmit lighting into the body and estimate blood circulation by measuring the amount of light absorbed. Instead of using adherent metal electrodes to monitor heart and muscle electrical activity, PPG monitoring is performed at peripheral body areas. This tracking is less obvious, providing for less intrusive peripheral body monitoring. Therefore, PPG sensors are frequently used in heart rate monitoring devices, including smartwatches and wristbands.

Various deep learning models for categorization, regression, and generalized pattern classification have been designed for PPG-based human activity recognition. Recurrent neural networks (RNNs) and convolutional neural networks (CNNs) are the most commonly utilized strategies [2, 4, 13]. In this study, a hybrid deep learning model named SE-DeepConvNet is proposed. The model solves the HAR issue by thoroughly sharing data in the same layer through squeeze-and-excitation components. On a publicly available PPG benchmark dataset, we conduct extensive experiments. The empirical findings demonstrate that the proposed SE-DeepConvNet model excels the existing benchmark deep learning accuracy and F1-score.

The following sections comprise the remainder of this article. Section 2 discusses current scholarly publications on the area. The SE-DeepConvNet model is detailed in-depth in Sect. 3. Our empirical findings are summarized in Sect. 4. Finally, Sect. 5 concludes this work and proposes some interesting future endeavors.

## 2 Related Works

### 2.1 *Recognition of Human Activity*

Human activity recognition (HAR) attempts to comprehend human activities to allow computer systems to help individuals proactively depending on their needs. Human activities, such as stepping, seating, working, and jogging, could be described as a collection of operations achieved by the individual during a specific time frame under a particular procedure. Various investigations on HAR have been conducted during the last ten years.

Chen et al. [7] assembled an excellent collection of HAR-related research. Human behaviors of daily living (ADLs) as running, resting, sleeping, cycling, and exercising are categorized by HAR. The methods to address the HAR issue have developed from the early 2000s' popular machine learning (ML) algorithms [12] to solutions based on deep learning (DL) [9, 12, 19]. Every approach in various technical areas that use big data depends on or is switching to utilizing DL. For example, deep learning and neural networks have been used to tackle natural language processing and voice recognition problems and the processing of signals and images. Consequently, many publications [9, 18, 19] have been published demonstrating deep learning models capable of extracting and selecting features, recognizing human behaviors, and even using identified actions in real-world systems.

As this study focused on HAR that utilizes accelerometer and photoplethysmography sensors, we restrict our study to sensor-based methods. Because most contemporary smartphones have biometric sensors, this kind of HAR is the most extensively studied. According to Chen et al. [7], activity recognition is a complicated task that consists of three basic functions: (1) monitoring an individual's behavior while simultaneously changing the surroundings via the use of suitable sensors and devices; (2) capturing, storing, and processing occurrences using suitable data interpretation and conceptual modeling formalism; and (3) correspondingly, they recommend a way of determining whether an individual is engaging in an action. Frequently, the technique used for one activity is reliant on the approach used for another activity.

As previously stated, researchers categorized HAR into three groups according to the procedures carried out by the participants. The category is comprised of visual, sensor, and radiofrequency sensing. Technologies are all used: regarding sensing processes and tools. Data-driven activity recognition should be contrasted with knowledge-based activity recognition. Models or processes that scrutinize sensor information to identify activities could depend on patterns explored through data analysis or considerable prior knowledge of the subject of interest.

### 2.2 *Deep Learning Approaches in HAR*

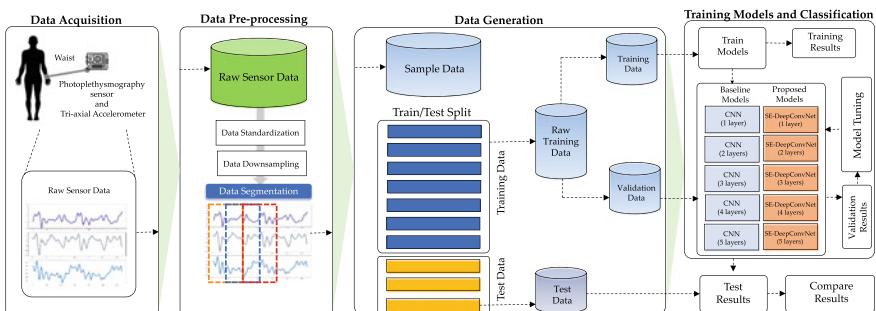
Historically, machine learning methods involved a high level of domain knowledge to execute, much alone develop [5, 11, 15]. The primary challenge with depending

on handcrafted features is the inability to identify a single precise characteristic or set of features that accurately distinguishes all activities [22]. Nevertheless, more sophisticated DL-based classification pipelines for HAR have been developed [19]. DL is distinguished from machine learning because it makes it easier to collect and categorize detailed data from various sensor inputs and modalities [25]. With DL methods, raw or preprocessed data is input into a trained DL model at the end. The classification outcome, including internal feature generation, is output at the other [1].

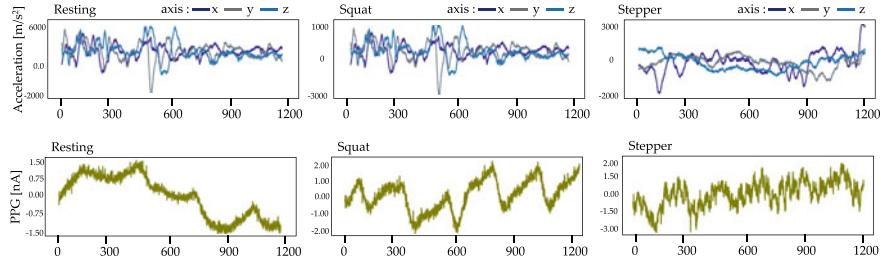
Cross-channel interaction offers several apparent benefits over certain previously suggested literature studies. Hu et al. [10] presented a squeeze-and-excitation (SE) module for calibrating the feedback of channel features. Chen et al. [6] and Dai et al. [8] utilized channel-based attention to perform semantic segmentation and image interpretation, respectively. Wu and He [24] utilized a technique known as group normalization. This approach could be considered as a customized version equipped with channel-specific communication. On the other hand, their interactions between channels are too simplistic, with just the standard deviation and mean of feature maps produced. In the computer vision field, interactions between channels at the same level have been introduced by Yang and colleagues [26], supporting channel communication inside the duplicate layer to improve efficiency.

### 3 The Proposed Framework of Sensor-Based HAR

As shown in Fig. 1, this study's sensor-based HAR methodology composes of four major processes: (1) data acquisition; (2) data preprocessing; (3) data generation; and (4) model training and classification.



**Fig. 1** Sensor-based HAR methodology based on wearable sensors



**Fig. 2** Some samples of accelerometer and PPG data

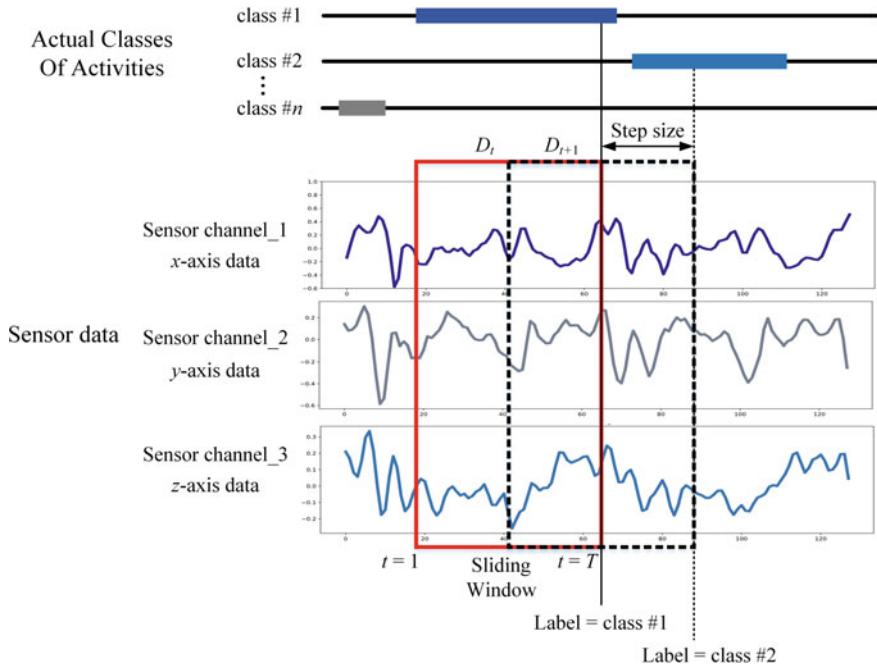
### 3.1 PPG Dataset

We utilized a recently published dataset [3] in this study. The dataset comprises PPG and tri-axial accelerometer data from seven diverse individuals completing five sequences of three specific activities (resting, squat, and stepper). The four signals are collected concurrently at 400Hz sample rate and contain 17,201 s of recorded data. Three men and four women between the ages of 20 and 52 years comprise the seven adult participants.

The PPG and accelerometer information was collected using the Maxim Integrated MAXREFDES100 health sensor platform from the wrist during the voluntary engagement. This platform includes a biopotential analog front-end solution (model: MAX30003 and MAX30004), a pulse oximeter and heart rate sensor (model: MAX30101), two human body temperature sensors (model: MAX30205), a three-axis accelerometer (model: LIS2DH), a three-dimensional accelerometer, a three-dimensional gyroscope (model: LSM6DS3), and an absolute barometric pressure sensor (model: BMP280). PPG signals were collected at the photodetector's ADC output with a 118  $\mu\text{s}$  pulse width, 16 bits resolution, and 8192  $n\text{A}$  full-scale scope illuminated by the green light. The parameters of the tri-axial accelerometer signal corresponding to the MEMS outcome with a ten-bit arrangement, the justified left scale of  $\pm 2 \text{ g}$ , and axes aligned with the  $z$ -axis pointing toward the experimenter's wrist. Figure 2 illustrates several accelerometers and PPG data samples.

### 3.2 Data Preprocessing

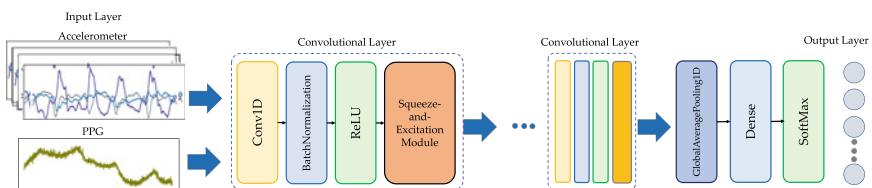
The following manipulations were performed to the raw sensor data preprocessing: standardization and downsampling. The preprocessed sensor data was then segmented using three-second fixed-width sliding windows with a 50% overlap, as shown in Fig. 3.



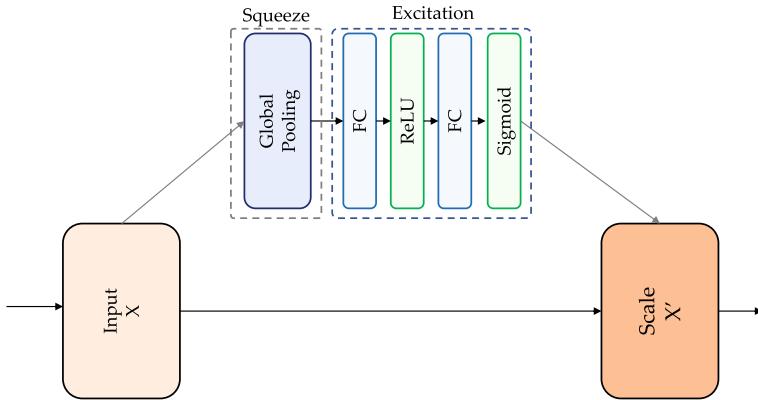
**Fig. 3** Data segmentation by overlapping temporal window

### 3.3 Training Deep Learning Models

The HAR challenge was addressed in this study by proposing a lightweight model of CNN-based deep learning that was relatively easy to implement. SE-DeepConvNet is the deep learning model that has been introduced. It automatically extracts features from data using convolutional layers. It then assesses channel-wise data using squeeze-and-excitation modules, which are both included in the package. Aside from that, we improved recognition outcomes of this model by additional layers of a batch normalization (BN) and a ReLU. As seen in Fig. 4, both layers accelerate network training while simultaneously decreasing vanishing of gradient and overfitting problems.



**Fig. 4** Introduced SE-DeepConvNet model



**Fig. 5** Squeeze-and-excitation module

Hu et al. [10] introduced the squeeze-and-excitation module as a computational unit for all transformations consisting of two major processes: squeeze and excitation. The squeeze process generates channel-wise statistics by utilizing contextual data further than the local receptive field and a global average pool. Following the squeeze process, the aggregated data is subjected to an exciting operation to capture channel-specific dependencies. A conventional gating mechanism is used to accomplish these operations, consisting of both effectively linked layers including ReLU and Sigmoid active functions, as shown in Fig. 5.

### 3.4 Performance Measurement Criteria

To assess the proposed deep learning effectiveness of the algorithm, four conventional evaluation criteria, namely accuracy, precision, recall, and F-measure, are determined through the tenfold cross-validation procedure. These four measurements have the same mathematical formula:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{F-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

The four evaluation metrics listed following are the most frequently utilized performance indicators for HAR in practice. True positive (TP) classes are those that are recognized for the examined group. In contrast, true negative (TN) classes are those that are identified for all other classes. Sensor data from other group could be falsely labeled as relating to that group, arising in a false negative (FN) characterization. In contrast, sensor data of one group could be mistakenly categorized as regarding to another, concluding in recognition of that class as a false positive (FP) characterization.

## 4 Conducting Experiments and the Findings

Providing empirical methods as well as the study results is the goal of this section. The analysis findings are utilized to evaluate the proposed SE-DeepConvNet model for HAR that is based on accelerometer and photoplethysmography data collected during the experiment.

### 4.1 *Experiment Setting*

Among the libraries used in developing the Python programming language are the TensorFlow (v.2.2.0), Scikit-Learn, Keras (v.2.3.1), Pandas (v.1.0.5), and Numpy (v.1.18.5). Several experiments have been conducted to evaluate the recognition of the proposed SE-DeepConvNet and to compare it to conventional CNNs. Every experimentation in this research is carried out using the platform of Google Colab Pro+ with a V100-Tesla as a computing device. Between 1 and 5 layers are regarded to represent the difference between the baseline network and the SE-DeepConvNet.

### 4.2 *Experimental Results*

The research results were obtained using a tenfold cross-validation procedure. As indicated in Table 1, this research conducted assessment series to evaluate the classifier's performance of the baseline CNN models, which were used in this investigation. The recommended SE-DeepConvNet is shown in Table 2 using a variety of metrics, including accuracy, precision, recall, and F1-score, to demonstrate its effectiveness.

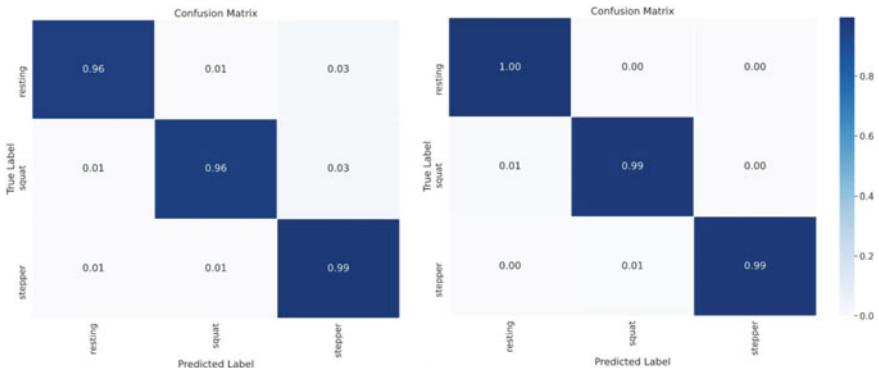
These experimental findings demonstrate that the SE-DeepConvNet surpasses existing deep learning networks with a 99.32% accuracy and a 99.32% F1-score. Comparing the number of convolutional layers in the proposed SE-DeepConvNet network applied in this study, we discovered that the proposed SE-DeepConvNet with

**Table 1** Performance metrics of baseline CNN models using PPG dataset

Model	Accuracy	Precision	Recall	F1-score
1 layer	50.30% ( $\pm 18.28\%$ )	46.35% ( $\pm 25.06\%$ )	50.30% ( $\pm 18.29\%$ )	45.99% ( $\pm 21.36\%$ )
2 layers	97.14% ( $\pm 00.23\%$ )	97.17% ( $\pm 00.21\%$ )	97.14% ( $\pm 00.23\%$ )	97.15% ( $\pm 00.22\%$ )
3 layers	91.13% ( $\pm 17.26\%$ )	88.77% ( $\pm 24.43\%$ )	91.13% ( $\pm 17.26\%$ )	89.43% ( $\pm 22.40\%$ )
4 layers	73.90% ( $\pm 28.20\%$ )	64.36% ( $\pm 39.90\%$ )	73.90% ( $\pm 28.20\%$ )	67.05% ( $\pm 36.60\%$ )
5 layers	96.80% ( $\pm 00.09\%$ )	96.84% ( $\pm 00.14\%$ )	96.80% ( $\pm 00.10\%$ )	96.82% ( $\pm 00.12\%$ )

**Table 2** Performance metrics of SE-DeepConvNet models using PPG dataset.

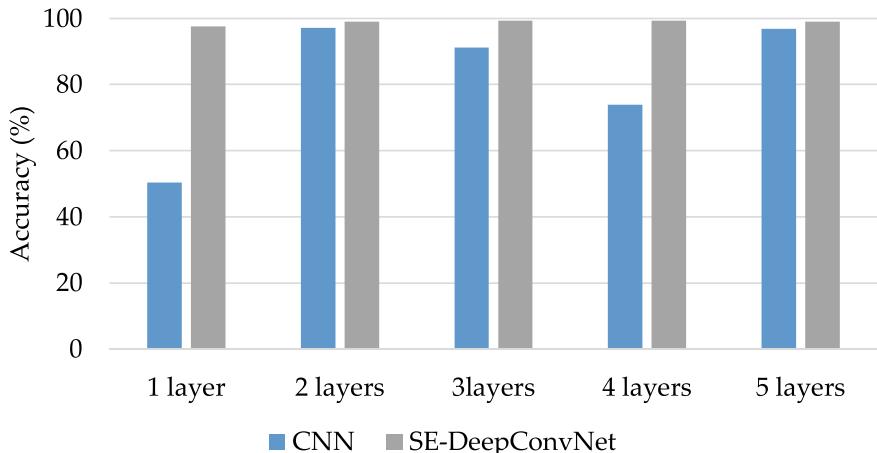
Model	Accuracy	Precision	Recall	F1-score
1 layer	97.55% ( $\pm 0.52\%$ )	97.59% ( $\pm 0.49\%$ )	97.55% ( $\pm 0.52\%$ )	97.57% ( $\pm 0.51\%$ )
2 layers	98.96% ( $\pm 0.52\%$ )	98.97% ( $\pm 0.50\%$ )	98.96% ( $\pm 0.52\%$ )	98.97% ( $\pm 0.51\%$ )
3 layers	99.32% ( $\pm 0.13\%$ )			
4 layers	99.31% ( $\pm 0.30\%$ )			
5 layers	98.96% ( $\pm 0.86\%$ )	98.99% ( $\pm 0.78\%$ )	98.96% ( $\pm 0.86\%$ )	98.98% ( $\pm 0.82\%$ )

**Fig. 6** Confusion matrices of DL model at the same number of layers **a** CNN **b** SE-DeepConvNet

three layers provides the highest results. Figure 6 illustrates the confusion matrices for the CNN and the SE-DeepConvNet with three layers.

#### 4.3 Discussion of the Results

From demonstrated data in Tables 1 and 2, the experimental outcomes evident that the squeeze-and-excitation mechanism added in the convolutional neural networks can increase the recognition performance significantly. Especially, the accuracy of the SE-DeepConvNet model with 1 layer is more than the CNN model with 1 layer 47.25%, as shown in Fig. 7.



**Fig. 7** The recognition performance by adding the squeeze-and-excitation mechanism

## 5 Conclusion and Future Works

According to our findings, the SE-DeepConvNet model, which is a CNN-based neural network that integrates squeezed-and-excitation modules, was proposed in this research. This proposed model uses particular squeeze-and-excitation module benefits to incorporate channel-specific data for each convolutional layer. We validated the proposed deep learning model using a publicly available benchmark dataset for wearable sensors, namely the PPG dataset. The SE-DeepConvNet was advanced than the CNN baseline model in terms of accuracy, with a maximum of 99.32%.

Additionally, future studies may involve validating the proposed DL models using a more comprehensive sample size of individuals with varying exercise activity patterns. Further improvements could be achieved by developing more sophisticated and lightweight deep learning networks and new time–frequency-based data structures.

## References

1. Abdel-Basset, M., Hawash, H., Chakrabortty, R.K., Ryan, M., Elhoseny, M., Song, H.: ST-DeepHAR: deep learning model for human activity recognition in IoHT applications. *IEEE Internet Things J.* **8**(6), 4969–4979 (2021)
2. Alessandrini, M., Biagetti, G., Crippa, P., Falaschetti, L., Turchetti, C.: Recurrent neural network for human activity recognition in embedded systems using PPG and accelerometer data. *Electronics* **10**(14) (2021)
3. Biagetti, G., Crippa, P., Falaschetti, L., Orcioni, S., Turchetti, C.: Human activity recognition using accelerometer and photoplethysmographic signals. In: Czarnowski, I., Howlett, R.J., Jain, L.C. (eds.) *Intelligent Decision Technologies 2017*, pp. 53–62. Springer International Publishing, Cham (2018)

4. Boukhechba, M., Cai, L., Wu, C., Barnes, L.E.: Actippg: Using deep neural networks for activity recognition from wrist-worn photoplethysmography (PPG) sensors. *Smart Health* **14**, 100082 (2019)
5. Chen, J., Chang, J.T.: Applying a 6-axis mechanical arm combine with computer vision to the research of object recognition in plane inspection. *J. Artif. Intell. Capsule Netw.* **2**, 77–99 (2020, May)
6. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.: Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6298–6306. IEEE Computer Society, Los Alamitos, CA, USA (2017, July)
7. Chen, L., Hoey, J., Nugent, C.D., Cook, D.J., Yu, Z.: Sensor-based activity recognition. *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)* **42**(6), 790–808 (2012)
8. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 764–773 (2017)
9. Guo, L., Wang, L., Liu, J., Zhou, W., Lu, B.: Huac: Human activity recognition using crowd-sourced WiFi signals and skeleton data. *Wirel. Commun. Mobile Comput.* **2018**, 1–15 (2018, January)
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
11. Kumar, T.S.: Video based traffic forecasting using convolution neural network model and transfer learning techniques. *J. Innov. Image Process.* **2**, 128–134 (2020, June)
12. Lin, J.J., Mamykina, L., Lindtner, S., Delajoux, G., Strub, H.B.: Fish'n'steps: Encouraging physical activity with an interactive computer game. In: Dourish, P., Friday, A. (eds.) *UbiComp 2006: Ubiquitous Computing*, pp. 261–278. Springer, Berlin, Heidelberg (2006)
13. Mekruksavanich, S., Jitpattanakul, A.: Biometric user identification based on human activity recognition using wearable sensors: an experiment using deep learning models. *Electronics* **10**(3) (2021)
14. Mekruksavanich, S., Jitpattanakul, A.: Deep convolutional neural network with RNNs for complex activity recognition using wrist-worn wearable sensor data. *Electronics* **10**(14) (2021)
15. Mekruksavanich, S., Jitpattanakul, A.: Deep learning approaches for continuous authentication based on activity patterns using mobile sensing. *Sensors* **21**(22) (2021)
16. Mekruksavanich, S., Jitpattanakul, A.: Lstm networks using smartphone data for sensor-based human activity recognition in smart homes. *Sensors* **21**(5) (2021, June)
17. Mekruksavanich, S., Jitpattanakul, A., Youplao, P., Yupapin, P.: Enhanced hand-oriented activity recognition based on smartwatch sensor data using lstms. *Symmetry* **12**(9) (2020)
18. Mutgeki, R., Han, D.S.: Feature-representation transfer learning for human activity recognition. In: 2019 International Conference on Information and Communication Technology Convergence (ICTC), pp. 18–20 (2019)
19. Mutgeki, R., Han, D.S.: A cnn-lstm approach to human activity recognition. In: 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC), pp. 362–366 (2020)
20. Qi, W., Su, H., Yang, C., Ferrigno, G., De Momi, E., Aliverti, A.: A fast and robust deep convolutional neural networks for complex human activity recognition using smartphone. *Sensors* **19**(17) (2019)
21. Schutzer, K.A., Graves, B.: Barriers and motivations to exercise in older adults. *Prev. Med.* **39**(5), 1056–1061 (2004)
22. Vijayakumar, T.: Posed inverse problem rectification using novel deep convolutional neural network. *J. Innov. Image Process.* **2**, 121–127 (2020, June)
23. Warburton, D.E.R., Nicol, C.W., Bredin, S.S.D.: Health benefits of physical activity: the evidence. *Can. Med. Assoc. J.* **174**, 801–809 (2006)
24. Wu, Y., He, K.: Group normalization. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) *Computer Vision—ECCV 2018*, pp. 3–19. Springer International Publishing, Cham (2018)

25. Xu, C., Chai, D., He, J., Zhang, X., Duan, S.: Innohar: a deep neural network for complex human activity recognition. *IEEE Access* **7**, 9893–9902 (2019)
26. Yang, J., Ren, Z., Gan, C., Zhu, H., Parikh, D.: Cross-channel communication networks. In: Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc. (2019)

# A Review Towards Research in Multi-robot Coordination System



M. Pavithra and T. Kavitha

**Abstract** A collection of more than two autonomous moving robots which work together is named as group of mobile robots. The simple robots have to coordinate each other which is important in multi-robot systems. The multi-robots have developed a good potential in many of the applications such as in military works, surveillance in the battlefield, searching of weeds in case of farming and searching for survivors in an accidental area, transportation of vehicles, and foraging techniques. This paper gives information about various coordination methods involved in multi-robot systems including many parameters and is important to achieve the goal and completing the given task.

**Keywords** Coordination method · Multi-robot systems · Communication · Localization · Collision avoidance · Architecture of multi-robotic systems

## 1 Introduction

A huge number of single robots work in group to perform a task is multi-robot system. They are used in task where traditional robots are not well suitable, like disaster area surveillance, contaminant clean-up. Robots should start from home source location, continue in search of target, and return back to source location to do any other tasks. Multiple robots give smaller problems to single robots which make them to communicate with each other to solve tedious problems. The communication concept in multi-robot system has been on study since the inception of distributed robotics research [1]. The multi-robot systems have more potential with large set of applications. It stresses upon many coordination and controlling techniques when robots try to communicate with each other to move from source to destination.

---

M. Pavithra (✉)  
RNSIT, Bengaluru, India  
e-mail: [mpavikanth@gmail.com](mailto:mpavikanth@gmail.com)

T. Kavitha  
AMC College of Engineering, Bengaluru, India  
e-mail: [drtkavitharaj@gmail.com](mailto:drtkavitharaj@gmail.com)

Multi-robot research work is not used widely as there are many issues to be resolved like designing the required environment for any application [2]. The focus of the study is on developing the environment for the foraging application. There are many parameters on this aspect. The review on these parameters helps to design suitable environment and implement algorithms for multi-robot coordination.

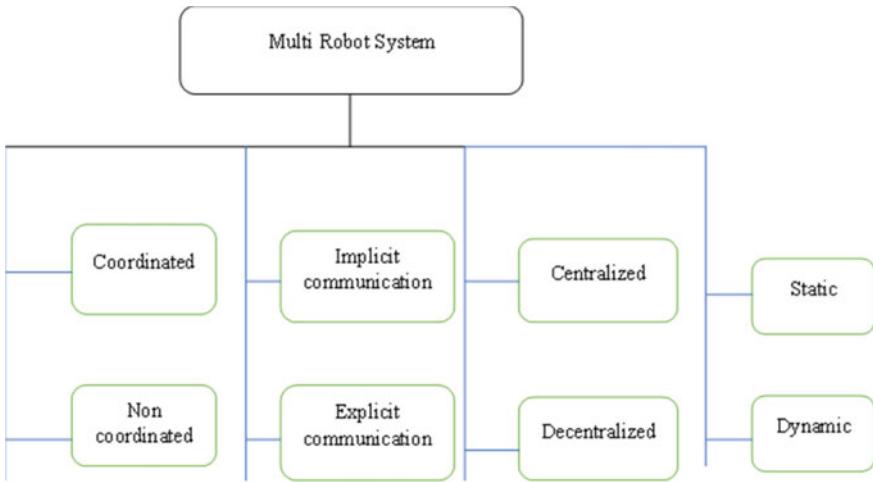
## 2 Multi-robot Systems

Developing a multi-robot system (MRS) does not cost more in comparison with developing a costly single robot which consists of more capabilities. The multi-robot systems have produced more potentiality in many of the applications. It requires more coordination methods and controlling techniques in case where robots communicate with each other in order to move from source to destination in formation of flock of birds or insects using many algorithms for coordination.

In 1986, Reynolds [1] made us aware of three rules which lead to the first computer animation for flocking. The author gave rules for flock centering: trying to stay near the flock mates. Obstacle avoidance: The collisions should be avoided among flock mates. Matching velocity: This makes an effort in matching the velocity in flock mates.

The work done by Chang et al. [3] used gyroscopic force for avoiding collisions. The work proposed checked decentralized law of control that made a group of vehicles to involve specific control to be achieved by avoiding the collision with each other and also the hurdles. A distributed control law technique with a leader robot is introduced by Carpin and Parker [4] to perform well in terms of team and each robot level actions. They helped in implementing a multithreaded based work, which takes care of heterogeneous multi-robot system using many sensors for exchanging information between each other. Canepa et al. [5] suggested flocking algorithm without a leader. In method of asynchronous, the leader among the robots is selected by using a proper algorithm; here, the robot's position is as per the specific formation. The disadvantage of the work done is it allowed the formation only to move in a single straight direction. The other disadvantage of the proposed work is that the leader chosen at once cannot be changed.

The challenges in MRSs are to include strategies for proper coordination between the robots that make robots to do all the tasks better in terms of time and environment used. The techniques used in the work include machine learning methods. Chang et al. [3] gave many topics for research in MRS like biological inspirations, communication methods, various architectures, localization and mapping/exploration, transportation of objects, motion coordination, and many more. Currently, there is no much research carried based on the concepts of the multi-robotic coordination. Agmon et al. [6] referred to three coordination mechanisms in multi-robot system, which includes no coordination, tight coordination, and loose coordination. They could prove that a robot with no coordination did a better performance than a coordinated robot. The tight coordination method is more suitable but tough to practice. Kaminka et al. [6]



**Fig. 1** Multi-robot classification [7]

showed the algorithm selection by using reinforcement learning methods in multi-robot coordination. The proposed work developed a reward function called effectiveness which reduces the time and resources used in coordination, and it increases the time between conflicts. Figure 1 gives the classification of multi-robot system based on some of the parameters [7].

### 3 Parameters for Coordination of Robots

Parameters for coordination of robots are coordination and control techniques, communication, mapping and localization, architecture of MRS.

#### A. Coordination and control techniques

The coordination in multi-robots is the main task of MRSs. The system efficiency is based on the way in which the coordination is obtained. Coordination can be either of the ways, static/dynamic. Static coordination is called as deliberative or offline coordination [8], it mentions the adoption of current environment than prioritizing the environment of the given work. For example, the rules in controlling traffic like stop, keeping right, and distance between target and the robot [9].

The dynamic coordination is called as reactive coordination [10] or online coordination [8], it happens when the task execution completely depends on the analysis of the available information and synthesis of it. The information is given by the way of interaction that can be done. Dynamic coordination is categorized as two types, such as implicit and explicit coordination. Gerkey and Matarić [11] define explicit coordination as a technique based on intentional communication, collaboration methods

are used to work with high ability robots, and implicit coordination is based on the techniques which uses the dynamic exchange of information among the robots and based on environment to obtain better performance. Explicit coordination consists of explicit communication, due to active behavior of a robot. Implicit coordination is mainly involved with implicit communication. Combination of explicit and implicit communication information will help in betterment of the performance using coordination in the complete robotic system. The static way manages complex tasks, but controlling of it in real time would be difficult. The dynamic way in real-time is more capable, but it is difficult in dealing the complex tasks. For multiple mobile robot system (MMRS), the outer environment would be tedious which is not known. The best thing is to use the combination of both methods accordingly, based on the particular characteristics of the task.

#### B. Communication

There is a difference between implicit/explicit communications, implicit communication is predominant in social insects like ants. Implicit communication is through the environment. Explicit way of communication is observed in honey bee workers. This is a better method to be used when the number of robots is less and during speed of execution is considered more, for example, when a dangerous situation occurs, counter measure has to be taken. McPartland et al. [12] implemented on two different swarms of robots to check the difference between communications by giving a known environment in small time duration. Paul et al. [13] presented strategies on system of simple communication that was implemented on implicit/explicit communication.

#### C. Localization

A robot has to move in given environment, so building of a map in given environment is the most important job. The mapping is about the environment of the robot which is not known before. It involves more timing in convergence and formation of a flock pattern. A multi-robot system will make calculation of the coordinates for given environment. Finding the proper location of itself is localization. To construct the map in centralized technique, the information by all the robots is necessary [14, 15]. In decentralized technique, the process of sub-map of a robot is not dependent on other robots [16–18]. The difficulty in centralized approach is single robot can create the failure in mapping, the complete system pauses even if other robots are in good condition. The decentralized approach is stronger and can withstand the faults as robots provide the local sub-map.

#### D. MRS Architecture

Multi-robot systems develop a good coordinated behaviors by making use of the environment provided. While multi-robot system is designed, many things are included like the group architecture [19]. The complete communication details [20], the abilities of a robot and the other models [21] are to be known to find if the controlling system is decentralized or centralized. In centralized method, a single robot provides directions to all robots which help in performing cooperative task. This architecture

**Table 1** The symmetry of parameters [7]

Parameters	Objective
Coordination and control	To check how the coordination approach is useful in controlling the robots, centralized
Communication	It explains the way robots exchange information, implicit or explicit, or decentralized
Mapping and localization	The information about known and unknown environment of robot
Architecture of MRS	The kind of architecture which is efficient for multi-robots

is considered as more efficient in controlling group of robots [22]. In group architecture, a computer system with good vision ability can check all the positions of robot, perform tasks, and help to control every robot. The control action of every robot in a system by any external controller will be tedious in case the colony size is increased [20]. The insect societies cannot have a single agent who can control the complete system in case of decentralized architecture and are considered on behavior-based control on multiple robots [20]. Decentralized architectures are distributed whereas all robots are equal with a control system, in locally centralized architecture [23–25].

Many of the coordination algorithms are proposed, they are leaderless and with leader approach. The observation is that the approach without a leader is more compatible in cooperation when compared to greater number of autonomous robots and heterogeneous robots, they do not make the leader more responsible in controlling group of robots. Table 1 refers to the comparison of parameters [7] and gives information about the objective of these parameters. Our contribution includes multi-robot coordination in performing a task and motion planning. We analyze the communication parameters like, composition, coordination techniques, environment, architectures, etc., for coordination approaches in many applications using MRS.

## 4 Conclusion

The communication between robots in an MRS system is major discussion in research community. For tasks such as obstacle detection and collision avoidance, explicit communication is more suitable when an immediate reaction is required. Situations in real time or where an immediate response is not needed, we can go for implicit communication, example pattern formation, path planning, etc. The new approach is one which includes both implicit and explicit communication. To know, the environment of the application is a must for a robot. If it is a case of dynamic environment, robots should do mapping and also self-localization is required. Some of the methods use robots which do localization and mapping at the same time (SLAM) [26]. A group of robots will increase the speed of performance by the use of parallel SLAM. Individual robot will create the sub-map and that is combined to form global map for the

environment. In a multi-robot system, the group architecture is of high importance. The applications based on real time can use centralized architectures, and these are preferred than decentralized architecture.

Not much research in this area has been carried out in the laboratories. The demonstrations are more as it is still uncommon to have implementations involving more than a dozen of physical modules.

## References

1. Arai, T., Pagello, E., Parker, L.E.: Editorial: advances in multi-robot systems. *IEEE Trans. Robot. Autom.* **18**(5), 655–661 (2002)
2. Chowdhary, V., Sane, T.: Reusable Software Components for Multi-robot Foraging. Worcester Polytechnic Institute (2018)
3. Chang, D.E., Saber, R.O., Marsden, J., Shadden, S.: Collision avoidance for multiple agent systems. In: Proceedings of the IEEE Conference on Decision and Control, Maui, Hawaii USA, Dec 9–12, pp. 539–543, 2003
4. Carpin, S., Parker, L.E.: Cooperative leader following in a distributed multi-robot system. In: Proceedings of IEEE International Conference on Robotics and Automation, pp. 2994–3001, 2002
5. Potop-Butucaru, M.G., Canepa, D.: Stabilizing and flocking via leader election in robot networks. In: Proceeding of 9th International Symposium on Stabilization, Safety, and Security of Distributed Systems, Nov 14–16, pp. 52–66, 2007
6. Kamimura, G., Noa, A., Elmaliach, Y.: Ann. Math. Artif. Intell. **57**(3), 293–320. <https://doi.org/10.1007/s10472-010-9193-y> SourceDBLP
7. Verma, J.K., Ranga, V.: Multi-robot coordination analysis, taxonomy, challenges and future scope. *J. Intell. Robot. Syst.* **102**, 10 (2021)
8. Todt, E., Suárez, R., Raush, G.: Analysis and classification of multiple robot coordination methods. In: Proceedings of ICRA'00, San Francisco, CA, USA, April 2000, pp. 3158–3163
9. Gerkey, B.P., et al.: The player/stage project tools for multi-robot and distributed sensor system. In: Proceeding, International Conference on Advance Robotics, pp. 317–323, 2003
10. Iocchi, L., Salerno, M., Nardi, D.: Reactivity and deliberation, a survey on multirobot systems. *Lect. Notes Comput. Sci.* **2103**, 9–32 (2001)
11. McPartland, M., Abbass, H.A., Nolfi, S.: Emergence of communication in competitive multi-agent systems—a Pareto multi-objective approach. In: Proceedings of GECCO, Washington DC, USA, June 25–29, pp. 51–58, 2005
12. Rybski, P.E., Veera Raghavan, H., Larson, A., Gini, M., LaPoint, M.: Communication strategies in multi-robot search and retrieval, experiences with MinDART. In: Distributed Autonomous Robotic Systems (Springer), Japan, 2007
13. Fenwick, J.W., Leonard, J.J., Newman, P.M.: Cooperative concurrent mapping and localization. In: Proceedings of the 2002 IEEE International Conference on Robotics and Automation, pp. 1810–1817, 2002
14. Williams, S.B.: Efficient solutions to autonomous mapping and navigation problems. Ph.D. Dissertation, University of Sydney, 2001
15. Rodriguez-Losada, D., Matia, F., Jimenez, A.: Local maps fusion for real time multirobot indoor simultaneous localization and mapping. In: Proceedings, IEEE International Conference on Robotics and Automation, 2004
16. Koller, D., Thrun, S., Ng, A.Y., Liu, Y., Durrant-Whyte, H., Ghahramani, Z.: Simultaneous localization and mapping with sparse extended information filters. *Int. J. Robot. Res.* **23**(7–8), 693–716 (2004)

17. Cao, Y.U., Fukunaga, A.S., Kahng, A.: Cooperative mobile robotics: antecedents and directions. *Auton. Robots* **4**, 1–23 (1997)
18. Arkin R.C.: Behaviour-Based Robotics, pp. 393–409. The MIT Press, USA (1998)
19. Bekey G.A.: Autonomous Robots From Biological Inspiration to Implementation and Control, pp. 294–314. The MIT Press, London, England (2005)
20. Kawauchi, Y., Fukuda, T.: Cellular robotic system as one of the realizations of self-organizing intelligent universal manipulator. In: Proceedings of IEEE International Conference on Robotics and Automation, Cincinnati, OH, May 13–18, pp. 662–667, 1990
21. Arkin, R.C., Nitz, E., Balch, T.: Communication of behavioural state in multi-agent retrieval tasks. In: Proceedings IEEE/RSJ International Conference on Robotics and Automation, Atlanta, GA. (ICRA'93), pp. 588–594
22. Dorigo, M., et al.: Evolving self-organizing behaviours for a Swarm-robot. *Auton. Robots* **17**, 223–245 (2004)
23. Beni, G.: From swarm intelligence to swarm robotics. In: Sahin, E., Spears, W. (eds.) *Swarm Robotics Workshop: State-of-the-Art Survey*, vol. 3342, pp. 1–9. Springer, Berlin Heidelberg (2005)
24. Simmons, R., Burgard, W., Apfelbaum D., Fox, D., Moors, M., Thrun, S., Younes, H.: Coordination for multi-robot exploration and mapping. In: Proceedings of the AAAI National Conference on Artificial Intelligence, Austin, TX, 2000
25. Jacobs, I.S., Bean, C.P.: Fine particles, thin films and exchange anisotropy. In: Rado, G.T., Suhl, H. (eds.) *Magnetism*, vol. III, pp. 271–350. Academic, New York (1963)
26. Xu, Z., Rong, Z., Wu, Y.: A survey: which features are required for dynamic visual simultaneous localization and mapping? *Vis. Comput. Ind. Biomed. Art* **4**, 20 (2021). <https://doi.org/10.1186/s42492-021-00086-w>

# Feature Extraction and Representation Learning via Deep Neural Network



**T. Anuradha, Arun Tigadi, M. Ravikumar, Paparao Nalajala, S. Hemavathi, and Manoranjan Dash**

**Abstract** Selection of a text characteristic is a prerequisite for text mining and information retrieval. Traditional techniques of feature extraction demand the use of custom features that must be made by hand. For new applications, deep learning allows the acquisition of new effective feature representations from training data rather than having to spend a lengthy time developing an effective feature by hand. Deep learning has made tremendous strides in text mining as a new feature extraction technique. This means that instead of using handmade features that strongly rely on designers' prior knowledge and cannot be fully utilised with big data, deep learning employs features from massive datasets. Massive datasets, containing millions of parameters, can be used to train deep learning models automatically to represent those datasets' features. This work initially describes the most prevalent text feature extraction approaches and then goes into greater depth on how deep learning is regularly utilised in text feature extraction, as well as how it can be used in future.

**Keywords** Deep learning · Text feature · Massive datasets · Feature representation

---

T. Anuradha (✉)

Department of Electrical and Electronics Engineering, KCG College of Technology, Chennai, India

e-mail: [tanura1872@gmail.com](mailto:tanura1872@gmail.com)

A. Tigadi

K.L. E Dr M.S.Sheshgiri College of Engineering and Technology, Constituent College of KLE Technological University, Hubballi, Belagavi Campus, Hubballi, India

M. Ravikumar

Department of Mechanical Engineering, New Horizon College of Engineering, Bangalore, India

P. Nalajala

Department of ECE, Institute of Aeronautical Engineering, Hyderabad, India

S. Hemavathi

Battery Division, Central Electrochemical Research Institute (CECRI), Chennai, India

M. Dash

Faculty of Management Sciences, Siksha O Anusandhan (Deemed to be University), Bhubaneswar, India

## 1 Introduction

Depending on how the data is represented and interpreted, some information processing activities might be easy or difficult. Although 210 by 6 may be easily divided using the long division method, the situation becomes more problematic if the digits 210 and 6 are converted to roman numerals as CCX divide by VI. Most people will convert CCX to Arabic numeral form first and then utilise that to begin the division step. This broad approach can be applied to a wide range of fields, including daily life, computer science, and deep learning in particular [1]. In this sense, supervised learning-trained feedforward networks do representation learning. A linear classifier, such as SoftMax regression, is employed as the network's last layer [2]. The rest of the network teaches the classifier how to represent itself. Consequently, supervised training improves the classification process by giving the representation at each hidden layer new attributes. Classes that were not previously linearly separable, for example, could become so in the last hidden layer. Finally, a model like a nearest neighbour classifier might theoretically be used as the final layer. In the penultimate layer, depending on the type of final layer, different qualities should be learned by the features there. No explicit conditions are imposed on learned intermediate features during supervised training of feedforward networks. Algorithms for other types of representation learning are generally expressly created to mould the representation in a specific way. Take for instance learning a representation that simplifies density estimation. To make modelling easier, we can devise an objective function that promotes the representation vector's elements to be more self-sufficient. Unsupervised deep learning algorithms, like supervised networks, develop a representation as a by-product of their primary training objective [3]. It makes no difference how anything is represented when it comes to communication. You can learn more than one task simultaneously with a shared internal representation (some supervised, some unsupervised). Unsupervised and semi-supervised learning can both be accomplished via representation learning. We frequently have a lot of unlabelled data and a little amount of labelled data for training. Experiments on the labelled subset using supervised learning approaches frequently lead to overfitting. Unlabelled data can be used in semi-supervised learning to help alleviate the overfitting issue. If the unlabelled data is properly represented first, we can use it to solve the supervised learning problem [4].

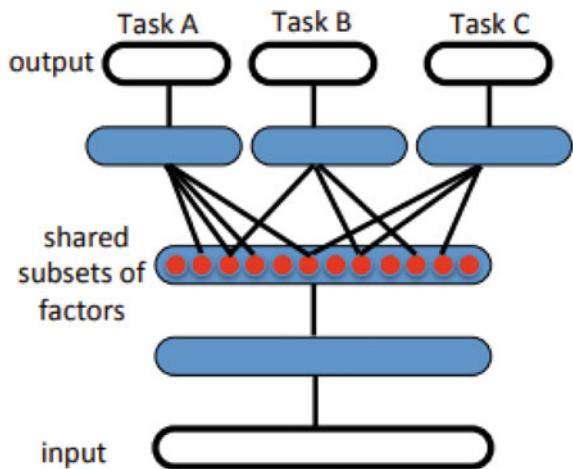
## 2 Learning Representation

A new conference, ICLR1, has been created to focus on representation learning, which has now become a separate subject within the machine learning community. As we will see in the next section, learning a representation is an important part of the storey, so framing the problem as one of learning a representation makes sense, while representation learning research has exploded, an impressive run of empirical

achievements has occurred both in academia and in industry, fuelling the recent growth in the field [5]. Some of the high highlights will be highlighted shortly here.

- **Speech recognition and signal processing:** Speech recognition was one of the first applications of deep learning networks. There has been a recent reversal in neural networks, deep and representation learning, and this has had a significant impact on speech recognition, with many companies releasing new versions of their MAVIS speech systems based on deep learning. Word mistake rates were cut in half compared to a previous model using Gaussian mixtures and training on the same amount of data as today's state-of-the-art model. The state of the art in polyphonic transcription has been significantly surpassed by representation learning algorithms in music, with relative error improvements ranging from 5 to 30% on a common benchmark of four datasets.
- **Object recognition:** This dataset had a 1.4% error rate while using support vector machines (SVMs). This dataset was used to kick-start deep learning in 2006. There is now a 0.81% inaccuracy in MNIST's knowledge-free version, which is state of the art in terms of unconstrained challenges (e.g. employing a convolutional architecture). Deep networks continue to hold the most recent records in this field. Recently, deep learning has progressed beyond just recognising digits to recognising objects in photographs, with the most recent accomplishment occurring on the ImageNet dataset, where the error rate was decreased from 26.1 to 15.3%.
- **NLP:** In addition to speech recognition, representation learning has numerous other uses in natural language processing (NLP). Word embeddings are distributed representations that are learned for each word. It was only by including a convolutional architecture and semantic role labelling that the SENNA system was established, which shares representations across a variety of tasks like as language modelling and part-of-speech tagging. SENNA is as good as or better than current state-of-the-art predictors, yet it is simpler and faster. When learning word embeddings and visual representations, it is possible to link the two together. As a result of using enormous volumes of data to map photographs and searches in the same area, in a short period of time, Google's picture search has grown substantially. This new neural net language model could outperform current state of the art in terms of both perplexity and speech recognition word error rate, decreasing i. Perplexity and BLEU scores have been boosted using statistical machine translation models similar to these (SMT). It was found that researchers could beat the present state of the art in whole sentence paraphrase identification by nearly doubling the F1 score using recursive autoencoders. On improve word sense disambiguation accuracy, representation learning can be applied to a subset of Senseval-3 and see an increase in accuracy from 67 to 70%.
- **Multi-task and transfer learning, domain adaptation:** When a learning algorithm can use commonalities across distinct learning tasks to share statistical strength and transfer knowledge between them, this is referred to as "transfer learning" (Fig. 1).

**Fig. 1** Explaining representation learning



## 2.1 Advantages of Representations

- **Distributed representations**

As long as a learnt representation is large enough, it can capture many different possible input combinations. To represent huge input regions with a limited number of parameters, there are a number of strategies including RBMs, sparing coding, and neural networks with several layers (where  $k$  denotes the number of nonzero components in a sparse representation and  $N$  denotes the number of zero components). All of these representations are dispersed or sparse. As a result of multi-clustering, the generalisation of clustering to distributed representations can be thought of as a form of object recognition that uses a histogram of cluster categories to find similar objects across different patches of an image [6]. This is an extremely popular method for extracting hierarchical features for object recognition. For example in sparse code or with a restricted Boltzmann machine, each parameter can be reused in many other cases that are not merely next to each other. However, with local generalisation, separate parts of input space are basically associated with their own unique set of parameters. The number of hidden units or features in a distributed representation can be activated by a single input, and this number grows exponentially over time. In a single-layer model, an input hyperplane corresponds to a code or representation for each feature, and the pattern of activation for that input corresponds precisely to the code or representation for that input [7]. The most common clustering algorithm does not employ a non-distributed representation, such as k-means, which uses a one-hot code to decide which of several cluster centroids best reflects an input vector.

- **Depth and abstraction**

In this study, we discuss representation learning strategies in depth, which is an important consideration. As we will see, deep architectures are notoriously difficult

to train properly, despite recent advances in the field. Despite these difficulties, deep architectures provide two key advantages that keep us interested in finding new training procedures.

- ***Disentangling factors of variation***

We want our representations to be distributed and invariant, but we also want them to separate the variables that cause variation. The input distribution tends to change independently of other explanatory components when studying a sequence of real-world inputs.

Many sources interact richly to provide complex data. Object classification, for example, might be made more difficult by the interactions of these variables. Object shapes and material properties interact to create a picture, for instance. All of these elements come together to form an image. Complex patterns can be produced when shadows from items in a scene fall on top of one another, giving the impression of object boundaries when there are not any. It is our belief that to overcome these problems the method we use will have to rely on a large number of unlabelled samples to create models that distinguish among the various explanatory sources [8]. As a result, a representation for AI-related tasks should be significantly more resilient to the complex and richly structured variations that can be found in natural data sources.

- ***Good criteria for learning representations***

The production of a clear training objective or target is one of the challenges of representation learning. Other machine learning activities such as categorisation cannot do this, thus, it is distinct from them. In dealing with classification, it is obvious (at least conceptually) that we want to limit the amount of wrong classifications. Representation learning has no connection to the final goal of learning a classifier or other predictor, which is frequently the case. This is similar to the credit assignment problem that can be seen in reinforcement learning programmes [9]. A good representation, according to our theory, separates out the fundamental causes of variation, but how can we put it into practise? We can incorporate priors like those described above (potentially data-dependent ones) that assist the representation better do this disentangling, even if we do not optimise the likelihood under a decent model.

## **Deep representations**

Feature learning and deep learning saw a breakthrough in 2006, thanks to Geoff Hinton, Lee, and a slew of other researchers. Gluttonous layer-by-layer unsupervised pre-training was a key concept. In order to train a feature hierarchy, it was necessary to combine previously learned transformations with unsupervised feature learning iterations, each of which contributed weight to the deep neural network one step below. Lastly, a neural network classifier or a deep Boltzmann machine might be created by combining the layers to create a deep supervised predictor.

There was still a significant difference between the unsupervised pre-training results and the results obtained with no pre-training. You could use a previous layer's

results as new inputs for a subsequent layer (on top of the raw input). Iterative pre-training is another option, which involves pre-training all previously added layers in a supervised manner at each stage of the iteration.

Using an unsupervised model, integrating pre-trained layers from unsupervised learning is less evident than combining single layers to produce an improved model [10]. First, pre-trained RBMs in DBNs were presented as the top layer of a DBN, with the lower layers being read as a directed sigmoid belief network and the lower layers as an RBM. This generative model could be improved further, but it is unclear how.

### 3 Text Feature Extraction Methods

Extracting text features is critical since it has a direct impact on text classification accuracy. A sentence is seen as a dot in N-dimensional space in vector space model (VSM). Each dot's datum dimension indicates a different (digitised) text characteristic. In addition, keyword sets are frequently used in text features. Meaning that a set of predetermined keywords is used to compute the weights of the textual terms, and a digital vector is then formed, which is the text's feature vector. Methods for extracting text features that are already available include those described below, such as text filtration, fusion, mapping, and clustering.

#### A. *Filtering method*

Fast and efficient filtering is the best method for extracting text features on a big scale. Word frequency, information gain, and a mutual information strategy are among the text feature extraction filtering strategies used.

- **Word frequency:** To measure a word's frequency, you count how many times it appears in a passage of text. Using word frequency to pick features reduces the dimensionality of feature space by excluding words with frequencies below a predetermined threshold. Words with low frequency have little effect on filtration, which is why this strategy is based on it. Information retrieval researchers, on the other hand, believe that words with a lower frequency of occurrences can occasionally hold more information. As a result, in the feature selection process, deleting large numbers of terms based only on their frequency is improper.
- **Mutual information:** Computational linguistics models often use the mutual information (MI) method for measuring the mutuality of two objects. It is used in filtration to check for feature distinction across different themes. Mutual information and cross-entropy have similar definitions. It is usual practise to count the number of mutual terms shared by a feature word and a class in order to estimate the amount of mutual information the two have. Nothing is presupposed about the link between feature words and classes in this strategy. Hence, it is ideal for registering text classification features and class descriptions in data bases.

- **Information gain:** There are many machine learning techniques that use information gain (IG) [11]. To determine how much of the topic's projected material is actually included in the text, look for a well-known feature inside a text on the subject. Computing information gain allows researchers to identify qualities that are more common in positive samples than negative ones. There are numerous mathematical ideas and sophisticated theories and formulas involving entropy involved in the evaluation approach known as information gain. The quantity of information a feature item can supply without considering the entropy of any other features, but the difference in entropy values between features is described as the item's ability to provide overall categorisation information. Items with little information gain are deleted, and the rest are sorted in descending order using the information received from each feature item. This is done using training data.

### B. Fusion method

Fusion necessitates the use of specialised classifiers, and the search must be undertaken with a time interval that increases exponentially [12]. There is a lot of variability in terms of timing. As a result, it should not be used to extract features from big texts. Fusion techniques that use the weighting method fall under a distinct category. It assigns a weight (0, 1) to each feature so users can practise using it while making tweaks. The linear classifiers' weighting mechanism is quite efficient. Example-based learning methods like the K-closest neighbours (KNN) algorithm.

- **Weighted K-nearest neighbours (KNN):** As part of the KNN classifier weighted feature extraction challenge, Han applied several of his earlier ideas. For each categorisation of continuous cumulative data, the approach has a strong classification influence. KNN's absence of parameters and statistical pattern recognition-based text categorisation capacity may lead to higher accuracy and recall rates in classification.
- **The centre vector weighted method:** It is suggested by Shankar that a weighted centre vector classification approach be used, which first establishes a method of characterising abilities to distinguish between right and wrong and then generates a new centre vector. Algorithm requires numerous weighted techniques

### C. Mapping method

Text classification has used mapping frequently and successfully. In latent semantic index (LSI) and PCA, it is commonly employed.

- **Latent semantic analysis:** It is a theory or method of computation used to acquire and demonstrate knowledge [13]. There is no link between words and texts because of the statistical computation approach used to evaluate a large number of text sets. This statistical computation method then utilises the latent semantic structure extracted from the text sets to represent the words and texts. By mapping texts from high-dimensional VSM to lower-level latent semantic space, the basic premise of latent semantic analysis is established. **Least squares mapping method:** Jeno studied high-dimensional data reduction from the centre vector and least squares perspectives.

#### D. *Clustering method*

Text feature comparison is taken into account in the clustering process because it is critical to text feature comparison. As a result, the core of each class is used to replace the class's individual features. As a result of its low compression ratio and steady categorisation accuracy, this approach has several advantages. It has the drawback of being exceedingly time-consuming.

- **CHI (chi-square) clustering method:** Instead of the usual algorithm's pattern of each word having a corresponding one-dimension, CHI clustering computes the contribution of each feature word to each class and groups those words together. This approach has the advantage of being quite simple to implement.
- **Concept indexing:** Text categorisation uses a basic but effective dimensionality reduction technique called concept indexing (CI). A basic vector structure subspace (CI subspace) is utilised for each class's centre, and each text vector is mapped to that subspace to represent it. When there is more classification contained in training sets than in the text vector space, it reduces vector space dimensionality because the CI subspace is smaller in dimensionality [14]. Text vector mapping can be viewed as an indexing procedure in this concept space, with each class centre serving as a generalisation of text contexts within that classification.

## 4 Deep Learning Approach

Hinton et al. introduced a new category of unsupervised learning in 2006 called “deep learning”. Studies on artificial neural networks inspired the idea for this project. A deep learning structure is a multi-layer perceptron with many implicit layers. Dispersed feature representation can be found via deep learning, which uses lower-level attributes to create higher-level property categories or features.

In contrast to surface-based algorithms, which have a number of advantages, deep learning algorithms have numerous disadvantages, such as a limited capacity to generalise for challenging classification tasks when using few samples of complex function. If the input data is characterised according to their distribution, then the implementation of complex function approximation is called deep learning. However, while dealing with samples, the essence of each dataset's feature is rarely studied. Instead of using handmade features, deep learning automatically learns new features from huge data, which makes it superior to standard pattern recognition approaches. Only one well-known good feature has developed in the history of computer vision progress in the last five to ten years. Semantic parsing, retrieval, semantic role labelling, sentimental analysis, question answering, machine translation (including named entity recognition), text classification (including summarisation), and text generation are all common natural language processing (NLP) tasks in which deep learning technology is used. Two prominent models used in this study are the convolution neural network and the recurrent neural network.

Following that, a number of text feature extraction methods, enhancement methods, and stages are discussed.

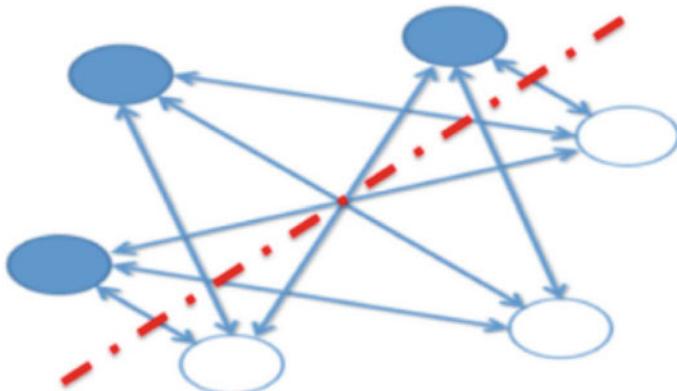
### A. Autoencoder

Using Rumelhart et al. autoencoder's as a feedforward network, researchers were able to learn a compressed and distributed representation of data for the first time. The input and output of an autoencoder are often separated by a secret layer. The hidden layer's representation is smaller than either the input or output layers' because it contains fewer units. It is possible to train an autoencoder without supervision by feeding it the same input data over and over again until you get the desired results. The training procedure is identical to that of a typical neural network with backpropagation, with the exception of the error, which is calculated by comparing the output to the input data. The deep counterpart of an autoencoder is a stacked autoencoder, which is constructed by stacking layers on top of each other. There are several layers in a neural network, and each one takes in the previously learnt representation as input and output. Gravelines et al. discussed a stacked sparse autoencoder, which is an autoencoder with regularisations for sparsity to learn a sparse representation.

### B. Restricted Boltzmann machine

Restricted Boltzmann machine (RBM) is a Boltzmann machine invented by Smolensky that has no connections between any of its visible or hidden units. It was originally known as harmonium. Visible units (i.e. data samples) are part of this network, as are some concealed ones (correspondingly visible vectors) (correspondingly hidden vectors). Binary velocities, such as the visible vector and the hidden vector, have states of 0 or 1. A bipartite graph represents the entire system. When comparing visible and hidden units, edges are only found between them; otherwise, no edge connections exist (Fig. 2).

### C. Deep belief network



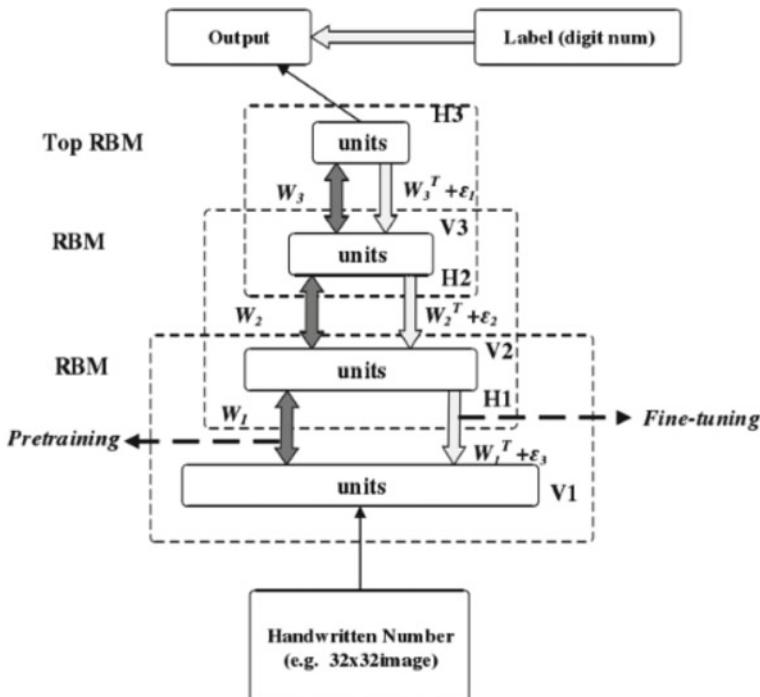
**Fig. 2** RBM structure

Since greedy RBMs may be taught, Hinton et al. proposed DBNs. In DBN's network topology, there is a layer on top of the layers that contains one of the constrained Boltzmann machines.

DBN's training paradigm is broken down into two stages: implementation and evaluation.

- RBM networks should be trained independently and without supervision for each layer to ensure that feature information is preserved as vectors are mapped to various feature spaces.
- Input feature vectors from RBM are used as input feature vectors in the BP network, which then trains an entity relationship classifier under supervision using the output feature vectors from RBM. Layer-specific RBM networks may only optimise the weights of the feature vectors in their own layer, not the feature vectors of the entire DBN. The entire DBN network is fine-tuned by an RBM backpropagation network, which sends error information to each tier of RBM.

In deep learning terminology, step one is referred to as pre-training, while step two is referred to as fine-tuning. In the supervised learning layer, any classifiers based on a given application domain can be employed. BP networks are required to be used (Fig. 3).



**Fig. 3** Structure of DBN

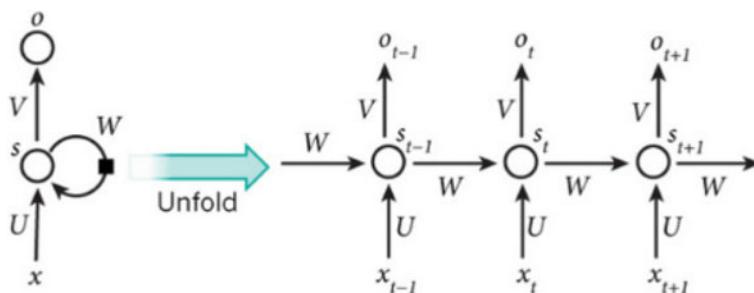
#### D. Convolutional neural network

As a new and extremely effective identification approach emerges, convolution neural network (CNN) has attracted considerable attention from researchers. Hubel and Wiesel proposed the idea of a receptive field in the 1960s based on their study of the visual cortex cells of cats. Fukushima was moved to provide neuropsychological ideas in the first deployment of the CNN network, and he also believed that a wild notion was first implemented in the artificial neural network sector. When it comes to pattern recognition, LeCun et al. found that the error gradient algorithm training in the convolutional neural network produced the best results when compared to other methods.

CNN is a type of artificial neural network because of its versatility and ability to extract local features from large amounts of input. Using shared network structure weights makes it more like biological neural networks, reducing network complexity by reducing weight numbers and allowing CNN to be employed in a variety of pattern recognition applications with excellent results. Shared network structure weights. The results were excellent. Combining local perception area with CNN ensures displacement invariability by sharing the weight and dropping the sample in space or time. This allows the data to be used to its greatest potential. There are other more applications for CNN that have been discovered over the course of many years of research, including as the identification of faces, documents, speech, and licence plates. By using permutation encoding technique, Kussul could identify faces, recognise handwriting digitally, and recognise objects with a certain level of performance in 2006.

#### E. Recurrent neural network

Sequential data is processed using RNNs. There are three layers in a typical neural network model: input layers, hidden layers, and output layers. The nodes in these layers are all disconnected from one another. For occupations requiring sequential inputs, such as voice and language, RNNs (Fig. 4) are usually preferred. They kept a “state vector” in their hidden units when working with input sequences, containing information about the prior history of each preceding component. Hidden unit outputs



**Fig. 4** RNN circuit

must be trained with backpropagation because they are analogous to the outputs of individual neurons in a deep multi-layer network.

Because the backpropagated gradients either rise or decrease at each step, it is been difficult to train RNNs because their dynamic capabilities often explode or vanish.

Hidden units with values  $s_t$  in prior time steps are fed into the artificial neurons (shown on the left by the black square, which represents a one-step delay in time). An input sequence of elements ( $x_t$ ) can be converted into an output sequence of elements, where the components in each  $o_t$  rely on the input sequence (for  $t'$ ), using this technique. The number of time steps is split by three, and the same set of parameters are used in each time step. A number of new RNNs have been developed, including the simple RNN (SRN), the bidirectional RNN, the deep (bidirectional) RNN, and the echo state network.

## 5 Conclusion

For text mining and information retrieval, the selection of a text feature item is a prerequisite step. If an extract metric, such as a reduction in the dimension of feature vector spaces, is met, it is applicable to initial feature subsets from test sets. Uncorrelated or unnecessary features will be removed during feature extraction. Feature extraction, as a preprocessing strategy for the learning algorithm, can enhance the learning algorithm's accuracy while decreasing training time. If you compare deep learning to other machine learning methods, you will find that the former can detect more complex interactions between features, learn lower-level features from nearly unprocessed original data, and mine characteristics that are difficult to detect. Although the recurrent neural network (RNN) has been widely employed in natural language processing (NLP), it is rarely used in text feature extraction for the simple reason that RNN focuses on time-sequenced input. The generative adversarial network model, first introduced in 2014 by Ian J. Goodfellow, has also achieved noteworthy accomplishments within the deep learning generative model field in under two years. It presents a novel frame that may be utilised to estimate and build an opponent process model compared to previous algorithms. This is a significant advancement in the field of unsupervised representation learning. It is now primarily used to create natural-looking photographs. However, in terms of text feature extraction, it has made little progress. Deep learning has some problems. In order to support supervised perception as well as reinforcement learning, significant volumes of data are required. Our dataset on diabetes now has data from 302 hospitals, and this data will let us employ deep learning in text feature extraction to better deal with medical issues. And, they are terrible at advanced plans, able to do nothing but very basic pattern discrimination tasks. Having unreliable, inaccurate, and unjust data necessitates further development in future. As a result of text feature

extraction's inherent properties, each approach has both advantages and insurmountable limitations. If at all possible, use a variety of extraction methods to get at the same piece of data.

## References

1. Peng, S., Sun, S., Yao, Y.-D.: A survey of modulation classification using deep learning: signal representation and data preprocessing. *IEEE Trans. Neural Netw. Learn. Syst.* <https://doi.org/10.1109/TNNLS.2021.3085433>
2. Jia, X., et al.: Semi-supervised multi-view deep discriminant representation learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**(7), 2496–2509 (2021). <https://doi.org/10.1109/TPAMI.2020.2973634>
3. Jiang, W., et al.: Statistical feature extraction and hybrid feature selection for material removal rate prediction in chemical mechanical planarization process. In: 2021 5th IEEE Electron Devices Technology & Manufacturing Conference (EDTM), 2021, pp. 1–3. <https://doi.org/10.1109/EDTM50988.2021.9421002>
4. Hao, Q., Li, S., Fang, L., Kang, X.: Multiscale feature extraction with Gaussian curvature filter for hyperspectral image classification. In: IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium, 2020, pp. 80–83. <https://doi.org/10.1109/IGARSS39084.2020.9323640>
5. Guo, H., Liu, Y., Zhao, J., Yang, D.: Research on feature extraction of Tai Le recognition. In: 2020 IEEE 3rd International Conference on Computer and Communication Engineering Technology (CCET), 2020, pp. 95–98. <https://doi.org/10.1109/CCET50901.2020.9213168>
6. Zhao, G., Li, T., Yang, Z.: An extended knowledge representation learning approach for context-based traceability link recovery: extended abstract. In: 2020 IEEE Seventh International Workshop on Artificial Intelligence for Requirements Engineering (AIRE), 2020, p. 22. <https://doi.org/10.1109/AIRE51212.2020.00010>
7. Wu, C., Qi, G., Zhao, H., Chen, Z.: Feature extraction of cultural gene image based on PCA method. In: 2020 International Conference on Computer Engineering and Application (ICCEA), 2020, pp. 860–863. <https://doi.org/10.1109/ICCEA50009.2020.00189>
8. Joy, A.A., Hasan, M.A.M.: A hybrid approach of feature selection and feature extraction for hyperspectral image classification. In: 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), 2019, pp. 1–4. <https://doi.org/10.1109/IC4ME247184.2019.9036617>
9. Yang, L., Zhang, J., Yang, Y.: A feature extraction technique in stereo matching network. In: 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2019, pp. 393–396. <https://doi.org/10.1109/IAEAC47372.2019.8998024>
10. Mestri, R., Limaye, P., Khuteta, S., Bansode, M.: Analysis of feature extraction and classification models for lip-reading. In: 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 911–915. <https://doi.org/10.1109/ICOEI.2019.8862649>
11. Zhang, S., Zhai, J., Xie, B., Zhan, Y., Wang, X.: Multimodal representation learning: advances, trends and challenges. In: 2019 International Conference on Machine Learning and Cybernetics (ICMLC), 2019, pp. 1–6. <https://doi.org/10.1109/ICMLC48188.2019.8949228>
12. Lu, M., Li, F.: Survey on lie group machine learning. *Big Data Min. Anal.* **3**(4), 235–258 (2020). <https://doi.org/10.26599/BDMA.2020.9020011>

13. Chen, J., Gong, Z., Wang, W., Liu, W., Dong, X.: CRL: collaborative representation learning by coordinating topic modeling and network embeddings. *IEEE Trans. Neural Netw. Learn. Syst.* <https://doi.org/10.1109/TNNLS.2021.3054422>
14. Sharma, V., Tapaswi, M., Sarfraz, M.S., Stiefelhagen, R.: Clustering based contrastive learning for improving face representations. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), 2020, pp. 109–116. <https://doi.org/10.1109/FG47880.2020.00011>

# Drowsiness Detection Using Facial Features, Image Processing and Machine Learning



S. Nandhini, Vaishnavi Venkatasubramanian, and C. Aparna

**Abstract** With an increase in population, the occurrence of automobile accidents has also seen an increase. Road traffic accidents are a common cause of trauma and loss of life in the community. Drowsiness or sleepiness is one of the key causes for traffic accidents and has a serious impact on road safety. Many fatal accidents can be avoided by warning sleepy drivers on time. According to World Health Organization, traffic accidents have gone up to 1.25 billion around the world and sensing driver's fatigue will be one of the main potential areas for avoiding many sleep-induced traffic accidents. There are different types of drowsiness detection methods that scan whether the person driving is fatigued or not while driving and alert the driver if he is not focused on driving. To prevent this from happening, we have proposed a facial recognition algorithm that detects driver drowsiness from facial features. The algorithm first detects facial features such as yawning and blinking frequency. The facial features determine the driver's condition. By combining the condition of the driver with the responsiveness of the eyes and mouth, the driver will be alerted.

**Keywords** Machine learning · Image processing · Drowsiness detection · Haar cascade algorithm · Convolution neural network

## 1 Introduction

Urbanization has increased tremendously over the past decade, resulting in an exponential increase in the number of vehicles on the roads. In this type of traffic, the number of road accidents has increased dramatically, and therefore, the number of deaths on the road has reached a record high. For this reason, road safety is one of the most focused areas in the traffic sector. A slip in this department in the slightest way would lead to catastrophic circumstances. New innovations are developed every day to circumvent these security gaps by integrating future technologies with common

---

S. Nandhini (✉) · V. Venkatasubramanian · C. Aparna

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai 600089, India  
e-mail: [nandhins1@srmist.edu.in](mailto:nandhins1@srmist.edu.in)

methods of solving problems. This is much-needed help because road accidents occur very frequently and so many people lose their lives every minute, particularly in a country like India.

In recent times, the need for a vehicle has become a necessity for people to commute from various places. The demand for vehicles is also increasing rapidly. On the contrary, the number of accidents happening on the roads mainly in the highways has also increased. India accounts for almost 11% of the accidents that happen around the world. Annually, 1.5 lakh people have died due to accidents. One of the major reasons was the driver's drowsiness while driving. According to a survey conducted in 2020, around 20,732 people have died due to the driver's sleep being deprived. Drivers, who tend to be very drowsy, while driving, are at fault for about 40% of the accidents that occur. This is proved by an analysis conducted through the Central Road Research Institute.

Drowsiness from long periods of driving is one of the most common and dangerous problems related to road safety. Many technologies and systems have been proposed to study this problem, but not many have been very successful. Some reasons for this are that the developed system has a subjective detection method in which the driver is forced to participate in activities such as completing questionnaires from which his drowsiness is detected. It can be observed that the symptoms of drowsiness when the driver is at the wheel. Those symptoms include the frequent yawning, inability to open the eyes, moving the head forward with a sudden jerk, etc.

In physiological measurements, electrooculogram (EOG), electrocardiography (ECG) and electroencephalography (EEG) are utilized to assess the condition of the driver. Although these appliances give accurate outputs due to their actual hindrances, they are not regarded favourably. For transport-based measurements, sleepiness is detected using steering wheel gestures and forms with respect to the holding of breaks. To identify the facial region, facial recognition algorithms were utilized. These are done by getting face images as inputs during the face recognition phase.

## 2 Related Work

(Reference [1]) In this paper, they used the perceptron multilayer classifier. They take videos from the NTHU database and then extract them as images. The combination of benchmarks is then extracted and submitted to the algorithm. Finally, the algorithm decides whether the driver is drowsy or not. This content is saved in a file format that can be used in the mobile app.

(Reference [2]) This paper introduces a new drivers' sleepiness detection method to address the effective issue of motorists' safety on the road. This paper uses convolutional neural networks to detect the driver's sleep state as a sign of drowsiness, taking into account the two real-time goals, which are high accuracy and responsiveness. A whole of three networks are used as classification networks. One of the networks used is a thoroughly fledged interconnected system. Those remaining use transfer learning in VGG19 as well as VGG16 accompanied by additional layers.

(Reference [3]) In this, driver drowsiness is built using IoT technology in conjunction with Raspberry Pi. While the driver is driving, a Pi camera is used to take pictures, and if drowsiness is detected, a voice speaker warns the driver and an email is sent to the vehicle owner. If an accident occurs during this time, the collision sensor records the data, and the nearest hospitals are alerted.

(Reference [4]) This article proposes a system called DriCare, which uses video images to detect driver fatigue such as yawning, blinking and how long it takes to close the eyes without equipping the body with any device. Based on 68 key points, a new recognition method for identifying the areas of the face is implemented. These facial regions are then used to assess the condition of the driver. DriCare will send a warning to the driver if the drowsiness is detected by combining the functions of the mouth and eyes.

(Reference [5]) This paper proposes a prototype to detect the drivers' drowsiness based on the conditions of the eye. An accuracy of 90% was achieved in this paper. A webcam is used to capture a series of images. This image can be saved as a single frame. That frame is considered as the input. Features, such as eyes, are drawn out from the snapshot. By developing the eye individually, the structure initiates a condition and advocates a definite number of frameworks with matching eye circumstances that may be recorded. The outcome of these snapshots can be held as input to get the degree of sleepiness that can be experienced by a driver at one time while driving a vehicle.

(Reference [6]) This article is about the development of driver drowsiness with the help of visually evoked potentials also known as VEPs. VEPs are calculated by detecting electroencephalogram signals from the visual cortex. A classification procedure is used to distinguish the eyes that are opened and the ones that are closed. For classification, radial basis neural networks are used, and power spectrum density features, Fourier transform and multilayer perceptron are utilized too.

(Reference [7]) In this paper, behavioural metrics and machine learning techniques were used to measure levels of sleepiness. Neural convolutional networks, support vector machines and hidden Markov models are used for detecting the drowsiness of the driver.

(Reference [8]) This paper uses deep learning algorithms to check if the handler is sleepy by recognizing the eye state of the driver while driving. The histogram equalization and Canny edge detection algorithms are used for the detection of the face and eye as well. An application is added as an extra feature to determine the active status of the driver with an alarm tone in the application.

(Reference [9]) This study proposes a system that uses ML algorithms to detect sleepiness, which cautions the driver to steer clear of misadventures. For the identification of the driver's eyes, this paper utilized the Haar cascade algorithm along with the OpenCV library. These eyes are detected by examining the coexistent video of the person driving. To determine if the eyeballs are opened or closed, eye aspect ratio (EAR) is used. A Raspberry Pi hybrid chip is used, along with a video equipment component and an alert system.

(Reference [10]) This paper proposes a sleepiness detection method for people who drive for long hours. There are three categories of process by which the detection can be detected. Those are EOG, EEG and driver image inspection. The writer of this paper checked the possibility of detecting the drowsiness based on the images taken when the driver is at the wheel. Then the drivers' eyes are checked if they are open, half-opened or closed. In this article, the writer investigated the likelihood of detecting the drowsiness or watchfulness of the driver based on the snapshot taken while moving and analysed the condition of the eyes of the person in the vehicle: open, partially open and completely closed. One invisible layer matrix and auto-associative matrix are the two types of artificial semantic networks used for this purpose.

(Reference [11]) In this thesis, ML is used for driver sleepiness. Here, they have detected the face and eyes along with the opening and closing of the eyes. This paper utilizes a greyscale image that works effectively day and night. Machine learning algorithms are used for face detection, and the geometric position of the face is used to minimize the range of eye searching. Based on the eye detection algorithm, it is detected irrespective of the fact that the person driving is wearing spectacles or not. In the end, the state of the drivers' eye is detected. If the motorist keeps the eyes closed for a prolonged span of time, and if he does not pay attention on steering or dozes his crown and moves it forward suddenly, an alarm is used to warn the driver.

(Reference [12]) In this paper, an Android application is developed to check the sleepiness of the driver while steering the vehicle. The driver's features are acquired. The facial landmark is then used to calculate features such as the ratio of the eye aspect, yawning and nose length based on which the drowsiness is detected. The most important of these features is their unobtrusive and affordable nature. A data set of 1200 users was used, and machine learning techniques were used for processing the same. An accuracy of 98% was achieved in this study with the utilization of a bagging classifier.

(Reference [13]) In this document, an inexpensive real-time fatigue detection system with allowable precision is developed. In this system, a web camera recorded the clips, and the motorist's face was acknowledged in every snapshot accompanied by the help of photo processing techniques. The facial points were envisioned on the recognized face, then the proportion of the eye, yawning and the nose length were computed. The fatigueness with respect to the formation of the dynamic or local threshold value were estimated.

(Reference [14]) The main idea behind this paper was to develop a non-intrusive data processing vision system with OpenCV that will extemporaneously check the driver's fatigue in a coexistent video feed. The driver is warned if they appeared sleepy by triggering a configured alarm. This system employs an algorithm that could detect the driver's eye reference points and generate an alarm by checking whether the driver is sleepy or not.

(Reference [15]) This paper proposes a framework that uses deep learning to check the fatigue of the motorist based on the state of the eye while steering the vehicle. To recognize faces and draw out the ocular area from the facial snapshots, in this work, the Viola-Jones facial recognition algorithm is used. A deep CNN was evolved to

withdraw dynamically identified mainframe characteristics from the digicam array and use them in the phase of learning. A softmax layer is utilized to grade the motorist as fatigued or not asleep in the CNN classifier. This system warns the motorist with an alert when he/she is sleepy.

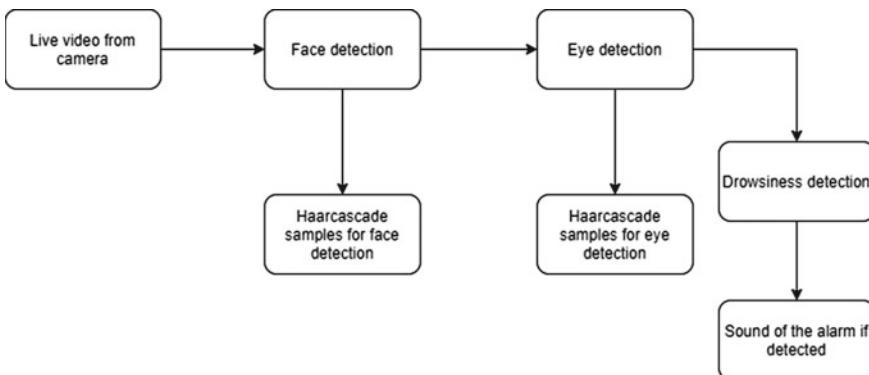
### 3 Proposed System

In this paper, a live video is taken from a webcam that is attached which captures the face and eyes of the driver and predicts the level of drowsiness. Preprocessing techniques such as conversion to greyscale images are applied to the data set. Then the data set is split into two, i.e. training set and testing set.

Convolutional neural networks are applied to encounter driver fatigue. Every drowsy picture requires a feature vector to correlate with real features in the data set to notice the drowsy image. A CNN is utilized in the recommended system for driver drowsiness recognition because an attribute vector is required for every fatigueness snapshot to correlate the characteristics present in the data set to recognize sleepiness. CNN is a deep learning algorithmic rule that can take snapshots as input, assign attention to different characteristics in the snapshot and can distinguish one from another.

Haar cascade algorithm is used in this implementation. It is a machine learning object detection algorithm which was put forward by Paul Viola and Michael Jones in their 2001 thesis “Fast Object Detection Using an Improved Cascade of Simple Functions”. It is an ML-based method where the waterfall function is instructed by a sequence of positive and negative perception. This is then used to identify objects in other images.

The system architecture of the driver’s drowsiness detection system is depicted in Fig. 1. In this system, first the image input is taken from the camera with the help of



**Fig. 1** Architecture diagram for driver’s drowsiness

a never-ending loop which will picturize every snapshot. It is converted to greyscale for image detection because the OpenCV algorithmic rule for object identification captures greyscale snapshots in the input. The same method is used to identify the eyes. CNN is used for predicting eye status. A score value will be displayed based on the eye status. The score is a number that is utilized to find out how long the eyes are shut. The score increases when the eyes are kept closed. When the eyes are open, the score goes down. Drowsiness is recognized when the eyes are shut for a long duration. If the score is higher than 15, an alarm beep sound will ring.

## 4 Implementation

The data set applied here consists of 7000 images of human eyes under various lighting conditions. This data set is separated and labelled as “Open” and “Closed” and is stored. The images are resized to 24 \* 24 pixels. The images from the data set are converted to greyscale images which would be taken as input. Greyscale images are used instead of coloured images to process less information in each pixel, and a lot of unrelated information is present in coloured images which are not required for the drowsiness detection of the driver. The images are then converted to an array to be able to enter the model. It is also scaled by dividing by 255. After the preprocessing, the data is split into a training set and test set in the proportion 70:30.

Convolution neural network is the classifier used in this particular paper. Convolutional neural network is a deep learning algorithm. It is a type of neural network that can take images as input, allocate weights to different aspects of the image and can distinguish one from another. They generally consist of neurons of learned weights and biases. They contain one or more hidden layers, an input layer and an output layer. CNN is used in this paper because the feature extraction process is done on its own and saves the work of extracting the individual features manually, for the rest of us.

In this paper, a sequential model as a model to predict the driver’s drowsiness has been used. CNN model has two convolutional layers which has 32 nodes and a kernel size 3 and a third and final layer consisting of 64 layers and a kernel size of 3. Pooling layers are also used in order to reduce the computations that are carried out in the network and also decreased the number of parameters that are used for learning. Here, a MaxPooling2D to reduce the convolution matrix is utilized. The final entirely interlinked layer has 128 nodes altogether. All the convolutional layers use a rectified linear activation function or a ReLU function. It is a linear function that immediately outputs an input immediately if it is positive, if not it will return zero.

This has become a default activation feature for many sorts of neural networks as the models that are used are easier to train and often perform better. The dropout function is used to randomly change the input values as zero with a frequency of the value of rate used during the training of the model. It drops many parameters too. This is done to prevent overfitting. Since only the output for a classification problem is

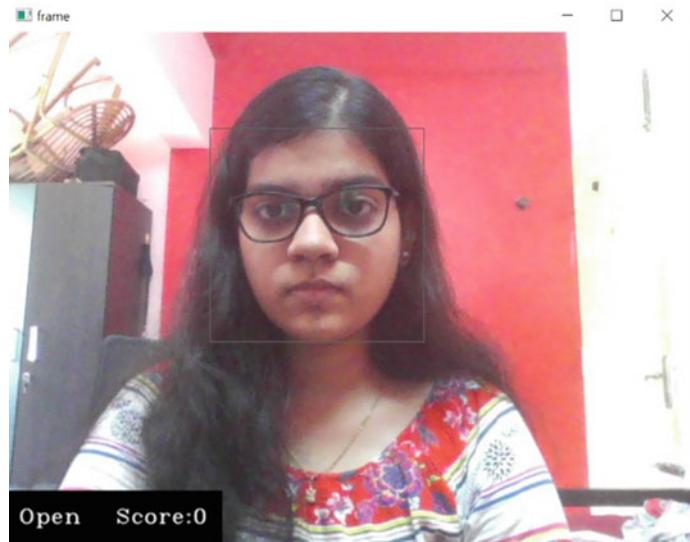
required, the dimensions are reduced using a flatten function. Next, the dense layers are utilized to build the finally connected layer. Softmax function is a function that is used in the final dense layer to normalize the output, converting it from a weighted sum value to a probability that adds to one. The values that have been obtained are either one or zero which corresponds to the eye being open or closed, respectively. The whole model is then compiled using Adam optimizer, and epochs are set to train the model and test the same. Around 15 epochs are used in this model. This is the training of the model.

After the training, the live feed from the camera needs to be considered for the prediction of the driver's drowsiness. The video clip is split up into frames, and every snapshot is read and saved, which is used as the input. All these images stored are then converted to greyscale images since colour is not necessary for the prediction. Next, Haar cascade classifiers are utilized to detect the face. The eyes are also detected using a different set of Haar cascade samples. The eyes are detected using the help of the detected face. The area of interest of the face is highlighted so that the amount of processing of the whole image is reduced. It is well known that the eyes are found mostly in the upper part of the face. This fact is considered for the detection of the eyes, and the Haar cascade XML files are also utilized for the detection of the same. After the detection of the face, the state of the eye is checked to see if the driver has his eyes opened or closed. This is done by checking the pixel values. If the value is low, then the eye is considered as open else it is considered as closed. The scores are also calculated for the same. It is a value that determines how long the eyes of the person driving are closed. If the eyes are shut, the scores will keep shooting up, and when the eyes are opened, the scores will go down. An alarm is used if the value of the score reaches beyond a particular threshold value. The alarm starts to ring continuously if the score is higher than the threshold value.

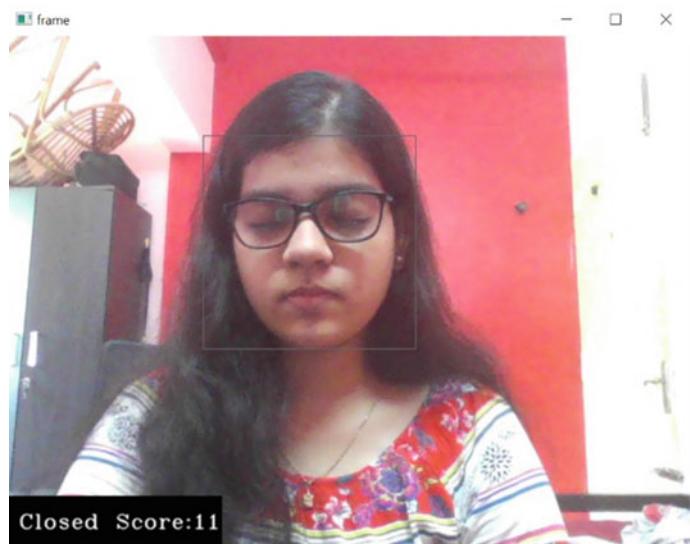
Figure 2 shows the score value with the eye open. Figure 3 shows that the person's eye is closed and a little sleepy. If the person does not open their eyes in a certain period of time, the score will continue to increase. This is illustrated in Fig. 4. As soon as the score exceeds 15, an alarm will beep to warn the driver. This is a very useful way to avoid accidents.

## 5 Result

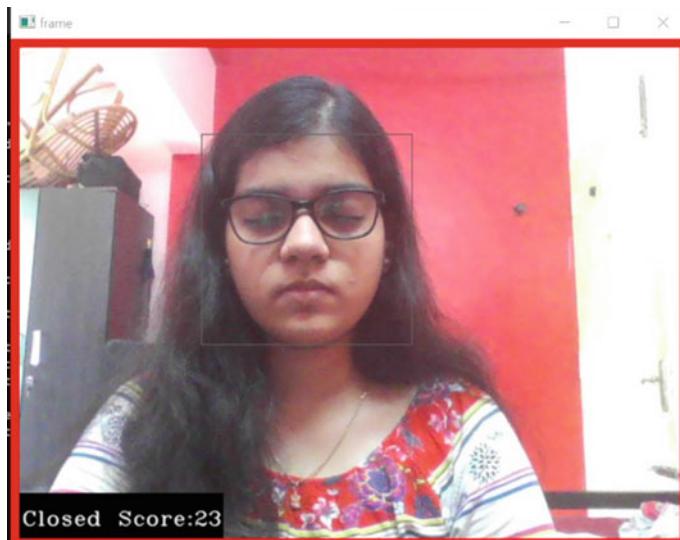
This paper proposes the driver's drowsiness detection system using CNN and Haar cascade algorithm. The data set goes through a set of preprocessing techniques, a model is created, preprocessed data is fed to it, and the model is trained. At the initial stage, a score of zero is displayed. The score continuously gets incremented as the driver dozes off, and once it crosses the threshold value, i.e. score of 15, an alarm beeps which will alert the driver. The score will come down to zero (which is the minimum value) when the driver's eyes are open.



**Fig. 2** Open eye detection with a score 0



**Fig. 3** Closed eye detection



**Fig. 4** Sleep alert with an alarm sound (when score goes above 15)

## 6 Conclusion and Future Scope

In this paper, a new method that detects driver fatigue with respect to the eye condition has been suggested. This decides whether the driver's eyes are fatigued or not. If the eyes are drowsy, an alarm is triggered. The facial and eye regions are recognized with the Haar cascade algorithm. The stacked deep CNN is conceived to bring out functions and is used in the phase of learning. The proposed system achieves a precision of (95%). The suggested system successfully recognizes the condition of the person driving and activates an alert if the framework constantly predicts the initial position of drowsiness. The existing parameters can be updated by adding more complex parameters which give better results. Mouth tracking can also be used as a parameter to measure the level of fatigueness with the help of yawning. Hence, there is still scope for future enhancements like detecting the facial features even with minimum light.

## References

1. Rate Jabbar, Khalifa Al-Khalifa, Mohamed Kharbeche, Wael Alhaj yaseen, Mohsen Jafari ShanJiang - Real-Time Driver-Drowsiness Detection System Using Facial Features **130**, 400–407 (2018)
2. Maryam Hashemi, Alireza Mirrashid, Aliasghar Beheshti Shirazi - Driver Safety Development: Real-Time Driver Drowsiness Detection System Based on Convolutional Neural Network - SN Computer Science volume 1, Article number: 289 (2020) Published: 31 August 2020

3. Anil Kumar Biswal, Debabrata Singh, Binod Kumar Pattanayak, Debabrata Samanta, Ming-Hour Yang - IoT-Based Smart Alert System for Drowsy Driver Detection - Volume 2021 | Article ID 6627217 - 10 Mar 2021
4. Ruoxue Wu, Wanghua Deng - Real-Time Driver-Drowsiness Detection System Using Facial Features - DOI <https://doi.org/10.1109/ACCESS.2019.2936663>, IEEE Access
5. Asad Ullah, Sameed Ahmed, Lubna Siddiqui, Nabiha Faisal - Real Time Driver's Drowsiness Detection System based on Eye Conditions - Volume 6, Issue 3, March-2015
6. Amjad Hashemi, Valiallah Saba, Seyed Navid Resalat - Real Time Driver's Drowsiness Detection by Processing the EEG Signals Stimulated with External Flickering Light - International Journal of Scientific & Engineering Research, Volume 6, Issue 3, March-2015
7. Ngxande, M., Tapamo, J.-R., Burke, M. (2017). Driver drowsiness detection using behavioral measures and machine learning techniques: A review of state-of-art techniques. Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech)
8. S. E. Viswapriya , Singamsetti Bala Balaji , Yedida Sireesha - A Machine-Learning Approach for Driver-Drowsiness Detection based on Eye-State - Published (First Online): 13–04–2021 - ISSN (Online) : 2278–0181 - IJERT Volume 10, Issue 04 (April 2021)
9. Shivani Sheth, Aditya Singhal, V.V. Ramalingam - Driver Drowsiness Detection System using Machine Learning Algorithms - International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277–3878, Volume-8 Issue-6, March 2020
10. T. Vesselényi, S. Moca, A. Rus, T. Mitran, B. Tătaru - Driver drowsiness detection using ANN image processing - IOP Conf. Series: Materials Science and Engineering 252 (2017) 012097 doi:<https://doi.org/10.1088/1757-899X/252/1/012097>
11. Cyun-Yi Lin, Paul Chang, Alan Wang, Chih-Peng Fan - Machine Learning and Gradient Statistics Based Real-Time Driver Drowsiness Detection - 2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)
12. Sukrit Mehta, Parimal Mishra, Arpita Jadhav Bhatt, Parul Agarwal - AD3S: Advanced Driver Drowsiness Detection System using Machine Learning - 2019 Fifth International Conference on Image Information Processing (ICIIP)
13. Muhammad Ramzan, Hikmat Ullah Khan, Shahid Mahmood Awani, Amina Ismail, Mahwish Ilyas, Ahsan Mahmood - A Survey on State-of-the-Art Drowsiness Detection Techniques - DOI <https://doi.org/10.1109/ACCESS.2019.2914373>
14. Jongseong Gwak, Akinari Hirao, Motoki Shino -An Investigation of Early Detection of Driver Drowsiness Using Ensemble Machine Learning Based on Hybrid Sensing
15. Venkata Rami Reddy Chirra, Srinivasulu Reddy Uyyala, Venkata Krishna Kishore Kolli - Deep CNN: A Machine Learning Approach for Driver Drowsiness Detection Based on Eye State - Vol. 33, No. 6, December, 2019, page 461–466

# Linear Separability as a Condition for Solving Multiple Problems by a Single Threshold Neuron



Kostadin Yотов, Emil Hadzhikolev, and Stanka Hadzhikoleva

**Abstract** The paper discusses the linear separability of data classes and the relationship of threshold neurons with class classifiers. The possibility of constructing a neuron that can solve two different problems without the need for an intermediate change in its parameters and architecture is shown theoretically. The idea is illustrated with a specific example of a neuron solving problems simultaneously with both Boolean functions “AND” and “OR”. A conclusion has been drawn for the existence of a neuron that can solve a class of an infinite number of problems. A necessary condition for this is that the domain of the problems is linearly separable from the surface in the input data space and the existence of parallel classifiers for separability for each individual problem.

**Keywords** Threshold neuron · Perceptron · Linear separability

## 1 Introduction

Let a set of objects be given

$$A = \{A_1, A_2, \dots, A_m\}, m \in N, \text{ in which}$$

each object  $A_i, i = 1 \dots m$  is characterized by “ $n$ ” different attributes of a completely random type. In the general case, when considering the linear separability, it is not necessary to define the type of the individual attributes in advance, but to draw attention to the possibility of grouping the objects into classes through these attributes.

---

K. Yотов · E. Hadzhikolev · S. Hadzhikoleva (✉)  
University of Plovdiv “Paisii Hilendarski”, Plovdiv, Bulgaria  
e-mail: [stankah@uni-plovdiv.bg](mailto:stankah@uni-plovdiv.bg)

K. Yотов  
e-mail: [kostadin\\_yотов@uni-plovdiv.bg](mailto:kostadin_yотов@uni-plovdiv.bg)

E. Hadzhikolev  
e-mail: [hadjikolev@uni-plovdiv.bg](mailto:hadjikolev@uni-plovdiv.bg)

One attribute can be qualitative—for example, “color” or “shape,” another—quantitative, for example, “weight” or “temperature,” and if necessary, for each qualitative characteristic we could give some quantitative expression.

Following this line of thought, if we introduce quantitative correspondences of qualitative characteristics, then we can describe each object only with quantitative characteristics. For example, if for the colors we assume green = 28, and for the shapes—polygon = 123, then the object green polygon is:

$$(\text{color} = \text{"green"}, \text{shape} = \text{"polygon"})$$

and it can be represented as:

$$(x_1 = 28, x_2 = 123).$$

To simplify the reasoning, let us assume that all the characteristics of the considered objects are quantitative values. This in turn means that each object of the set A can be considered as a point in the  $n$ -dimensional space of the attributes (Fig. 1), i.e.,

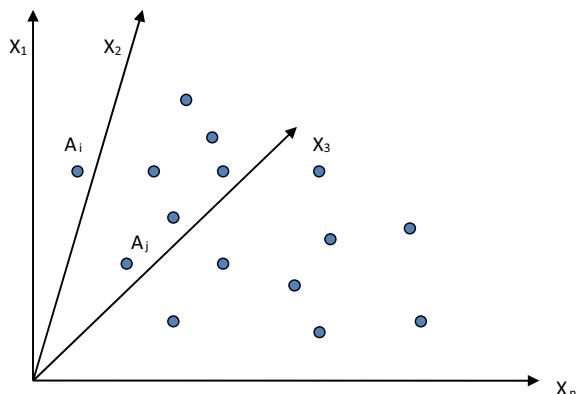
$$A_i = \{x_{i1}, x_{i2}, \dots, x_{in}\} \in E^n, \quad \forall A_i \in A, i = 1, 2, \dots, m$$

As for the linear separability of A, it can be represented as a union of two linearly separable classes [1, 2]:

$$\begin{aligned} B &= \{\forall B_i(x_{i1}, x_{i2}, \dots, x_{in}) / B_i \in A, i = 1, 2, \dots, m\} \text{ and} \\ C &= \{\forall C_j(x_{j1}, x_{j2}, \dots, x_{jn}) / C_j \in A, j = 1, 2, \dots, m\}, \\ &\text{as } B \cap C = \emptyset, \end{aligned}$$

if there exists a plane  $\alpha$  with representation:

**Fig. 1** Elements of the sum A, presented as points in  $E^n$



$$\alpha : \sum_{i=1}^n w_i x_i + b = 0, \quad (\forall w_i \in R, i = 1, 2 \dots n) \quad (1)$$

such as for  $\forall B_i(x_{i1}, x_{i2}, \dots x_{in}) \in B$ , the inequality is fulfilled

$$\sum_{k=1}^n w_k x_{ik} + b < 0 \quad (2)$$

and  $\forall C_j(x_{j1}, x_{j2}, \dots x_{jn}) \in C$ —respectively:

$$\sum_{k=1}^n w_k x_{jk} + b > 0. \quad (3)$$

Let us consider a two-dimensional space of attributes, assuming that all objects of the set  $A$  are characterized by only two qualities. In this case,  $n = 2$  the plane (1) is reduced to its two-dimensional analogue—straight line with the equation:

$$\alpha : w_1 x_1 + w_2 x_2 + b = 0, w_i \in R, i = 1, 2 (1')$$

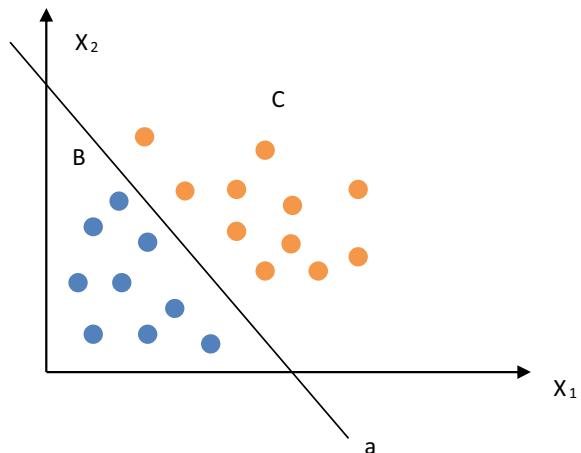
where  $x_1$  and  $x_2$  are the coordinates of an arbitrary point of  $\alpha$ .

Transferred to  $E^2$  (Fig. 2), based on the two-dimensional variant of conditions (2) and (3), classes  $B$  and  $C$  are linearly separable if for  $\forall B_i(x_{i1}, x_{i2}) \in B$  the equation is fulfilled:

$$w_1 x_{i1} + w_2 x_{i2} + b < 0 \quad (2')$$

and  $\forall C_j(x_{j1}, x_{j2}, \dots x_{jn}) \in C :$

**Fig. 2** Elements of the set  $A$ , presented as points in  $E^2$



$$w_1x_{j1} + w_2x_{j2} + b > 0 \quad (3')$$

Since the conditions (2') and (3'), classifying the belonging of an object to one of the two classes are based on equation (1'), the line  $\alpha$  is a linear classifier for  $B$  and  $C$ .

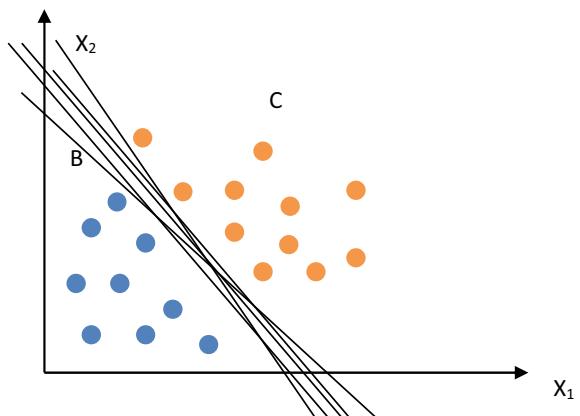
There are many optimization algorithms for finding a suitable classifier that can effectively separate classes [3, 4] and some testing methods for linear separability [5]. In our case, however, we are interested not so much in finding the most efficient separation, but in finding suitable parallel classifiers. Let us pay attention to the fact that if two classes are bounded and linearly separable, an infinite number of linear classifiers can be indicated (Fig. 3), which separate them in the way described by (1)–(3).

Let us look at the two-dimensional Boolean function “AND” (Table 1). Its domain consists of 4 points:

$$D = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$$

Let us introduce the class of points  $B$ , containing the elements of  $D$ , for which  $X_1 \wedge X_2 = 0$ , and the class  $C$ , containing the points from D.O, for which  $X_1 \wedge X_2 = 1$ . Given the essence of the Boolean function “AND”, we look for a classifier type (1'), for which the following system is implemented:

**Fig. 3** Multiple classifiers separating the two classes  $B$  and  $C$  linearly



**Table 1** Truth table for Boolean functions “AND” and “OR”

$X_1$	$X_2$	$X_1 \wedge X_2$	$X_1 \vee X_2$
0	0	0	0
0	1	0	1
1	0	0	1
1	1	1	1

$$\begin{cases} w_10 + w_20 + b < 0 \\ w_10 + w_21 + b < 0 \\ w_11 + w_20 + b < 0 \\ w_11 + w_21 + b > 0 \end{cases}$$

It is obvious that the free member  $b$  must meet the conditions:

$$\begin{cases} b < 0 \\ b < \min\{-w_1, -w_2\} \\ b > -(w_1 + w_2) \end{cases} \quad (4)$$

Then for every two specific positive numbers  $w_1$  and  $w_2$ , we can find a corresponding value of  $b$ , by which we can define a classifying straight line of the type (1'), which will be only one of an infinite number of members of the same family of classifiers. If, for example,  $w_1 = 0.3$  and  $w_2 = 0.7$ , then according to the system (4) we can choose arbitrarily  $b$ :

$$-1 < b < -0.7$$

One possible solution is  $b = -0.8$ . Thus, this particular representative of the whole possible class is given by the equation:

$$\alpha_1 : 0.3x_1 + 0.7x_2 - 0.8 = 0$$

For the coordinates of each point  $(x_1, x_2) \in \alpha_1$  is met

$$x_2 = -\frac{0.3}{0.7}x_1 + \frac{0.8}{0.7}$$

This is a straight line passing through point  $(0, 1.143)$  and having an angular coefficient  $k = -(0.3/0.7) = -0.428$ . Figure 4 shows an illustration of other possible solutions:

$$\alpha_2 : 0.3x_1 + 0.7x_2 - 0.9 = 0,$$

which has the same angular coefficient as that of  $\alpha_1$ , and

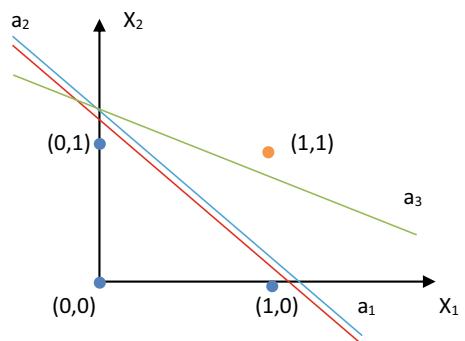
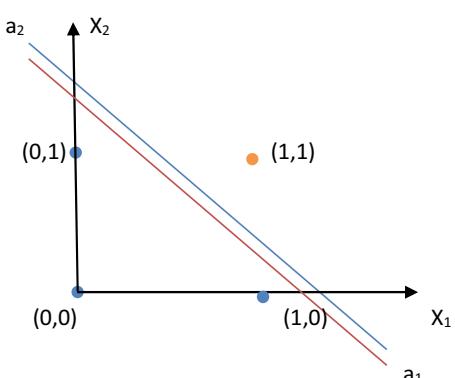
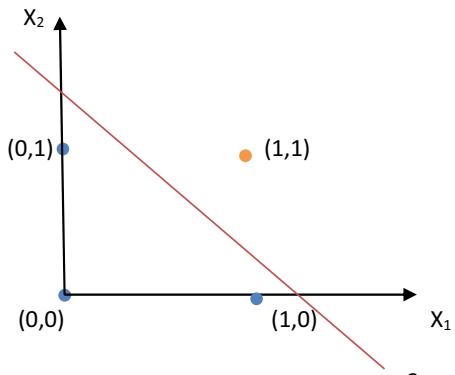
$$\alpha_3 : 2x_1 + 9x_2 - 10 = 0,$$

with angular coefficient  $k = -0.222$ .

## 2 Threshold Neuron and Its Relationship with Linear Classifiers

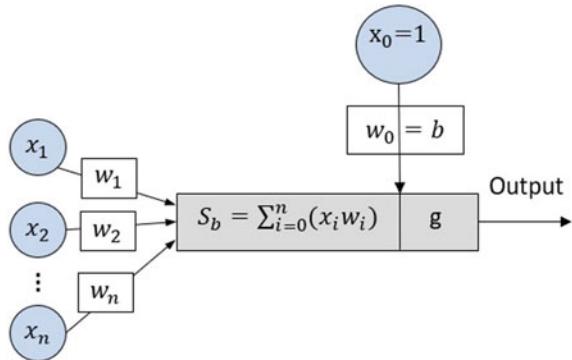
Let us consider a threshold neuron with  $n$  inputs which are activated through the following step function:

**Fig. 4** Three straight lines classifying the domain of the Boolean function “AND”



$$g(S) = \begin{cases} 0, & \text{at } S < 0 \\ 1, & \text{at } S \geq 0 \end{cases} \quad (5)$$

If the weight vector  $\vec{W}$  has coordinates  $(w_1, w_2, \dots, w_n)$ , at input stimuli  $(x_1, x_2, \dots, x_n)$  along the axon of the neuron, a signal will propagate with the following

**Fig. 5** Threshold neuron

value:

$$\text{Output} = g \left[ \sum_{i=1}^n w_i x_i + b \right],$$

where  $b$  is the threshold of the neuron [6, 7]. Given the peculiarities of the activating function, it is clear that this signal has a binary character—it will be “0” or “1”. The neuron thus defined is called the threshold logic or TLU (Fig. 5), or just threshold neuron [8]. The term perceptron is often used as a synonym for threshold logic unit, although the perceptron is generally much more than a simple threshold logic unit [9].

Based on the way the TLU is constructed, it is clear that if the stimuli ( $x_1, x_2, \dots, x_n$ ) appear in its dendritic tree, at fixed weights ( $w_1, w_2, \dots, w_n$ ), only the following results are possible:

$$g \left[ \sum_{i=1}^n w_i x_i + b \right] = \begin{cases} 0 \\ 1 \end{cases}$$

Thus, unlike other neurons, TLU easily realizes the concept of “all or nothing,” which makes it especially suitable for solving logical problems [10].

We will consider a special case when using threshold neurons, namely the one in which the set of input data is linearly divisible by a given attribute.

Thus, let the domain of the input variables be linearly divisible with respect to the ordered n-tuple  $(x_1, x_2, \dots, x_n)$  leading to the appearance of “1” at the output of the neuron, and those  $(x_1, x_2, \dots, x_n)$ , which generate the end result “0”. From the assumed linear separability, it follows that there is a plane:

$$\alpha : \sum_{i=1}^n w_i x_i + b = 0 (\forall w_i \in R, i = 1, 2 \dots n), \text{ such that}$$

for  $\forall(x_{01}, x_{02}, \dots, x_{0n})$ , for which at the output we have

$$g\left[\sum_{i=1}^n w_i x_{0i} + b\right] = 0$$

the inequality is fulfilled

$$\sum_{i=1}^n w_i x_{0i} + b < 0,$$

and for  $\forall(x_{11}, x_{12}, \dots, x_{1n})$ , for which at the output we have

$$g\left[\sum_{i=1}^n w_i x_{1i} + b\right] = 1, \text{ we have: } \sum_{i=1}^n w_i x_{1i} + b > 0$$

The presented information about the linear classifiers outlines their connection with the threshold neurons. If we look at the total signal in the body of the artificial neuron, we will notice the obvious classification form:

$$S = \sum_{i=1}^n w_i x_i + b,$$

while the classification itself is done through the activating function (5). Thus, the action of the trained threshold neuron could be considered as an act of classification of the linearly separable domain of its input variables. This essential relationship between TLU and linear classifiers allows us to draw two very important conclusions:

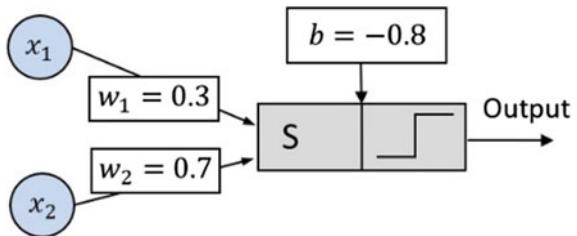
- (1) Based on each mathematically constructed linear classifier, we can form a threshold neuron that is genetically prepared to solve the classifier problem. If we consider, for example, one of the classifiers for the Boolean function “AND”, which we built above:

$$\alpha_1 : 0.3x_1 + 0.7x_2 - 0.8 = 0$$

It is clear that we could form the neuron as shown in Fig. 6. Along the axon, we have a signal which, given the weights and thresholds, fully corresponds to the Boolean function “AND”.

- (2) We mentioned that if two classes are linearly separable, then there are infinitely many linear classifiers separating the objects in each of the classes. And since the classifier is uniquely determined by the coefficients  $w_1, w_2, \dots, w_n, b$ , with which we could subsequently form a corresponding threshold neuron, this means something very important to us, namely ***For each specific problem***

**Fig. 6** Threshold neuron prepared to solve the Boolean function “AND”



*with a linear separable compact domain of the input variables, there are an infinite number of threshold neurons that are able to solve it.*

On the other hand, from the existence of infinitely many, but let us emphasize now, **parallel** linear classifiers, follows the existence of neurons:

$$\begin{aligned} H_1 : Out_1 &= g \left[ \sum_{i=1}^n w_{1i} x_i + b_1 \right] \\ H_2 : Out_2 &= g \left[ \sum_{i=1}^n w_{2i} x_i + b_2 \right] \\ \dots & \\ H_r : Out_r &= g \left[ \sum_{i=1}^n w_{ri} x_i + b_r \right] \\ \dots & \end{aligned}$$

with weights for which

$$\frac{w_{1p}}{w_{1q}} = \frac{w_{2p}}{w_{2q}} = \cdots = \frac{w_{rp}}{w_{rq}} = \cdots, p \neq q; p, q = 1, 2, \dots n$$

and associated with planes

$S_j = 0$ , where

$$S_j = \sum_{i=1}^n w_{ji} x_i + b_j, j \in N.$$

### 3 Solving Multiple Problems from a Single Threshold Neuron

An interesting question is about the possibility of the same neural network to solve a set of several tasks. There are various researches on this issue. For example, Kirkpatricka and team apply a special type of neural network regularization, which is associated with sequential training of the network to solve two tasks—A and B [11]. In the second task B, the weights required for the first task A are retained and the gradient descent continues. On the other hand, Yang and team train single network models to perform 20 cognitive tasks that depend on working memory, decision making, categorization, and inhibitory control [12]. The authors find that after training, recurrent units can be grouped into clusters that are functionally specialized for different cognitive processes and introduce a simple but effective measure to quantify relationships between single-unit neural representations of tasks. In the present study, we focus on the possibility of a separate threshold neuron to solve two logical functions simultaneously, without the need for sequential training for both tasks, as in the networks of Kirkpatricka and team.

The connection of the threshold neurons with the linear classifiers creates preconditions for searching for ways to apply linear separation in neural networks. We have already mentioned that for each specific problem with a bounded, closed and linearly separable area of the input data, an infinite number of neurons can be constructed, which are genetically prepared to solve it. Let us look at things from a different perspective and ask ourselves the question—*is it possible for a threshold neuron to be constructed in a way that allows it to solve several different problems?*

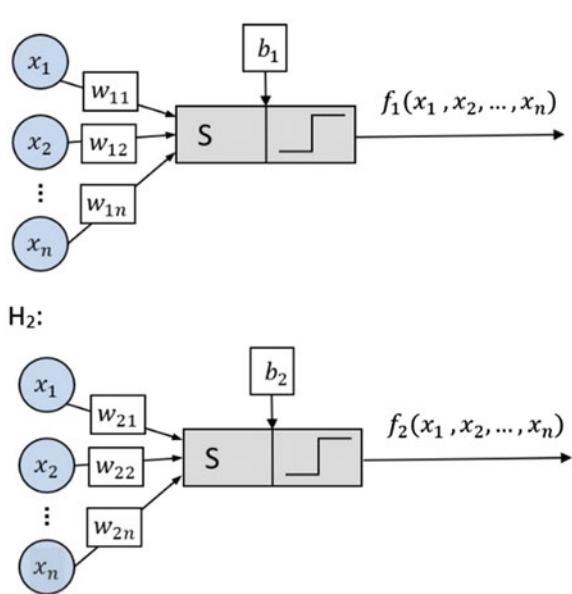
Let two problems be given related to finding the binary solutions of different functions of n-tuples:  $f_1 = f_1(x_1, x_2, \dots, x_n)$  и  $f_2 = f_2(x_1, x_2, \dots, x_n)$ . Let the domains D.O<sub>1</sub> and D.O<sub>2</sub> of the two functions be compact and linearly separable with respect to the n-tuples  $(x_1, x_2, \dots, x_n)$ , for which  $f_1$  and  $f_2$  return “0” and, respectively, “1”. Under these conditions, it follows that there are threshold neurons  $H_1$  and  $H_2$ , which successfully solve the two problems (Fig. 7) in the following way:

$$H_1 : f_1(x_1, x_2, \dots, x_n) = g\left[\sum_{i=1}^n w_{1i}x_i + b_1\right]$$

$$H_2 : f_2(x_1, x_2, \dots, x_n) = g\left[\sum_{i=1}^n w_{2i}x_i + b_2\right]$$

However, is there a vector  $(w_1, w_2, \dots, w_n)$ , and a value for  $b$ , with which both problems can be solved by a single neuron? What would such a threshold neuron look like? If we submit only the variables  $(x_1, x_2, \dots, x_n)$  at the input, will the sought TLU know what to do with them? Should we use this data to calculate the function  $f_1(x_1, x_2, \dots, x_n)$ , or should we use the inputs provided to calculate the value on  $f_2(x_1, x_2, \dots, x_n)$ ? Obviously, along with the input data, the neuron needs

**Fig. 7** Threshold neurons  
 $H_1$  and  $H_2$ , solving two problems



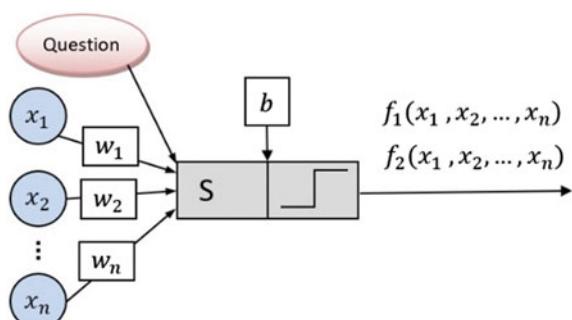
another input to get a question: “How much is  $f_1 = f_1(x_1, x_2, \dots, x_n)$ ?” or the task “Calculate  $f_2 = f_2(x_1, x_2, \dots, x_n)$ .”

The purpose at first glance is to find the weights  $\{w_i\}_{i=1}^n$  and the threshold  $b$ , for which:

$$g\left[\sum_{i=1}^n w_i x_i + b\right] = \begin{cases} f_1(x_1, x_2, \dots, x_n), & \text{Question: “How much is } f_1(x_1, x_2, \dots, x_n)\text{?”} \\ f_2(x_1, x_2, \dots, x_n), & \text{Question: “How much is } f_2(x_1, x_2, \dots, x_n)\text{?”} \end{cases}$$

Now we have another input for the neuron, and that is “Question” (Fig. 8). It is also a variable input value to the sought neuron and is fed to its body through the

**Fig. 8** Structure of a neuron solving both problems simultaneously



dendritic tree, along with the other variables. Thus, with the emergence of the need for a question, the input vector now takes the shape  $\vec{X} (x_1, x_2, \dots, x_n, x_{n+1} = \text{question})$ , and the weight— $\vec{W} (w_1, w_2, \dots, w_n, w_{n+1})$ . That is why we are actually looking for

$$g \left[ \sum_{i=1}^{n+1} w_i x_i + b \right] = \begin{cases} f_1(x_1, x_2, \dots, x_n), & x_{n+1} = \text{"How much is } f_1(x_1, x_2, \dots, x_n)?" \\ f_2(x_1, x_2, \dots, x_n), & x_{n+1} = \text{"How much is } f_2(x_1, x_2, \dots, x_n)?" \end{cases}$$

We have already discussed that the qualitative characteristics of objects can be represented by quantitative values. And since at the input of the neuron there is no way to ask the question  $x_{n+1}$  directly by using some linguistic construction, we need a quantitative interpretation that the neuron can process correctly. So, let

$x_{n+1} = 0$ , If we want to ask the question “How much is  $f_1(x_1, x_2, \dots, x_n)$ ?”

and

$x_{n+1} = 1$ , if the question is “How much is  $f_2(x_1, x_2, \dots, x_n)$ ?”

This way, we look for weights  $\{w_i\}_{i=1}^{n+1}$  of a neuron with  $n + 1$  inputs  $x_1, x_2, \dots, x_n, x_{n+1}$  and a threshold  $b$ , for which:

$$g \left[ \sum_{i=1}^{n+1} w_i x_i + b \right] = \begin{cases} f_1(x_1, x_2, \dots, x_n), & x_{n+1} = 0 \\ f_2(x_1, x_2, \dots, x_n), & x_{n+1} = 1 \end{cases}$$

The ability of the neuron  $H_1$  to calculate correct values for the function  $f_1(x_1, x_2, \dots, x_n)$  means that there is a linear classifier of the domain D.O<sub>1</sub> of the input variables of the first problem, represented by a plane

$$S_1 : \sum_{i=1}^n w_{1i} x_i + b_1 = 0,$$

where  $x_i$  are the coordinates of any point  $M \in S_1$ .

Adding the question  $x_{n+1}$  requires a transition to  $(n + 1)$ —tuple dimensional space of attributes and a requirement for linear separation of D.O<sub>1</sub> through

$$S_1 : \sum_{i=1}^{n+1} w_{1i} x_i + b_1 = 0, \quad \text{for } x_{n+1} = 0. \quad (6)$$

In a similar way for the plane  $S_2$ , which divides linearly D.O<sub>2</sub> in the  $(n + 1)$ —tuple dimensional space of the attributes, we have:

$$S_2 : \sum_{i=1}^{n+1} w_{2i} x_i + b_2 = 0, \quad \text{for } x_{n+1} = 1 \quad (7)$$

where  $x_i$  are the coordinates of any point  $M \in S_2$ .

The domain for the threshold neuron we are looking for is:

$$D.O = D.O_1 \cup D.O_2 \quad (8)$$

The neuron  $H$  that solves the two problems would exist only if this common D.O is linearly separable. However, given that D.O consists of points in the coordinate plane  $Ox_1x_2 \dots x_nx_{n+1}$ , as  $x_{n+1} = 0$  and points in its parallel  $Ox_1x_2 \dots x_nx_{n+1}$ , as  $x_{n+1} = 1$ , from (6)–(8) it follows that for neurons  $H_1$  and  $H_2$  a very important condition must be imposed: The existence of parallel linear classifiers of the type (6) and (7), through which to build the sought plane dividing the common D.O. This condition means that it is necessary to have surfaces  $S_1$  and  $S_2$ , with the representation (6) and (7), respectively, for which the condition of parallelism is fulfilled:

$$\frac{w_{11}}{w_{21}} = \frac{w_{12}}{w_{22}} = \dots = \frac{w_{1n}}{w_{2n}}, b_1 \neq b_2 \quad (9)$$

allowing us to construct new planes separating D.O of the same class of parallel planes

$$S : \sum_{i=1}^{n+1} w_i x_i + b = 0, x_{n+1} = \lambda, \lambda \in R$$

In other words, from the existence of parallel surfaces  $S_1$  and  $S_2$ , with the representation (6) and (7) follows the possibility of constructing a family of parallel planes containing  $S_1$  and  $S_2$  ( $\lambda = 0$  or  $1$ ) and linearly separating D.O.

Let us look at a specific example—the Boolean functions “AND” and “OR” with the truth tables given in Table 1. The domains of the two functions are linearly separable with respect to the pairs  $(X_1, X_2)$  returning “0” or “1”. Let  $S_1$  be one of the infinite sets separating D.O<sub>1</sub> lines:

$$S_1 : 0.3x_1 + 0.2x_2 - 0.4 = 0.$$

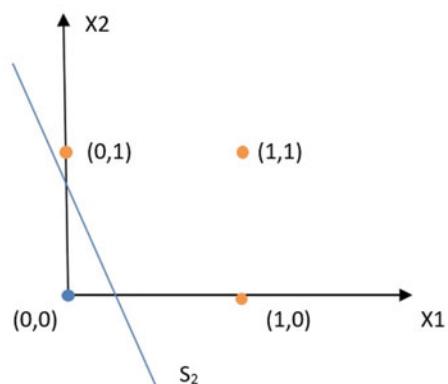
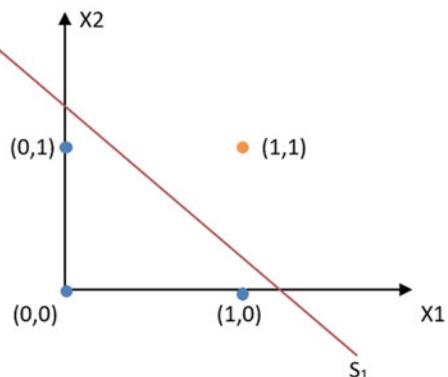
Then there is a threshold neuron  $H_1$ , solving the function  $f_1(x_1, x_2) = x_1 \wedge x_2$ , with weights and threshold indicated by the coefficients of  $D_1$  (Fig. 9). Similarly, if we look at the Boolean “OR” function and use the classifier:

$$S_2 : 0.75x_1 + 0.9x_2 - 0.3 = 0$$

Through it we can form a neuron  $H_2$ , solving the function  $f_2(x_1, x_2) = x_1 \vee x_2$ . So, we have two problems

$$f_1(x_1, x_2) = x_1 \wedge x_2 \text{ and } f_2(x_1, x_2) = x_1 \vee x_2,$$

**Fig. 9** Linear separability of D.O of Boolean functions “AND” and “OR”

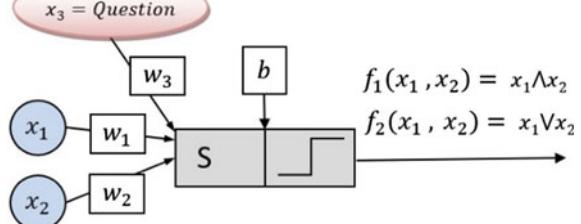


which we can solve quite precisely through two different neurons. Our goal is to build a neuron H, which is capable of solving both problems (Fig. 10).

The classification lines  $S_1$  and  $S_2$  have angular coefficients, respectively,  $k_1 = -\left(\frac{w_{11}}{w_{12}}\right) = -1.5$ , and  $k_2 = -\left(\frac{w_{21}}{w_{22}}\right) = -0.8333$ . Obviously, they are not parallel. We cannot build a plane through them that divides in an appropriate way

$$\text{D.O} = \text{D.O}_1 \cup \text{D.O}_2$$

**Fig. 10** Structure of a neuron that can solve two problems related to Boolean functions “AND” and “OR”



One possible solution is to keep looking for other linear classifiers for the two functions until we find parallel ones. Another solution is to use the condition of parallelism (9), which in our case has the following representation:

$$\frac{w_{11}}{w_{21}} = \frac{w_{12}}{w_{22}}, b_1 \neq b_2$$

Therefore, we can choose the linear classifier  $S_2$ , in such a way that

$$w_{22} = w_{21} \left( \frac{w_{12}}{w_{11}} \right) \quad (10)$$

So with the correction (10) of  $w_{22}$ , we have:

$$S_2 : 0.75x_1 + 0.5x_2 - 0.3 = 0$$

We can easily see that this line is still a linear classifier for the Boolean “OR” function, with an angular coefficient  $k_2 = -\left(\frac{w_{21}}{w_{22}}\right) = -1.5$ , i.e.,  $S_1 \parallel S_2$ .

Let point  $M_1$  and point  $M_2 \in S_1$ , while point  $M_3$  and point  $M_4 \in S_2$ . The choice of these points does not matter much. It is enough to concretize them clearly, so that belonging to the two lines, through them to form the vectors with which we will construct a plane containing  $S_1$  and  $S_2$ . Let:

$$\begin{aligned} & \text{point } M_1 \left( x_1 = 0.5, x_2 = \left( (-1) \frac{w_{11}}{w_{12}} \right) x_1 - \frac{b_1}{w_{12}} = 1.25, x_3 = 0 \right) \\ & \text{point } M_2 \left( x_1 = -0.3, x_2 = \left( (-1) \frac{w_{11}}{w_{12}} \right) x_1 - \frac{b_1}{w_{12}} = 2.45, x_3 = 0 \right) \\ & \text{point } M_3 \left( x_1 = 0.1, x_2 = \left( (-1) \frac{w_{21}}{w_{22}} \right) x_1 - \frac{b_2}{w_{22}} = 0.45, x_3 = 1 \right) \\ & \text{point } M_4 \left( x_1 = 0.6, x_2 = \left( (-1) \frac{w_{21}}{w_{22}} \right) x_1 - \frac{b_2}{w_{22}} = -0.3, x_3 = 1 \right) \end{aligned}$$

Let us now form the vectors  $\vec{p} = \overrightarrow{M_3 M_1}$  and  $\vec{q} = \overrightarrow{M_4 M_2}$ . We have

$$\begin{aligned} & \vec{p}(p_1 = -0.4, p_2 = -0.8, p_3 = 1), \text{ and} \\ & \vec{q}(q_1 = 0.9, q_2 = -2.75, q_3 = 1). \end{aligned}$$

Then, if the plane  $S$  is defined by the vectors  $\vec{p}$  and  $\vec{q}$ , and the point  $M_4$ , then

$$S : Ax_1 + Bx_2 + Cx_3 + D = 0,$$

as

$$A = p_2 q_3 - p_3 q_2, B = p_3 q_1 - p_1 q_3, C = p_1 q_2 - p_2.$$

For the free member D, we have:

$$D = -Ax_1 - Bx_2 - Cx_3,$$

where  $(x_1, x_2, x_3)$  are the coordinates of point  $M_4$ .

Given the specific values of the coordinates of the vectors  $\vec{p}, \vec{q}$  and the point  $M_4$ , it follows that we can construct a linear classifier of

$$D.O = D.O_1 \cup D.O_2,$$

which has the representation:

$$S : 1.95x_1 + 1.3x_2 + 1.82x_3 - 2.6 = 0,$$

and the corresponding threshold neuron has the construction as shown in Fig. 11.

Let us recall that

$$\text{Out} = g \left[ \sum_{i=1}^3 w_i x_i + b \right] \quad (11)$$

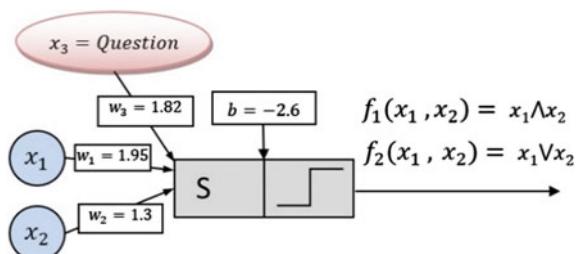
where  $x_3 = 0$ , if we want to ask a question to the neuron “How much is  $f_1(x_1, x_2) = x_1 \wedge x_2$ ”, and  $x_3 = 1$ , to demand calculation of the value of  $f_2(x_1, x_2) = x_1 \vee x_2$ .

We will do a check with two specific examples. Let us pass the pair of logical variables  $(x_1 = 0, x_2 = 1)$  at the input of the neuron and pose the question “How much is  $f_1(x_1, x_2) = x_1 \wedge x_2$ ”. In the dendritic tree, we have input stimuli:

$$x_1 = 0, x_2 = 1, x_3 = 0.$$

Then, according to (11), a signal propagates along the axon of the neuron

**Fig. 11** Threshold neuron corresponding to the found classifier  $S$



$$\text{Out} = g \left[ \sum_{i=1}^3 w_i x_i + b \right] = g(-1.3) = 0$$

Now let us ask the question to calculate the value of  $f_2(x_1, x_2) = x_1 \vee x_2$ . At the same values of  $x_1$  and  $x_2$ , at the input of the neuron, we have:

$$x_1 = 0, x_2 = 1, x_3 = 1.$$

and on the axon—a signal:

$$\text{Out} = g \left[ \sum_{i=1}^3 w_i x_i + b \right] = g(0.52) = 1$$

Other cases can be checked in a similar way.

## 4 Conclusion

Examining the linear separability of the data and the relation of the threshold neurons with the classifiers of the classes, we showed that it is possible to construct a neuron that can solve two different problems without the need for an intermediate change in its weights. But how effective and useful is a neuron that calculates several functions? On the one hand, the use of such a neuron saves the use of neural structures; therefore—memory, and on the other hand—this leads to an increase in the number of operations in the body of the neuron by two.

In conclusion, we should note that summarizing the presented ideas and results, we can find a neuron that solves a whole class of an infinitely number of problems with domains that are linearly separated from the found surface S. The only condition is that for each individual problem there is separability with linear classifiers which are parallel to each other.

**Acknowledgements** The work is partly funded by the MU21-FMI-004 project at the Research Fund of the University of Plovdiv “Paisii Hilendarski.”

## References

1. Gelig, A., Mateev, A.: Introduction to the Mathematical Theory of Learned Recognition Systems and Neural Networks. St. Petersburg State University (2014)
2. Bauckhage, C., Cremers, O.: Lecture Notes on Machine Learning: Linear Separability, University of Bonn, Bonn (2019)
3. Kowalczyk, A.: Support Vector Machines Succinctly, Syncfusion, Inc. (2017)
4. Picton, P.: Threshold logic: is there finally a solution? In: 2016 International Joint Conference on Neural Networks (IJCNN), pp. 45–51 (2016). <https://doi.org/10.1109/IJCNN.2016.7727179>
5. Elizondo, D.: The linear separability problem: some testing methods. IEEE Trans. Neural Netw. **17**(2) (2006)
6. El-Shahat, A.: Advanced Applications for Artificial Neural Networks. Georgia Southern University (2018)
7. Gorzałczany, M.: Essentials of Artificial Neural Networks. Physica, Heidelberg Online ISBN 978-3-7908-1801-7 (2002)
8. Gurney, K.: An Introduction to Neural Networks. Taylor & Francis e-Library, London and New York (2004)
9. Kruse, R., Borgelt, C., Braune, C., Mostaghim, S., Steinbrecher, M.: Threshold logic units. In: Computational Intelligence. Texts in Computer Science. Springer, London (2016). <https://doi.org/10.1007/978-1-4471-7296-3>
10. Minsky, M.L., Papert, S.A.: Perceptrons. MIT Press, Cambridge (1969)
11. Kirkpatricka, J., Pascanua, R., Rabinowitz, N., Veness, J.: Overcoming catastrophic forgetting in neural networks, DeepMind, N1C 4AG. United Kingdom, London (2017)
12. Yang, G., Joglekar, M., Song, H., Newsome, W., Wang, X.-J.: Task representations in neural networks trained to perform many cognitive tasks. Nat Neurosci **22**, 297–306 (2019)

# Chemicals Informatics: Search Structural Factors and Optimal Composites



Takashi Isobe and Yoshihiro Okada

**Abstract** Chemical industry pays much cost and long time to develop new compounds or composites that have the aimed properties. The developers need to efficiently discover initial candidates before simulation, actual synthesis, optimization, and evaluation. To meet their needs, we have developed chemicals informatics (CI) to efficiently discover potential candidates based on public literatures. Our system now has the data of 117 M existing compounds, and 61 properties extracted by analyzing a public chemical database linked to worldwide 33 M papers and 30 M patents in addition to 11 M new structures generated based on existing compounds. The compounds are shown as vectors including 41 organic and 70 inorganic features while the composites are presented as the combinations of them. The properties are extracted from the literatures using rule-based natural language processing (NLP). We also made our system evolve to predict 61 properties by each space of compound crossover for composite in addition to neighbor and structural crossover for compound written in the past paper. In the evaluation, CI could predict 16 properties at R<sub>2</sub> of 0.5–1.0 that was the same as or higher accuracy than a past paper. CI could also execute the prediction at 2 min over single GPU, which was 43 times faster than a past paper. We further showed the actual use case where users could extract structural factors and highly probable combinations of compounds for the binder of lithium-ion battery from 128 M to the fourth power. The combination had a discharge capacity of 20 Ah/g at maximum and had no patents.

**Keywords** Chemicals informatics · Materials informatics · Machine learning · Lithium-ion battery · Binder

---

T. Isobe (✉)

Hitachi High-Tech America, Inc, Pleasanton, CA 94588, USA  
e-mail: [takashi.isobe.sw@hitachi-hightech.com](mailto:takashi.isobe.sw@hitachi-hightech.com)

Y. Okada

Hitachi High-Tech Solutions Corporation, Chuo-ku, Tokyo 104-6031, Japan  
e-mail: [yoshihiro.okada.rj@hitachi-hightech.com](mailto:yoshihiro.okada.rj@hitachi-hightech.com)

## 1 Introduction

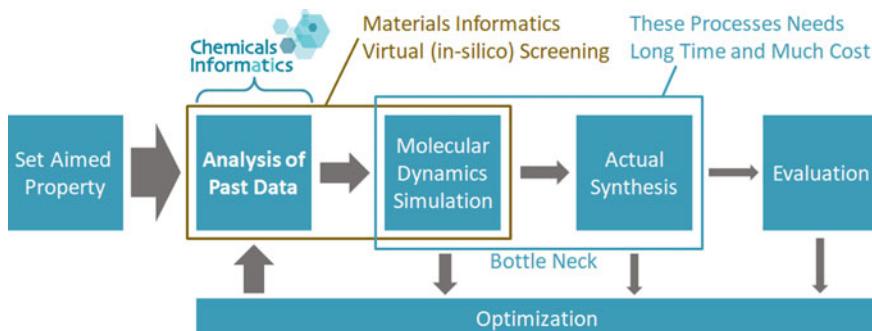
With the progress of globalization, the race to develop new chemicals is getting more fierce in recent years. The attempt to improve the development efficiency and keep the competitive advantage using materials informatics (MI) is becoming more popular. MI includes the wide range of techniques such as machine learning based on past data and dynamic molecular simulation based on quantum physics.

Chemical manufacturing industry has developed new chemicals through huge number of trials and errors while repeating experiments before MI attracts a lot of attention. Therefore, chemical industry often has paid the cost of 10–100 M dollars in addition to decades for developing new chemicals that have the aimed properties.

The chemical researchers and developers specify the aimed properties at first and then discover candidates based on analysis of past data. After preparing them, they simulate, actually synthesize, optimize, and evaluate them (Fig. 1). Analysis of past data and molecular dynamics simulation are categorized as virtual (*in-silico*) screening and also called as materials informatics. Molecular simulation and actual synthesis need long time and much cost for large scale of computer or manufacturing equipment. On average, 10,000 candidates are prepared and tested for each successful compound [1, 2]. The developers need to efficiently discover highly probable initial candidates before simulation, actual synthesis, optimization, and evaluation.

To meet the developers' needs, we have developed chemicals informatics (CI) to efficiently discover highly probable candidates of chemicals based on large number of public literatures.

Conventional analysis of past data [3] used private data or small scale of data limited to a specific field. For that reason, the applicable area is limited to the biased extension of existing chemicals included in the specific field. In addition, it was not available for the discovery of composites combining multiple compounds. Therefore, the practical use remains narrow scope. Moreover, no matter how hard chemical researchers used MI that focused on tuning the condition of manufacturing or the



**Fig. 1** Chemical development process and the need for “chemicals informatics”

ratio of mixture, they often had difficulty in approaching good results when the combination of compounds used for composites originally had low properties.

Chemicals informatics (CI) uses analysis based on less biased data using the wide range and large number of public literatures. In this paper, we also made our system evolve to predict wide range of 61 properties by each space of compound crossover for composite in addition to neighbor and structural crossover for compound written in past paper [4]. The newly developed techniques made it possible for researchers to efficiently discover the highly probable candidates of not only compounds beyond the extension of existing compounds but also the combinations of compounds used for composites that originally contributed to good properties.

## 2 Related Work

Virtual screening is classified into analysis of past data and molecular dynamics simulation.

Analysis of past data predicts properties and new structures using the set of data combining structures and properties. QSPR [5, 6]/QSAR [3, 7] (quantitative structure–property/activity relationship) is well known. QSPR/QSAR discovers probable candidates based on structural similarity to existing compounds that have good properties. Structural similarity is calculated as the distance between the numeric arrays that show structural features of compounds. Traditionally, the technique of fingerprint is used as the numeric array [8]. Recently, the technique of deep learning using neural network is also reported [3, 9, 10]. QSPR/QSAR is available for wide range of fields only if it can gather the set of data combining structures and properties. It can also achieve the screening at higher speed compared to molecular dynamics simulation. In biochemistry, it is classified as ligand-based virtual screening (LBVS) [11] using only ligands and chemical genomics-based virtual screening (CGBVS) [12] using both ligands and target proteins. CGBVS predicts biological activity of compounds against target protein based on the set of data arrayed from structures combining ligand and target protein.

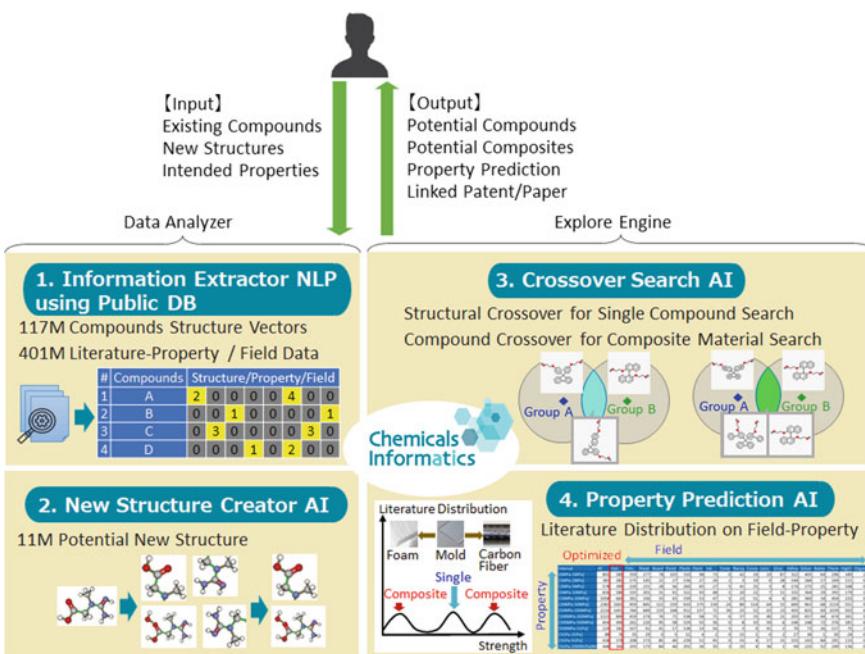
Molecular dynamics simulation predicts properties by calculating intermolecular forces based on potential energy function that describes the interactions between molecules while calculating position, speed, and acceleration by each molecular according to Newton's equation of motion. In biochemistry, 3D-docking simulation is used to judge the affinity of docking between ligands and proteins. It was classified into structure-based virtual screening (SBVS) [13, 14] using structural conformation and pharmacophore-based virtual screening (PBVS) [15, 16] using pharmacophore conformation. SBVS predicts the properties by simulating structural binding-conformation of ligands to binding-pocket of target protein whose biosynthetic pathway is known in details. PBVS creates the model of pharmacophore based on the forces of polarity, hydrogen bond, and Van der Waals between ligands and protein. They can predict the biological activity of new structural ligands whom biochemical researchers have no information about. On the other hand, their use is

limited to the case where the developers have the detailed information of detailed biosynthetic pathway and structure about target proteins or enzyme.

### 3 Chemical Informatics System

CI works as a cloud service of software as a service (SaaS) including various services [4, 17, 18] with network as a service (NaaS) [19]. Users input up to four groups of compounds or structural identifier with the aimed properties and fields into CI of cloud. CI comprises of a data analyzer and an explore engine. The explore engine predicts probable candidates by analyzing the user-input data based on the database that the data analyzer generates in advance. After then, users receive the output from the cloud.

The data analyzer comprises of our original information extractor natural language processor (NLP) and new structure creator artificial intelligent (AI). The explore engine comprises of our original crossover search AI and property prediction AI (Fig. 2).



**Fig. 2** Configuration of chemicals informatics

### ***3.1 Information Extractor NLP Using Public Database***

Our information extractor NLP runs our original processing engine using a rule-based algorithm over 20 core CPU. It preliminarily generates the data of 117 M compounds, 61 properties, and 119 fields extracted by analyzing a public chemical database [20] linked to worldwide 33 M papers and 30 M patents.

The properties and fields are extracted at rule base by analyzing worldwide papers and patents linked to compounds. Our NLP analyzes each phrase, sentence, paragraph, and document based on keywords or units that are registered in advance. It extracts the maximum and minimum values of properties with the field from each literature. The multiple units are unified into single one by each property. The extracted properties cover the wide range such as biological activity, toxicity, strength, hardness, thermal, electromagnetic, quantum, optics, water, gas, semiconductor, and synthesis. The extracted fields also cover the wide range such as polymer, molded, dye, liquid, electromagnetic, flexibility, thermal, water, luminescence, inorganic, metal, synthesis, agrochemical, pharmaceutical, and biochemistry. We manually sampled a few percent out of the extracted properties and fields and confirmed the accuracy of 95% or more.

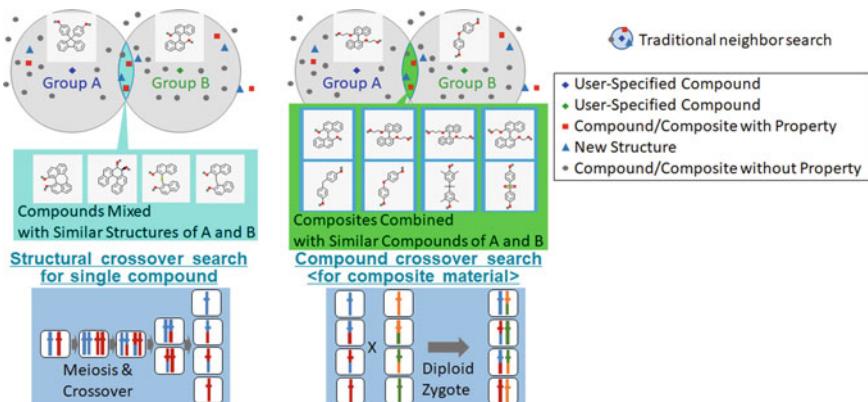
### ***3.2 New Structure Creator AI***

New structure creator AI generates new structures by adding a small functional group or replacing a part over existing compounds that have good properties. After then, it confirms whether the newly created structures coincide with all existing compounds or not. Only new structures that did not match all existing compounds are registered to CI database with existing compounds. CI database has 11 M new structures and 117 M existing compounds.

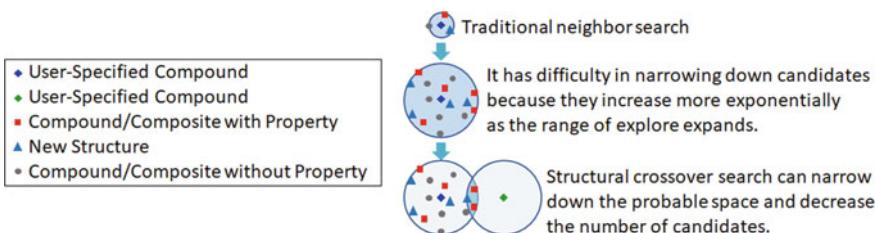
### ***3.3 Crossover Search AI***

Crossover search AI discovers the space of not only traditional neighbor compounds but also structural and compound crossover and predicts the properties by each space. These results can help users extract structural factors and highly probable combinations of structures and compounds that achieve the aimed properties and have no patents from the combination of 128 M to the fourth power.

Structural crossover searches the compounds mixed with similar structures of user-specified compounds using distances between vectors. It is similar to biological meiosis and crossover. Novel compound crossover searches the composites combined with similar compounds of user-specified compounds using distances between vectors. It is similar to biological diploid zygote (Fig. 3).



**Fig. 3** Comparison of structurally and compound mixed crossover search



**Fig. 4** Merit of structural crossover against traditional neighbor search

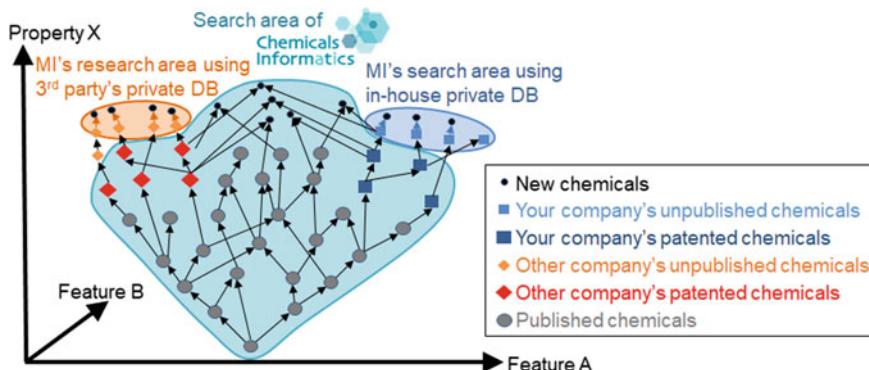
The traditional neighbor search had difficulty in narrowing down candidates because they increased more exponentially as the distance from the user-specified compound became longer (Fig. 4).

Structural crossover search can extract structural factors and highly probable combinations of structures that achieve the better properties by narrowing down the probable space and decreasing the number of candidates by structural crossover.

Newly developed composite search had difficulty in searching candidates from the large number of compounds' combinations because the number of compounds' combination increased exponentially as the power of combination increased.

Our compound crossover search for composite can rapidly discover highly probable combinations of compounds that achieve the aimed properties and have no patents from the combination of 128 M to the fourth power by searching the composites combined with the compounds similar to user-specified compounds. Execution time is 2–40 min depending on the number of combination that is 10 K—40 M.

The compound crossover search for composite is available for discovering highly probable combinations that contribute to the better properties in the wide range of fields such as copolymer, mixed materials, sinter, crystal, alloy, semiconductor, etching gas, electrolyte, and catalyst synthesis.



**Fig. 5** Crossover search following the tree diagram of evolution

CI's crossover search provides a fast and comprehensive way to search the vast space of compounds and composites that are not covered with private data and overlooked by both in-house and other companies. By repeating crossover search following the tree diagram of evolution, users can accomplish a comprehensive search for highly probable combinations of compounds and composites that have the better properties and no patents (Fig. 5).

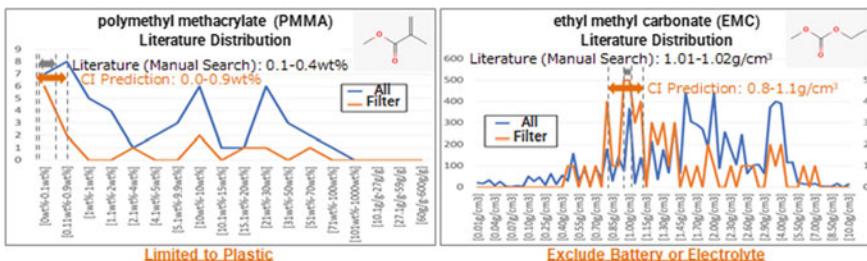
### 3.4 Property Prediction AI

Property prediction AI predicts 61 properties based on the distribution of the number of literatures linked to compounds or composites included within each space. The number of literatures is distributed between the variety of fields and the value range of properties. Users can improve the accuracy in predicting properties by specifying the filter to limit or exclude 119 fields based on their target (Fig. 6). Especially, the filter can improve the accuracy in predicting properties for single materials that are

		Field																			
		Opt.	All	Film	Sheet	Board	Panel	Plastic	Paint	Ink	Toner	Nano	Cosmi	Lens	Glue	Adhes	Solver	Batter	Therm	HighT	Organ
Interval																					
[0MPa-1MPa]	892	589	359	277	78	103	310	98	73	0	42	19	19	87	322	405	64	296	189	2	
[1MPa-2MPa]	383	282	175	145	22	27	136	27	20	2	3	19	0	28	144	184	17	169	111	0	
[2MPa-5MPa]	578	399	220	231	28	56	260	65	27	5	41	25	3	8	174	273	15	281	168	0	
[5MPa-10MPa]	836	580	333	301	31	92	322	93	48	1	20	22	7	11	332	364	29	391	179	1	
[10MPa-20MPa]	1058	665	361	336	33	43	339	53	47	5	23	51	4	6	281	360	34	454	233	0	
[20MPa-50MPa]	2383	1602	959	845	122	159	920	175	130	24	80	114	64	32	600	963	68	1114	551	9	
[50MPa-100MPa]	2216	1451	766	689	133	168	912	137	72	36	25	52	63	23	455	825	69	1039	503	6	
[100MPa-200MPa]	1223	849	419	274	76	71	538	58	71	6	37	33	40	21	203	437	44	674	341	5	
[200MPa-500MPa]	604	490	262	220	95	58	329	55	36	0	8	10	59	6	146	248	72	376	181	6	
[500MPa-1GPa]	224	182	107	92	41	17	138	14	16	5	3	0	41	2	70	73	26	117	75	1	
[1GPa-2GPa]	88	76	35	29	0	5	52	4	2	0	15	1	0	2	27	36	1	50	24	0	
[2GPa-5GPa]	438	378	208	173	85	40	239	53	45	0	55	4	17	31	155	243	84	291	133	0	
[5GPa-20000GPa(M)]	448	344	205	173	84	46	201	26	35	3	25	4	36	1	94	225	52	249	136	0	

Optimized by Filter

**Fig. 6** Field–property matrix used for property prediction



**Fig. 7** Literature distribution of PMMA water content and EMC density

used for wide range of fields. For composites that blend two groups of compounds, it is often unnecessary to specify the filter because their field is often limited when their blend is specified.

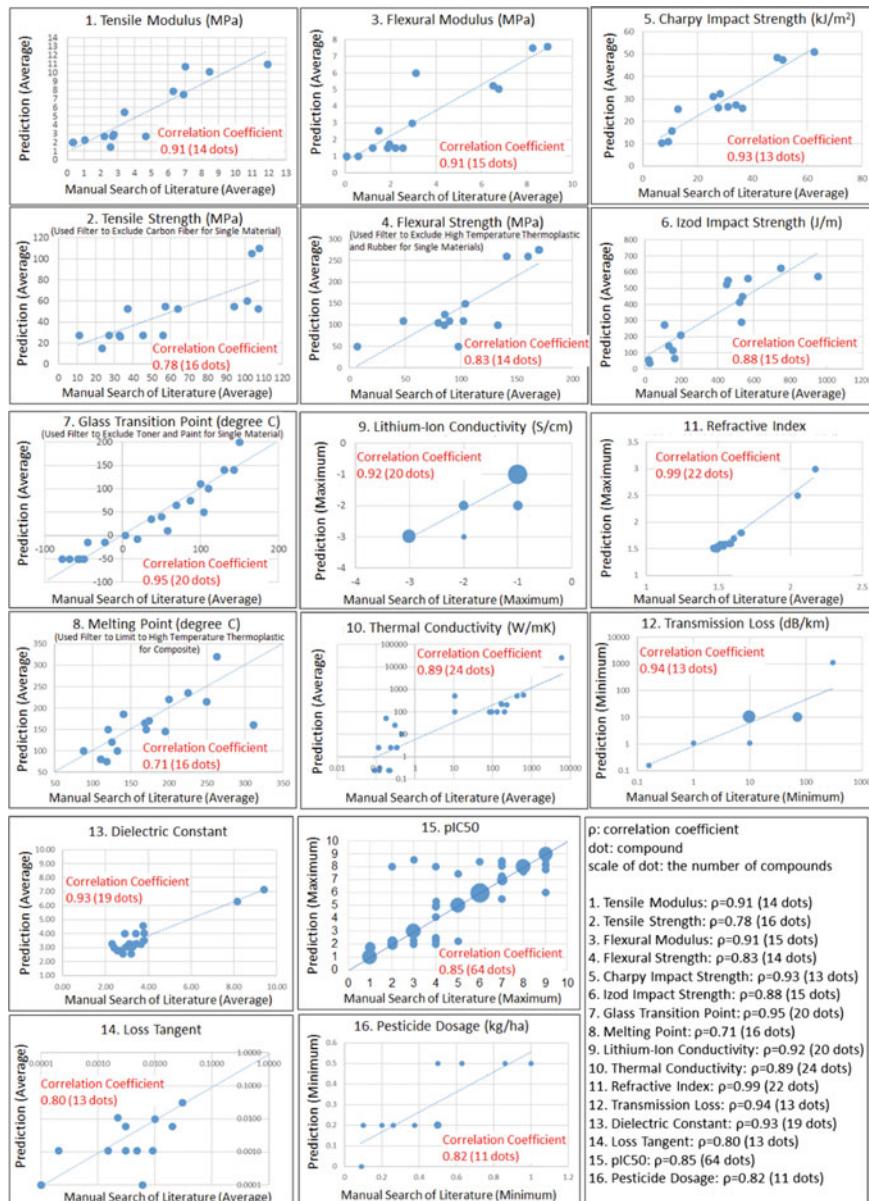
Figure 7 shows the distribution of the number of literatures to predict the water content of molded polymethyl methacrylate (PMMA) and the density of solvent ethyl methyl carbonate (EMC).

PMMA is used for manufacturing not only high absorbent materials such as contact lens, porous cosmetic particles, and water absorber but also low absorbent materials such as molded acrylic board. When users want to predict the water content for molded acrylic board of PMMA that has the wide range of uses, they can acquire the value of water content predicted for molded acrylic board of PMMA by specifying CI's filter to limit the field to plastic and using the literature distribution related to only plastic (Fig. 7). The range of predicted value included that of manually searched literatures.

EMC is mainly used as the electrolyte of lithium-ion battery including metals. On the other hand, it is also used for manufacturing some molded copolymers. If users want to predict the density for molded board made of EMC that has the wide range of uses, they can acquire the value of density predicted for molded board made of EMC by specifying CI's filter to exclude the field of battery or electrolyte and using the literature distribution not related to battery or electrolyte (Fig. 7). The range of predicted value included that of manually searched literatures.

Our past paper [4] has already reported the correlation coefficient of 0.85 between the predicted and literal values of IC<sub>50</sub> using 64 compounds. Currently, we have further confirmed that our CI additionally achieves the correlation coefficient of 0.71–0.99 (R<sup>2</sup> of 0.5–1.0) between the predicted and literal values of 15 properties that are often used in typical chemicals.

Figure 8 shows the relationship and correlation coefficient between the predicted and literal values using typical 13–29 chemicals such as polymer (polystyrene, polyethylene, polypropylene, polymethyl/ethyl/butyl/propyl acrylic/methacrylic acid, nylon-6/12, polyacetal, polycarbonate, polyvinylchloride, styrene butyl rubber (SBR), polyethylene/polybutylene terephthalate (PET/PBT)), inorganic materials (phosphorus sulfide/oxide, metal sulfide/oxide/nitride/halide/carbide), pesticides (Pyroxasulfone, Limsulfuron, Cycloxydime, Amicarbazone, Mefenacet, Imazapill,



**Fig. 8** Relationship between predicted and literal value

Metazachlor, Benomyl, Famoxadon, Zoxamide, Metominostrobin) and inhibitors having literal values acquired by manual search. The scale of dots shows the number of data. Each property used the common condition of filter for all evaluated compounds. If we tune the condition of filter for each compound, we can improve the accuracy much more.

A past paper [3] reported the average R<sub>2</sub> of 0.3–0.9 for the prediction of 15 biological activities. Our CI could predict 2 biological activities and 14 material properties at the same as or higher accuracy than the past paper.

Our CI executed this single prediction using 128 M compounds at 2 min over single GPU. Our screening achieved 43 times faster speed per single GPU compared to the result of 18 min for 1330 M compounds over 50 GPUs reported in the past paper [21] that did not support the prediction of biological activity or material property.

## 4 Evaluation in Actual Use Cases

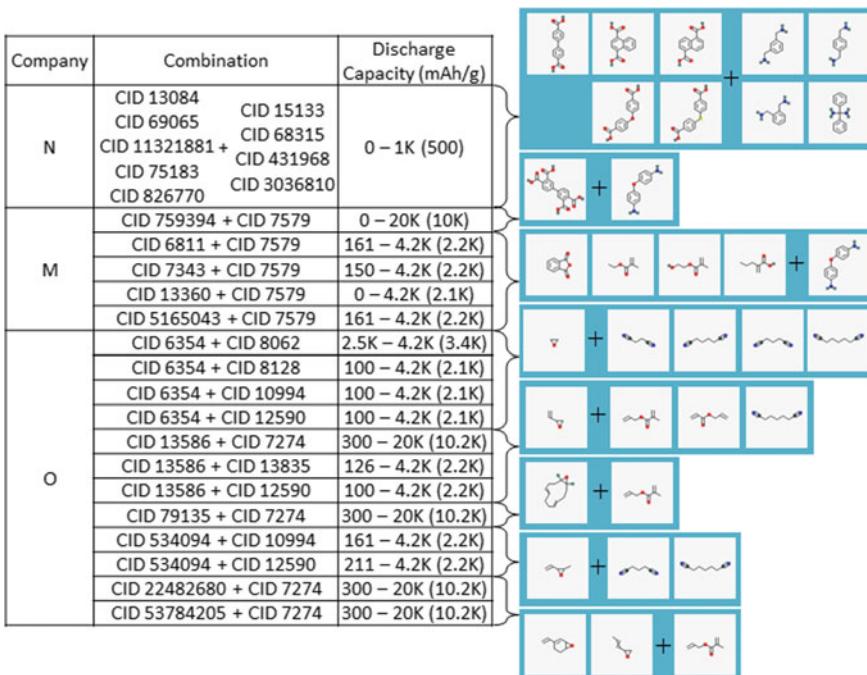
We evaluated CI in the actual use cases of binder for lithium-ion battery. Lithium-ion battery mainly comprises of anode, electrolyte, cathode, separator, and binder [22]. Especially, the binder is an important part to tie anode, electrolyte, and cathode tightly. Excellent binder can increase discharge capacity with discharge voltage and cycle count [23]. Furthermore, the needs for the binder of solid-state battery are increasing because of the impact and fire resistance required for vehicles.

In this paper, we explored structural factors and highly probable composites that had the new combination of compounds for the binder composite of solid-state battery that contributed to high discharge capacity using the combinations of compounds such as epoxy, acrylic acid, carboxylic acid, diamine, and polyimide written in patents filed by company T [24], M [25–27], O [28, 29], and N [30].

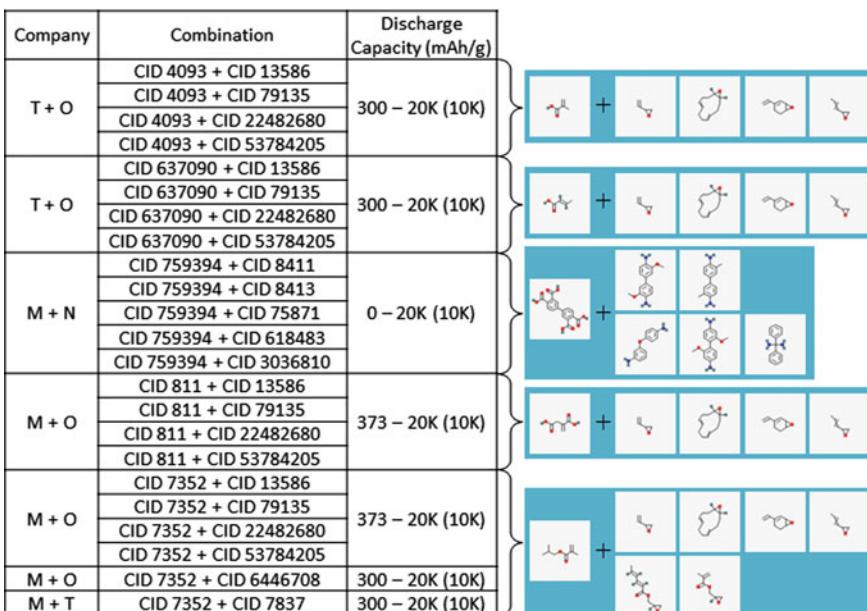
We executed not only crossover search, respectively, mixing compounds written in the patent filed by each company but also crossover search mixing all of them. From the result of CI, we extracted the dominant combination of compounds that contributed to high discharge capacity of 1000, 4000, and 20,000 mAh/g at maximum value in low voltage. Discharge capacity becomes larger as voltage becomes smaller.

The number of combinations with discharge capacity of 1000 mAh/g or more was 0 from the patent of company T, and 20 from the patent of company N. The number of combinations with discharge capacity of 4000 mAh/g or more was 5 from the patent of company M, and 12 from the patent of company O. The number of combinations with discharge capacity of 20,000 mAh/g or more was 1 from the patent of company M, and 8 from the patent of company O (Fig. 9).

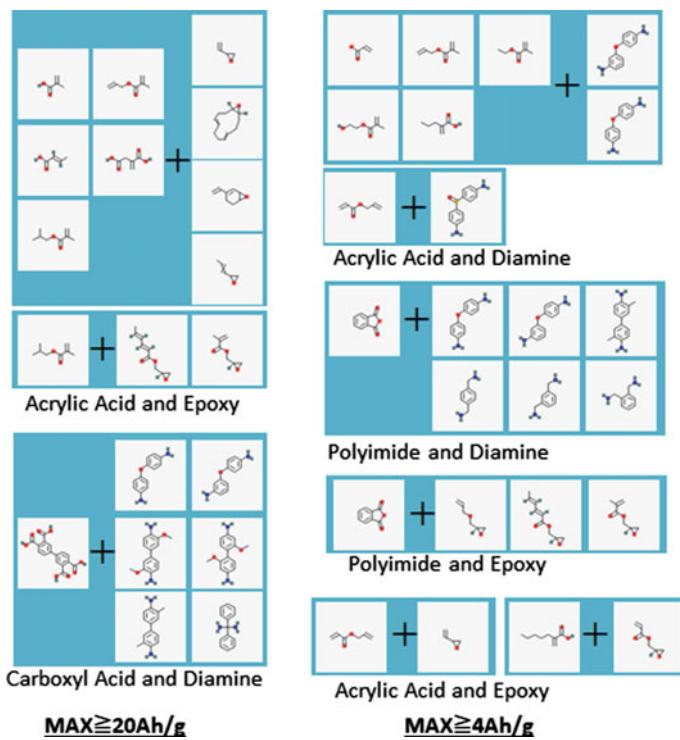
On the other hand, the number of combinations with discharge capacity of 20,000 mAh/g or more was 23 in the combinations mixed from patents of company T, M, O, and N (Fig. 10). The crossover search mixing the compounds of patents filed by multiple companies could extract much more highly probable candidates of combination that contributed to much better properties and was overlooked by four companies.



**Fig. 9** Combinations with good properties from patents filed by each company



**Fig. 10** Combinations with good properties from patents filed by multiple companies



**Fig. 11** Combinations that contributes good properties

Next, we extracted structural factors that contributed to the high discharge capacity. We extracted the combinations with maximum discharge capacity of 4000mAh/g or 20,000 mAh/g predicted by crossover (Fig. 11).

We found that the dominant combination that contributed to 4000 mAh/g was acrylic acid and diamine, polyimide and diamine, or polyimide and epoxy. We also found that the dominant combination that contributed to 20,000 mAh/g was acrylic acid and epoxy, or carboxylic acid and diamine.

From the neighborhood of dominant combinations of compounds that contributed to good properties (CID\_811 + CID\_79135), we could find new composites of the combination that had no literatures as shown in Fig. 12. They are highly probable candidates of composite with high discharge capacity. The manufacturing method can be acquired from patents linked in similar combination. If users confirm the aimed properties using the new combination, they can take competitive advantage by filing it as a new patent as soon as possible.

List of compounds in a promising space				Charge / discharge capacity (mAh / g) Interval, (number of patents), patent number, application	
No.	Promising compound structure diagram	Promising compound name (* 2)	Promising compound CID		
35		A Neighbor malonic acid	867	373mAh/g-200Ah/ [1]	Literatures exist in other combinations
		Combination			
36		B Neighbor (4E,8E)-13-oxa bicyclo[10.1.0]trideca-4,8-d iene	5357430	300mAh/g-200Ah/ [1]	Literatures exist in other combinations
53		A Neighbor 2-mercaptosuc cinic acid	6268	0mAh/g-200Ah/g 373mAh/g-600mAh 4.21Ah/g-200Ah/g [373mAh/g-600mA [373mAh/g-600mA [373mAh/g-600mA [373mAh/g-600mA	US2015357648 app-ZEON CORP US-10559828-B2 app↑ CN-104904042-A app↑ CN-104904042-B app↑
54		B Neighbor (4E,8E)-13-oxa bicyclo[10.1.0]trideca-4,8-d iene	5357430	300mAh/g-200Ah/ 373mAh/g-600Ah 4.21Ah/g-200Ah/g [373mAh/g-600mA [373mAh/g-600mA [373mAh/g-600mA [373mAh/g-600mA	US2015357648 app-ZEON CORP US-10559828-B2 app↑ CN-104904042-A app↑ CN-104904042-B app↑

**Fig. 12** New composite that has no literatures (literature blank zone)

## 5 Conclusion

We have developed CI to efficiently discover potential candidates of compounds or composites based on large number of public literatures. Our system has the data of 117 M existing compounds with 61 properties and 119 fields extracted by analyzing a public database linked to 33 M papers and 30 M patents in addition to 11 M new structures generated based on existing compounds. The compounds are shown as vectors including 41 organic and 70 inorganic features. The properties and fields could be extracted from literatures using our original rule-based NLP at the accuracy of 95% or more.

We also made our system evolve to predict wide range of 61 properties by each of crossover spaces combining compounds for composite in addition to spaces neighboring user-specified compounds and crossover spaces mixing their structures for compound written in our past paper [4]. These techniques newly enabled for users to extract structural factors and highly probable combinations of compounds for the composite that achieved aimed properties and had no patents from the combination

of 128 M to the fourth power. Execution time is 2 – 40 min depending on the number of combination that is 10 K – 40 M.

In the evaluation, CI could predict 16 properties at R<sub>2</sub> of 0.5–1.0 that was the same as or higher accuracy than a past paper [3]. CI could also execute the single prediction at 2 min over single GPU, which was 43 times faster than a past paper [21]. We could achieve the high-speed screening using small number of 111 features and the accurate prediction using simple rule-based NLP that analyzed each phrase, sentence, paragraph, and document based on keywords or units that are registered in advance.

The use case of binder for lithium-ion battery showed how to explore the new combination of compounds that contributed to high discharge capacity. As the initial combinations of compounds, we used composites written in patents filed by four companies. CI's crossover search mixing all compounds patented by these companies could extract much more highly probable candidates of dominant combinations that contributed to good properties and were overlooked by four companies. Moreover, we could find structural factors that contributed to good property in addition to new composite of combination that had no literatures.

CI supports not only the above-mentioned use case but also many other applications such as kinase inhibitor, medicine repurposed for COVID-19, low dielectric, optical resin lens, solid electrolyte, hard coating alloy, hydrophobic material, hydrophilic film, liquid black ink, ray-curable resin, fluorescent materials, pesticide, etching gas for semiconductor, natural adhesive, and biodegradable plastics. In the future, we will expand the scope of our CI's applications.

## References

1. Hambley T. W.: The Influence of Structure on the Activity and Toxicity of Pt Anti-Cancer Drugs. In: *Coordination Chemistry Reviews*, vol. 166, pp. 181–223 (1997)
2. Wolff, M., McPherson, A.: Antibody-Directed Drug Discovery. *Nature* **345**, 365–366 (1990)
3. Kato Y., Hamada S. and et. al.: Validation Study of QSAR/DNN Models Using the Competition Datasets. In: *Molecular Informatics* (2019)
4. Isobe T. and Okada Y.: Chemical XAI to Discover Probable Compounds' Spaces Based on Mixture of Multiple Mutated Exemplars and Bioassay Existence Ratio. In: *Bigdata 2020, LNCS*, vol. 12402, pp. 177–189 (2020)
5. Lu X. and et. al.: Quantitative Structure–Property Relationship (QSPR) Analysis of ZrO<sub>2</sub>-Containing Soda-Lime Borosilicate Glasses. In: *The Journal of Physical Chemistry B*, vol. 123 (6), pp. 1412–1422 (2019).
6. Neda A., Fatemeh S. and et. al.: Quantitative Structure- Property Relationship (QSPR) Investigation of Camptothecin Drugs Derivatives. In: *Combinatorial Chemistry & High Throughput Screening*, vol. 21(7), (2018).
7. Roy K., Kar S. and et. al.: Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment. In: *Academic Press*, pp. 455–462 (2015).
8. Willett P.: Similarity-based Virtual Screening Using 2D Fingerprints. In: *Drug Discovery Today*, 11 (23–24), pp. 1046–1053 (2006)
9. Zhavoronkov A., Ivanenkov Y. A. and et. al.: Deep Learning Enables Rapid Identification of Potent DDR1 Kinase Inhibitors. In: *Nature Biotechnology* 37, pp. 1038–1040 (2019)

10. Puri M., Pathak Y. and et. al.: Artificial Neural Network for Drug Design, Delivery and Disposition. In: Academic Press, pp. 3–13 (2016)
11. Kurczyk A., Warszycki D. and et. al.: Ligand-Based Virtual Screening in a Search for Novel Anti-HIV-1 Chemotypes. In: *J. Chem. Inf. Model.*, 55 (10), pp. 2168–2177 (2015)
12. Hamanaka M., Taneishi K. and et. al.: CGBVS-DNN: Prediction of Compound-Protein Interactions Based on Deep Learning. In: *Molecular Informatics*, 36 (1–2) (2017)
13. Johnson D. K. and Karanicolas J.: Ultra-High-Throughput Structure-Based Virtual Screening for Small-Molecule Inhibitors of Protein-Protein Interactions. In: *J. Chem. Inf. Model.* 56 (2), pp. 399–411 (2016)
14. Elovely K. M. and Doerkson R. J.: Docking Challenge: Protein Sampling and Molecular Docking Performance. In: *J. Chem. Inf. Model.*, 53 (8), pp. 1934–1945 (2013)
15. Matter H. and Poetter T.: Comparing 3D Pharmacophore Triplets and 2D Fingerprints for Selecting Diverse Compound Subsets. In: *J. Chem. Inf. Comput. Sci.*, 39 (6), pp. 1211–1225 (1999)
16. Wang J., Chen L. and et. al.: Pharmacophore-Based Virtual Screening and Biological Evaluation of Small Molecule Inhibitors for Protein Arginine Methylation. In: *J. Med. Chem.*, 55 (18), pp. 7978–7987 (2012)
17. Isobe T. and Okada Y.: Medical AI System to Assist Rehabilitation Therapy. In: ICDM 2018, LNCS, vol. 10933, pp. 266–271 (2018)
18. Isobe T. and Okada Y.: Rehabilitation XAI to Predict Outcome with Optimal Therapies. In: AIMS 2020, LNCS, vol. 12401, pp. 127–139 (2020)
19. Isobe T., Tanida N. and et. al.: TCP Acceleration Technology for Cloud Computing: Algorithm, Performance Evaluation in Real Network. In: ATC 2014, pp. 714–719 (2014)
20. Kim S., Thiessen P. A. and et. al.: PubChem Substance and Compound Databases. In: *Nucleic Acids Res.* 2016, 44 (D1), D1202–1213 (2016)
21. Christoph G., and et. al.: Virtual Screening in the Cloud: How Big Is Big Enough. In: *J. Chem. Inf. Model.*, vol. 60 (9), pp. 4274–4282 (2019).
22. Ritchie A. and Howard W.: Recent Developments and likely Advances in Lithium-Ion Batteries. In: *Journal of Power Sources*, vol. 162, issue 2, pp. 809–812 (2006)
23. Omar N. and et. al.: Lithium Iron Phosphate based Battery – Assessment of the Aging Parameters and Development of Cycle Life Model. In: *Applied Energy*, vol. 113, pp. 1575–1585 (2014).
24. Saito N. and et. al.: Method for Producing Crosslinkable Polymer or Salt Thereof. In: WIPO Patent, WO-2018180799-A1, (2017)
25. Ishida Y. and et. al.: Electronic Component. In: United States Patent, US-10319503-B2, (2015)
26. Morishita M. and et. al.: Lithium Ion Secondary Battery. In: WIPO Patent, WO-2016152505-A1, (2015)
27. Suzuki T. and et. al.: Conductive Paste, Multilayer Ceramic Electronic Component, and Method for Manufacturing Same. In: US Patent, US-9401244-B2, (2012)
28. Miura K. and et. al.: Inorganic Solid Electrolyte Secondary Battery Electrode and Inorganic Solid Electrolyte Secondary Battery. In: WIPO Patent, WO-2020110993-A1, (2018)
29. Miura K. and et. al.: Composite Solid Electrolyte, and Composite Solid Electrolyte Secondary Battery. In: WIPO Patent, WO-2020110994-A1, (2018)
30. Sugawara K. and et. al.: Binder for Forming Nonaqueous Battery Electrode, Binder Composition, Electrode Mixture Slurry Using Same, Electrode Structure and Nonaqueous Battery. In: WIPO Patent, WO-2016013611-A1, (2014)

# Analysis of Temperature Impacts on Material-Dependent Thermoelastic Damping in Simply Supported Rectangular Microplate Resonators Applying Size Effects



R. Resmi , V. Suresh Babu , and M. R. Baiju

**Abstract** In vibrating structures like resonators, various intrinsic and extrinsic energy dissipation mechanisms exist which limit the maximum achievable quality factor. Among the different types of energy losses, damping caused by thermoelastic effect is a crucial mechanism which arises due to the interaction among thermal and mechanical fields. Thermoelastic damping (TED) limits the maximum attainable quality factor ( $Q_{\text{TED}}$ ), and coupling between the strain and temperature fields varies with temperature. In this paper, the effects of temperature on energy dissipation due to TED in a simply supported rectangular microplate resonator is analyzed. When the devices are downsized, size effects should be incorporated, and instead of classical elasticity theories, modified couple stress theory (MCST) is applied to investigate the energy dissipations. The impact of temperature on energy dissipation with and without size effects is investigated by incorporating a length scale parameter ( $l$ ) which is made dimensionless by dividing it by the plate thickness ( $l/h$ ). The influence of temperature on  $Q_{\text{TED}}$  in rectangular microplates with and without size effects is analyzed using various structural materials (polySi, SiC, GaAs, diamond, and Si). The thermoelastic energy dissipation seems to be increased with rise in temperature, and with the incorporation of size scaling, owing to the impact of dimensionless length scale parameter, even at elevated temperatures, high  $Q_{\text{TED}}$  is achieved. Thermoelastic energy dissipation in rectangular plate is analyzed by pertaining MCST, and the impact of temperature on energy loss with size effects using  $l/h$  is numerically simulated using MATLAB 2015.

---

R. Resmi

University of Kerala, LBS Institute of Technology for Women, Thiruvananthapuram, Kerala 695012, India

e-mail: [resmi@lbsitw.ac.in](mailto:resmi@lbsitw.ac.in)

V. S. Babu

College of Engineering Trivandrum, APJ Abdul Kalam Technological University, Kerala 670644, India

M. R. Baiju

University of Kerala, Kerala, India

**Keywords** Quality factor · Thermoelastic energy dissipation · Size scaling · Length scale parameter · Microplate rectangular resonators

## 1 Introduction

Micro/nano-mechanical resonators have been evolved as the important components for various applications such as sensing, communication, and detection [1]. Micromechanical resonators have the inherent advantages such as high S/V ratio, large sensitivity, less mass, and low cost [2]. Quality factor (QF) is an important performance parameter to ensure high sensitivity and resolution of microstructures [3]. QF is a direct indication of dissipated energy due to various energy loss mechanisms such as thermoelastic damping, squeeze film damping, anchor damping, and Akhiezer damping. To achieve higher quality factors, it is critical to distinguish and understand the various energy losses and measures should be taken to reduce each loss mechanism. Accurate modeling of each energy loss mechanism is necessary, and QF associated to thermoelastic damping ( $Q_{TED}$ ) characterizes the impact of TED due to temperature dependence.

In resonators at micro-scales, damping due to thermoelastic energy loss is a prominent energy dissipation mechanism which is intrinsic in nature, and the maximum achievable quality factor limited by thermoelastic damping in the resonator is denoted by  $Q_{TED}$  [4]. In vibrating structures, regions of compression and elongation are formed in the material due to strain gradients which leads to temperature differences persuaded by coefficient of thermal expansion (CTE). Flow of thermal energy from regions of higher temperature to lower takes place due to nonuniform temperature distribution which leads to energy loss. The reason for the energy loss due to thermoelastic energy loss is the interaction between thermal and mechanical fields, and separate Eigen modes are generated. When the thermal expansion coefficient is zero, the coupling between the two fields can be eliminated leading to maximum value for  $Q_{TED}$ .

Thermoelastic damping is pointed out as a dominant energy loss mechanism in enormous research works [5]. In 1937, TED is first illustrated by Zener [6, 7]. Lifschitz and Roukes derived an exact closed-form expression for energy dissipation due to TED in slender beams [8]. The mechanic behaviors of other microstructures such as micro-plates and micro-shells have also been explored in various works [9].  $Q_{TED}$  and critical dimensions of circular plate resonators formulated on inplane and out of plane vibrations are analyzed by Resmi et al. [10]. Nayfeh and Younis derived an analytical expression for  $Q_{TED}$  of microplates regarding the structural mode shapes applying the perturbation method [11].

Conventional classical theories are insufficient to investigate the size scaling impacts on account of the absence of length scale parameter ( $l$ ) which is material dependent. The size dependencies in resonators are accurately predicted by higher-order continuum theories. When the devices are scaled down, modified couple stress theory (MCST) is evolved to accurately predict TED by Yang et al. MCST is the most

commonly used nonclassical elasticity theory which associates a single length scale parameter ( $l$ ) to include size scaling and its impacts [12]. Experimental observation of the size scaling effects of materials has been analyzed by downsizing the microstructures [13]. Razavilar et al. investigated the TED of a rectangular microplate on account of MCST and studied the significance of the length scale parameter [14]. Attenuation, thermoelastic frequency, and figure of merit associated with thermoelastic damping of rectangular plates using MCST are investigated by Resmi et al. [15]. Zhong et al. explored TED and the factors affecting in detail applying MCST in microplates [16]. Fang et al. presented analytical solutions for TED in microplates and analyzed three boundary conditions using Rayleigh's method [17]. TED in microplates based on the non-Fourier model like dual-phase-lag heat conduction was analyzed by Vahid Borjalilou et al. [18].

In this paper, the basic equations of thermoelastic energy dissipation and quality factor of microplates with size effects are given in Sect. 2. To include size effect, MCST is used in the analysis. The results and discussions are presented in Sect. 3 which analyzes the temperature effects on energy dissipation due to TED. The plots of the energy dissipation ( $Q^{-1}$ ) with temperature  $T$  for five various structural materials are illustrated in Sect. 3. The conclusions and future scope of the work are given in Sect. 4.

## 2 Energy Dissipation Related to TED of Rectangular Microplates

As far as classical theory is considered, stress and strain tensors are independent of the rotation vector; in nonclassical theories like MCST, the strain energy density is a function of both strain tensor and curvature tensor. The coupled thermoelastic fields lead to energy dissipation, and quality factor declines as a result, and the amount of energy loss can be estimated from the following expressions.

A microplate of width  $W$ , length  $L$ , and thickness  $h$  under plane stress condition is investigated with five structural materials at a temperature  $T_0 = 298$  K.

In the coupled thermoelastic fields, the equations to have the simple-harmonic vibrations as [16]

$$\begin{aligned} W(x, y, t) &= w_0(x, y)e^{i\omega t}, \\ \theta(x, y, z, t) &= \theta_0(x, y, z)e^{i\omega t} \end{aligned} \quad (1)$$

where  $t$  is the time,  $W$  denotes the mid plane displacement, and  $\theta$  is the parameter related to distribution of temperature in the microplate.

The relaxation strength of the material can be expressed as  $r' = E\alpha^2 T_0/Cp$  [19].

The microplate equation for representing the motion can be given by

$$\left[ M_0 + M_1 + \frac{M_0}{2} r'(1 + f(\omega)) \right] \nabla^4 W - \rho h \bar{\bar{W}} = 0 \quad (2)$$

where  $M$  is a parameter depending on material properties [19].

In Eq. (2),  $M_0 \nabla^4 W$  and  $\rho h \bar{\bar{W}}$  are the terms related to the CT model.

In Eq. (2), the size effect is represented by  $M_1 \nabla^4 W$ , and the effect of thermoelastic coupling due to downsizing is denoted by  $\frac{M_0}{2} r'(1 + f(\omega))$  [19]

The thermoelastic frequency

$$\omega_i = \pi^2 \sqrt{\frac{M'}{\rho h}} \left( \frac{m^2}{L^2} + \frac{n^2}{W^2} \right) = \omega_0 \sqrt{1 + \frac{M_1}{M_0} + \frac{\Delta_E}{2} [1 + f(\omega)]} \quad (3)$$

where  $\omega_0$  is the Eigen frequency pertaining to isothermal conditions [19].

From Eq. (3), real and imaginary parts in the presence of thermoelastic damping in the rectangular plate microresonators can be obtained.

To quantify energy dissipation, TED is expressed as [19]

$$Q^{-1} = 2 \left| \frac{\text{Im}(\omega)}{\text{Re}(\omega)} \right| \quad (4)$$

### 3 Results and Discussions

Rectangular plate resonators with width  $W = 200 \mu\text{m}$ , length  $L = 200 \mu\text{m}$ , and thickness  $h = 10 \mu\text{m}$  using various structural materials are analyzed, and numerical simulations are carried out with MATLAB 2015.

The material attributes of all the structural materials used in the rectangular plates are taken from [15], and all the properties are temperature variant. The damping associated with thermoelastic vibration changes with temperature; at lower temperatures, the vibrations are small. When an equilibrium temperature is selected ( $T_0 = 298 \text{ K}$ ), all mechanical and thermal properties can be almost approximated as constants. The thermal properties are very decisive in the temperature impacts analysis. When size effects are included, the mechanical properties are also remarkable in the analysis. The different structural materials used for the analysis of the impacts on energy dissipation are Si, diamond, polySi, GaAs, and SiC.

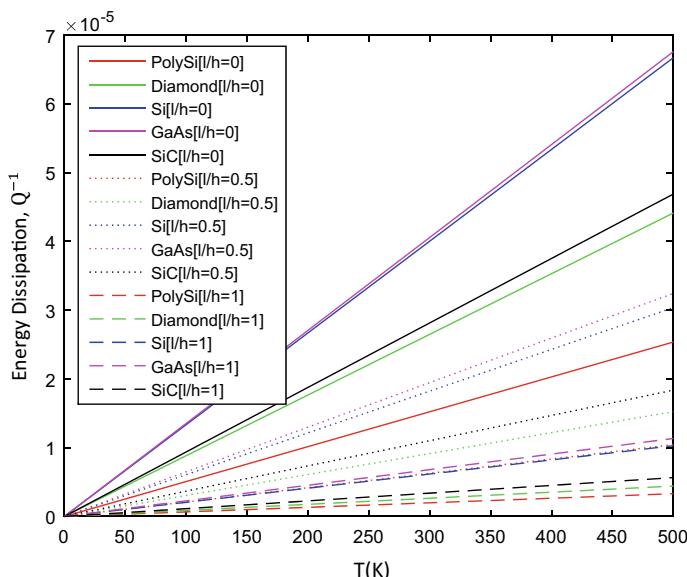
In our study, the size dependence effect is also included, which has been experimentally observed as an intrinsic property of materials when the structures are reduced to the order of microns and submicrons. The size effect in micro/nano-structures can be included by incorporating the material length scale parameter  $l$  and made dimensionless by taking  $l/h$ , which is known as the dimensionless length scale parameter.

The size effect study is characterized by changing the dimensionless length parameter  $l/h$  from 0 to 1, where  $l = 0$  denotes the classical plate model with thermoelasticity, and the varying  $l$  corresponds to the nonclassical continuum model of the modified couple stress theory (MCST). Size-dependent behavior is apparent as  $l/h$  increases and becomes maximized when the thickness is equal to the material length scale parameter, and the two theories converge as  $l/h$  decreases.

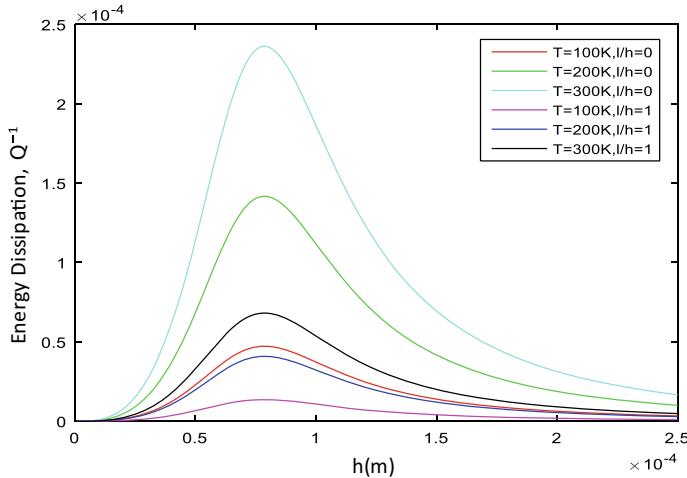
### 3.1 Effects of Temperature

When thermal impacts are considered, many material properties are explicitly temperature dependent. According to the expression of thermoelastic energy dissipation, maximum value of thermoelastic energy losses,  $Q_{\max}^{-1} = 0.494 r'$ , where  $r' r'$  is a temperature-dependent parameter. The effects of temperature are due to the temperature dependency of parameters such as Young's modulus, thermal expansion coefficient ( $\alpha_T$ ), and specific heat capacity at constant temperature,  $C_V$ . To study the temperature dependence of the energy loss on size effects, numerical simulations have been done, and the graphs are plotted as shown in Fig. 1.

Figure 1 shows the inverse of the quality factor versus temperature for different  $l/h$  values with five different materials. The value of  $Q^{-1}$  (inverse value of the quality



**Fig. 1** Variation of energy dissipation,  $Q^{-1}$ .vs. temperature, T for a rectangular microplate for different  $l/h$  ( $l/h = 0, 0.5, 1$ ) with vibrating mode (1,1);  $L = W = 200 \mu\text{m}$ ,  $h = 10 \mu\text{m}$ ; boundary condition-simply supported



**Fig. 2** Variation of energy dissipation,  $Q^{-1}$ .vs. thickness,  $h$  for a polySi-based rectangular microplate for different  $l/h$  ( $l/h = 0, 1$ ) and temperature,  $T$  ( $T = 100$  K, 200 K, 300 K) with vibrating mode (1,1),  $L = W = 200$   $\mu\text{m}$ ; boundary condition-simply supported

factor) presents the energy dissipation due to thermoelastic damping which enhances slightly by increasing the ambient temperature irrespective of the elasticity theories used.

As the temperature increases, the disparity between the two theories also increases. When size effects are included, TED is considerably lower than that owing to classical theories. While the length scale parameter is included, the microplate is stiffened more, and the TED is also diminished at elevated temperatures. When the size effect is included, a high quality factor is attained, even at high temperatures, using the current analysis. As  $l/h$  increases,  $Q_{\text{TED}}$  also increases, and, even at high temperatures, a high quality factor is obtained.

Figure 2 illustrates a comparison of the relation between the inverse quality factor and thickness for different fixed length scale parameters ( $l/h = 0, 1$ ) and temperatures ( $T = 100$  K, 200 K, 300 K) of a rectangular microplate. For a simply supported boundary condition, the temperature and quality factor are inversely proportional to each other, as depicted in Fig. 2.

The energy dissipation can be subsided and quality factor can be enhanced by selecting  $l/h = 1$  and operating at low temperatures. The sequence according to which the thermoelastic energy loss decreases in a rectangular microplate for distinct  $l/h$  value is  $l/h = 0 > l/h = 0.5 > l/h = 1$ . The impact of temperature is to shoot up the TED losses, and the effect can be dropped off by selecting high  $l/h$  values. The difference between the classical and nonclassical (MCST) theories increases at higher dimensionless length scale parameter values as illustrated in Fig. 2. According to the expression  $Q_{\max}^{-1} = 0.494 r'$ , the effects of temperature are traceable to the temperature reliance of  $r'$  which is associated to  $E$ ,  $\alpha_T$ , and  $Cv$ .

## 4 Conclusion

In micro/nano-scales resonators, due to different energy losses, the maximum attainable quality factor is limited. Thermoelastic damping is a prominent energy expending mechanism which depends on the association between temperature and strain fields. The energy dissipation on account of rise in temperature causes rise in thermoelastic damping which decreases the thermoelastic damping limited quality factor ( $Q_{TED}$ ). When the resonator sizes are scaled down, to precisely model the energy losses, a dimensionless length scale parameter ( $l/h$ ) is incorporated and found to be effective in reducing the thermoelastic damping. The related energy dissipation is estimated to be low even at higher temperature. According to the numerical results, when  $l/h$  is increased, the energy losses got diminished at higher temperatures also. In this work, by including a dimensionless length scale parameter, higher values of  $Q_{TED}$  were achieved even at elevated temperatures.  $Q_{TED}$  (the inverse of energy dissipation  $Q^{-1}$ ) diminishes with temperature for different materials for all values of dimensionless length scale parameters. The results obtained help engineers to fabricate microplate resonators with superior quality factors for high temperature applications.

## References

1. Unlu, M., Hashemi, M., Berry, C.W., Li, S., Yang, S.H., Jarrahi, M.: Switchable scattering meta-surfaces for broadband terahertz modulation. *Nat. Sci. Rep.* **4**, 5708 (2014)
2. Rebeiz, G.M.: RF MEMS. Wiley-Blackwell, Hoboken (2003)
3. Srikan, V.T., Swan, A.K., Ünlü, M.S., Goldberg, B.B., Spearing, S.M.: Micro-Raman measurement of bending stresses in micromachined silicon flexures. *J. Microelectromech. Syst.* **12**(6), 779–788 (2003)
4. Duwel, R.N., Candler, T.W., Varghese, K.M.: Engineering MEMS resonators with low thermoelastic damping. *J. Microelectromech. Syst.* **15**(6), 1437–1445
5. Kim, S.-B., Kim, J.-H.: Quality factors for the nano-mechanical tubes with thermoelastic damping and initial stress. *J. Sound Vibr.* **330**, 1393–1402 (2011)
6. Zener, C.: Internal friction in solids. I. Theory of internal friction in reeds. *Phys. Rev.* **52**(3):230–235 (1937)
7. Zener, C.: Internal friction in solids II. General theory of thermoelastic internal friction. *Phys. Rev.* **53**(1), 90–99 (1938)
8. Lifshitz, R., Roukes, M.L.: Thermoelastic damping in micro-and nanomechanical systems. *Phys. Rev. B* **61**(8), 5600 (2000)
9. Zuo, W., Li, P., Zhang, J., Fang, Y.: Analytical modeling of thermoelastic damping in bilayered microplate resonators. *Int. J. Mech. Sci.* **106**, 128–137 (2016)
10. Resmi, R., Suresh Babu, V., Baiju, M.R.: Analysis of thermoelastic damping limited quality factor and critical dimensions of circular plate resonators based on axisymmetric and non-axisymmetric vibrations. *AIP Adv.* **11**, 035108 (2021). <https://doi.org/10.1063/5.0033087>
11. Nayfe, A.H., Younis, M.I.: Modeling and simulations of thermoelastic damping in microplates. *J. Micromech. Microeng.* **14**(12), 1711–1717 (2004)
12. Yang, F., Chong, A., Lam, D.C.C., Tong, P.: Couple stress based strain gradient theory for elasticity. *Int. J. Solids Struct.* **39**(10), 2731–2743 (2002)

13. Park, S.K., Gao, X.-L.: Bernoulli-Euler beam model based on a modified couple stress theory. *J. Micromech. Microeng.* **16**, 2355–2359 (2006)
14. Razavilar, R., Alashti, R.A., Fathi, A.: Investigation of thermoelastic damping in rectangular microplate resonator using modified couple stress theory. *Int. J. Mech. Mater. Des.* **12**(1), 39–51 (2016)
15. Resmi, R., Baiju, M.R., Suresh Babu, V.: Thermoelastic damping dependent quality factor analysis of rectangular plates applying modified coupled stress theory. *AIP Conf. Proc.* **2166**, 020029 (2019). <https://doi.org/10.1063/1.5131616>
16. Zhong, Z.Y., Zhang, W.M., Meng, G., Wang, M.Y.: Thermoelastic damping in the size-dependent microplate resonators based on modified couple stress theory. *J. Microelectromech. Syst.* **24**(2), 431–445 (2015)
17. Fang, Y., Li, P., Zhou, H., Zuo, W.: Thermoelastic damping in rectangular microplate resonators with three-dimensional heat conduction. *Int. J. Mech. Sci.* **133**, 578–589 (2017)
18. Borjalilou, V., Asghari, M.: Small-scale analysis of plates with thermoelastic damping based on the modified couple stress theory and the dual-phase-lag heat conduction model. Springer-Verlag GmbH Austria, part of Springer Nature 2018. <https://doi.org/10.1007/s00707-018-2197-0>
19. Resmi, R., Suresh Babu, V., Baiju, M.R.: Impact of dimensionless length scale parameter on material dependent thermoelastic attenuation and study of frequency shifts of rectangular microplate resonators. In: 2021 IOP Conference Series: Materials Science and Engineering, vol. 1091, p. 012067

# Template Protection in Multimodal Biometric System Using Watermarking Approach



C. Vensila and A. Boyed Wesley

**Abstract** A biometric system examines the behavioral and physical traits of a person to recognize the unique qualities of that person. Multimodal biometric systems, which combine data from a variety of biometrics, have recently gotten a lot of interest because they can overcome the restrictions of unimodal biometric systems. Due to safety and privacy concerns, template preservation is necessary in biometric-based authentication systems. The privacy issues have been occurred because of the improper storing and usage of templates. Template protection techniques are being developed so that templates saved in databases remain secure and unauthorized users cannot access them. The main goals of these strategies are to achieve safety and privacy. This paper shows how to preserve templates in multimodal biometric systems via watermarking. Biometric features such as face and fingerprints are employed in this case. In this paper, discrete wavelet transform (DWT) watermarking technique is used to incorporate a fingerprint image into a face image, allowing for secure database storage. Both fingerprint and facial patterns are extracted from the watermarked image and matched against query images during authentication. The matching score acquired after fusing fingerprint and facial matching scores at the match score level determines the final choice. The testing results show that the watermarked image is perceptually identical to the original, and that the retrieved features are also same. The proposed method protects templates and generates a secure, dependable, and accurate authentication result.

**Keywords** Biometric security · Biometric template · Watermarking · Multimodal biometric system · Face recognition · Fingerprint recognition

---

C. Vensila (✉) · A. Boyed Wesley

Department of Computer Science, Nesamony Memorial Christian College, Marthandam Affiliated To Manonmaniam Sundaranar University, Tirunelveli, India

e-mail: [vensila\\_csa@nmcc.ac.in](mailto:vensila_csa@nmcc.ac.in)

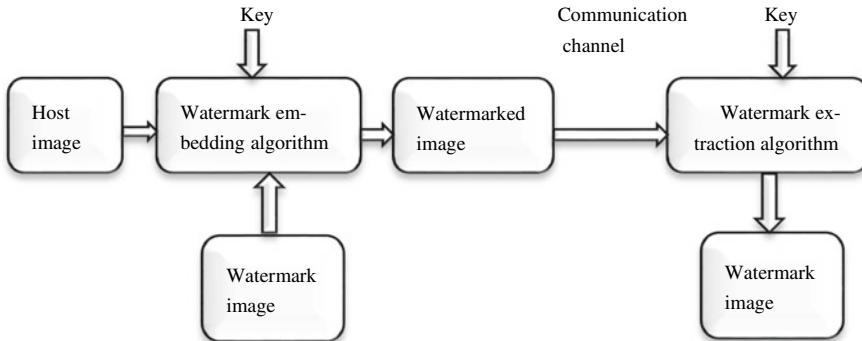
## 1 Introduction

Biometric technologies encompass an extensive variety of applications that can be used to authenticate a person's identity by analyzing physiological or behavioral features. Fingerprints, iris, facial features, signature, voice, finger knuckle, gait, palm print [1] and other biometric characteristics are being utilized in safety applications. Person identification is performed by unimodal biometric systems using only one source of biometric data. Spoof attacks, sensor noise, inter-class similarities, non-universality are all common difficulties in such systems. As a result of these practical issues, unimodal biometric systems have high error rates, making them unsuitable for implementation in safety-critical applications [2].

By adopting multimodal biometric systems, several of the issues that affect unimodal biometric systems can be avoided. They identify the problem of non-universality. It becomes progressively difficult for an impostor to fake various biometric features of an individual. It becomes increasingly hard for fraudsters to spoof multiple biometric traits of an individual. Furthermore, multi-biometric systems can be thought of as fault tolerant systems. Multi-biometric systems that combine data from multiple biometric sources are divided into six categories: multi-algorithm systems, multi-sensor systems, multi-sample systems, hybrid systems, multimodal systems, and multi-instance systems. The fusion scheme can be classified as feature level, sensor level, decision level, and score level fusion dependent on the amount of data that is fused. The score level fusion is the most often employed strategy in multi-biometric systems.

A biometric authentication system consists of modules like feature set extraction module, matcher module, sensor module, and decision module. Sensor module gets raw biometric data from numerous biometric traits (i.e., iris, face, fingerprint, finger knuckle print, hand geometry, voice, signature, and gait) that the person to be realized possesses essentially. The feature set extraction module takes the data received in the preceding module and extracts appropriate features from it. The matcher module compares these extracted traits to the templates stored in the system database at the time of acceptance in order to validate or recognize the person [3]. The highest damaging attack in a biometric system is the attack on templates stored in the system database.

The safety of biometric systems is at danger of numerous attacks which Ratha et al. [4] have divided into eight. The first type of attack includes the sensor module receiving a fake biometric trait. The second type of attack involves delivering previously collected data to the system. In the third type of attack, the feature set extraction module is used to build feature sets that the attacker chooses. The attacker replaces legitimate feature sets with those selected by the attacker in the fourth type of attack. The matcher module, which has been altered to give an artificially high identical result, is the target of the fifth category of attacks. Attacks on the template database are the sixth type of attack. The transmission channel between the matcher unit and the template database is targeted in the seventh type of attack, which leads to template misuse. Overriding the decision module's result is the eighth type of attack.



**Fig. 1** Watermark embedding and extraction

The process of digital image watermarking embeds data into an image in such a way that it cannot be easily erased or destroyed. Digital watermarking is a method of protecting copyright information in multimedia data by inserting data into the host media. With the help of the embedding algorithm and the key, the system takes the host image and embeds the watermark image into it. The system then obtains the watermarked image and transmits it across the communication channel. Using the watermark extraction algorithm and the key, the system finally extracts the watermark image. Figure 1 shows the above procedure.

The following are the primary contributions of this article:

- Watermarking is used to give protection to templates in a multimodal biometric system that includes fingerprint and face recognition.
- The most allowable biometric features are the fingerprint and the face. Because of its durability and distinctive patterns, the finger was chosen. Facial patterns are not constant, but they are distinct, and they work well with fingerprints in the development of a multimodal biometric system.

The remainder of the paper is organized as follows: Sect. 2 lists related works. The proposed method for integrating a fingerprint image into a face image is presented in Sect. 3. Section 4 describes the outcomes of the implementation. Finally, Sect. 5 describes the conclusion.

## 2 Related Works

This section discusses several related works on template security in biometric systems that have been mentioned in the literature. Khalil Zebbieche et al. [5] have suggested a wavelet digital watermarking technique to safeguard biometric data. They employed fingerprint images to conceal fingerprint minutiae points, ensuring that both hidden data and the host image were safe. This approach improves the security of fingerprint minutiae transmission and can also be used to secure the original fingerprint image,

but it is less resistant to attacks. Chander Kant et al. [6] developed a steganography-based approach to develop safety biometric systems. For encoding, a secret key is placed in the image itself, which can only be decoded by the legitimate user. This method increases the security of fingerprint minutiae transmission and can also be used to protect the original fingerprint image, but it is more vulnerable to attacks. Wadood abdul et al. [7] developed a multimodal biometric template protection technique combining fingerprint and facial templates with fusion at the score level. The discrete wavelet transform (DWT) is used to incorporate fingerprint characteristics in different directional sub-bands on facial images. Quantizing the mean values of the wavelet coefficients is used for watermark embedding and extraction.

Sanaa Ghouzali et al. [8] proposed a security approach for the storage of biometric templates in a multimodal biometric system. To produce cancelable biometric traits of the face and fingerprint minutia points, a logistic map and Torus Automorphism were used. This method yielded a 100% authentic acceptance rate but failed to detect image noise to maintain high performance rate. Ajai Kumar Gautam et al. [9] introduced the FLSL fusion approach and the modified deep learning neural network (MDLNN) classifier in order to improve the performance of the multimodal biometric (MMB) recognition system. The features level and scores level (FLSL) fusion method is used to combine the features in this case. As a result, the MDLNN receives the fused output of features and classifies the person as genuine or imposter. The accuracy of this approach is higher than any other classifier, but the system's performance is lower. Zebbiche et al. [10] developed a watermarking-based approach for protecting fingerprint images by placing a watermark within the area of ridges acquired after processing. The watermark is inserted in the DCT and DWT transform domains in this technique. In the face of attacks like mean filtering and additive white noise, this method provides increased robustness. Naima Bousnina et al. [11] developed an approach based on integrating face and fingerprint modalities via the combination of DTCWT and DCT frequency domain techniques to safe and improve the performance of the biometric identifications systems. This approach attaining the higher accuracy rate of 100%.

Naima Bousnina et al. [12] introduced DTCWT-DCT-based watermarking to integrate the fingerprint image into the face image. A 3D chaotic-map-based encryption approach is employed to protect the watermarked facial image in order to provide an extra layer of protection. This method achieves great performance while still meeting diversity and revocability requirements; however, it is unable to distinguish between real and false users. Mehwish et al. [13] proposed a convolutional neural network (CNN)-based model for the feature level fusion of fingerprint and online signature. This approach implements two different types of feature-level fusion algorithms for fingerprints and online signatures. The first technique, known as early fusion, combines fingerprint and online signature features before fully connected layers, whereas the second scheme known as late fusion, combines features after fully connected layers. The early feature fusion scheme achieved 99.10% accuracy, whereas the late feature fusion scheme achieved 98.35% accuracy, however, security is low.

### 3 Proposed Work

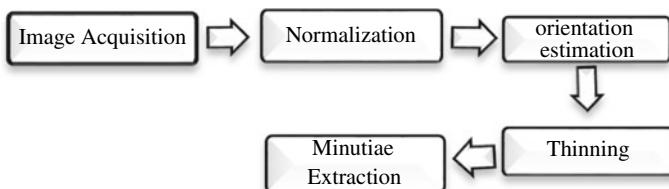
Natural characteristics that distinguish the identification of person are called biometric traits. They are persistent in nature, so they cannot be revoked and reissued if they are used by an unapproved person. Even though biometric systems are vulnerable to a variety of attacks, the most damaging attack is on biometric templates. In these systems, the lack of a replacement for stolen or duplicate templates poses a number of safety and isolation concerns. In order to address these concerns, this research proposes a watermarking-based solution for template protection in multimodal biometric systems. By inserting a fingerprint image into a facial image, the last template to be saved in a database is formed as a watermarked image. By preventing unwanted users from accessing secret fingerprint data and covering facial images, watermarked images produce high safety.

#### 3.1 Feature Set Extraction of Fingerprint

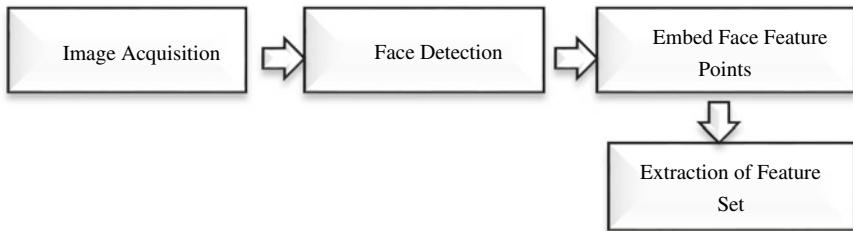
Preprocessing [14] of the captured image is carried out first in order to extract relevant information. The term normalization refers to the process of changing the range of gray-level values in an image such that it falls within a specified range of values. Normalization is a pixel-by-pixel process. The transparency of the ridge and indentation structures is unaffected (Fig. 2).

The major goal of normalization is to decrease the alteration in gray-level values across ridges and indentations, making succeeding computing steps easier. As a result, a variation threshold-based segmentation method can be applied. The orientation image is a fundamental aspect of fingerprint images that establishes unchanging coordinates for ridges and indentation in a given area. The directionality of ridges in a fingerprint image is represented by the orientation field. Based on median and increase in efficiency, the block orientation could be derived from the pixel gradient inclination.

Before extracting minutiae, the last step in the preprocessing procedure is thinning. Thinning is a morphological process in which the foreground pixels are gradually removed until they are only one pixel broad. The thinning operation is performed using two sub repetitions using a standard thinning algorithm. When the thinning



**Fig. 2** Methods for extracting fingerprint feature set



**Fig. 3** Methods for extracting face feature set

algorithm is applied to a fingerprint image, the association of the ridge design is preserved while a blueprint version of the binary image is created. The minutiae are extracted using this blueprint image. A  $3 \times 3$  window is used to scan the local region of every ridge pixel in the image to extract the minutiae. The crossing number (CN) concept is the most often used way of minutiae extraction. The CN value, which is described as half the summation of differentiation between pairs of adjoining pixels in the eight neighborhoods, is then calculated.

### 3.2 Feature Set Extraction of Face

Figure 3 depicts the basic techniques for extracting a face feature set. Face detection is done first to determine the position and boundaries of the face in the obtained image. The proportionate distances between the eyes, nose, mouth, and chin generate unique patterns in each facial image, making them essential feature points on the face. To realize faces, the spatial scattering among these important features, as well as the form of the face, is retrieved [15]. Eigenfaces are also known as main components are essential feature points. To obtain significant feature points from an obtained facia image, a technique based on eigenfaces could be utilized. It uses the principal component analysis (PCA) methodology to create a series of images [10]. The extracted feature set [16] or feature vector of some facial image is an estimation of the image on the decreased eigenface. The match score is calculated by computing the Euclidean length between the database template's eigenface coefficients and the query image.

### 3.3 Proposed Approach

As illustrated in Fig. 5, the suggested multimodal system comprises two phases: enrollment and verification. In the enrollment phase, the watermarking is applied to the user's biometric data. To create a watermarked image, insert a preprocessed fingerprint image into a preprocessed face image using the watermark inserting

method. Then, the watermarked face image is stored in the database as a reference image. During the authentication phase, recapture a watermarked image from database based on the claimed identification. To calculate the fingerprint matching score ( $MS_{Fingerprint}$ ), compare the fingerprints template retrieved from the watermarked image to the present fingerprint template. To obtain a face matching score ( $MS_{Face}$ ), compare the face template retrieved from the watermarked image to the present face template. Compute fusion of matching scores at the match score level, i.e.,  $(MS_{Fingerprint} + MS_{Face})$ , to generate the final matching score  $MS_{Final}$ . If the final matching score is greater than threshold, user is authentic. Otherwise, user is not authentic.

#### **Algorithm 1: Pseudocode of Enrollment Process**

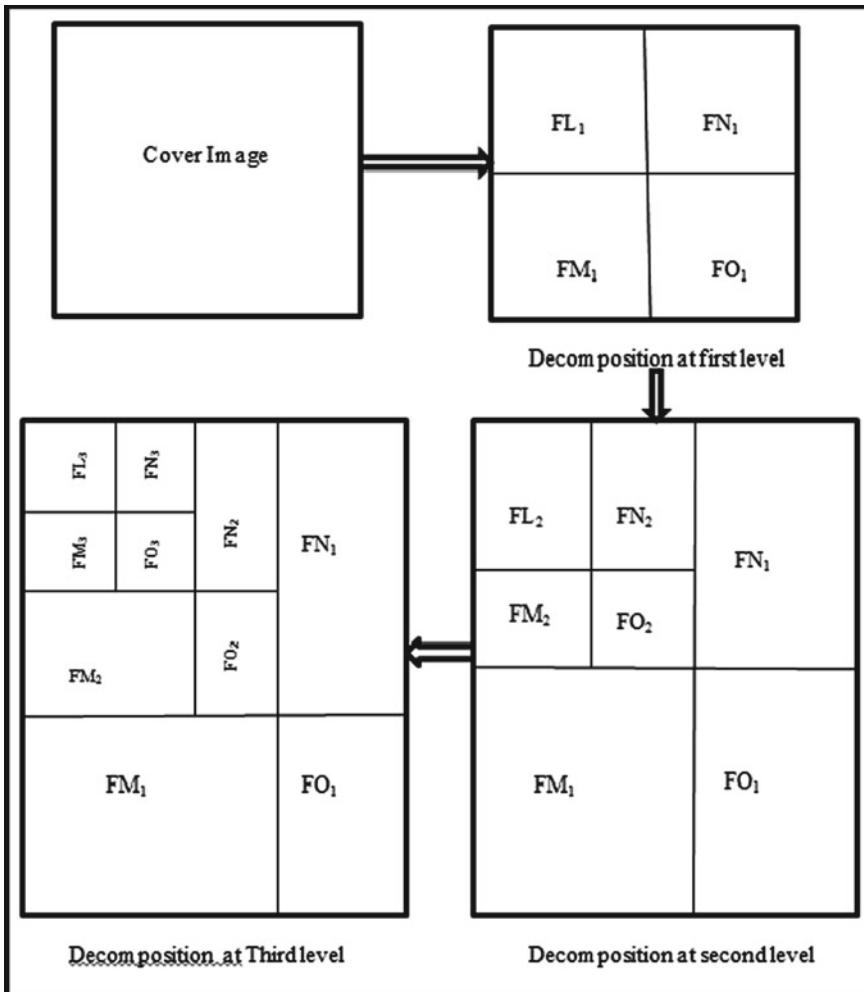
1. Input  $f_p, f_a$
2. Output  $w$
3. Initialize  $L$
4. For each  $f_p, f_a \in L$  do
5. Get  $pf_p, pf_a$
6. Insert  $pf_p$  to  $pf_a$
7. Get  $w'$
8. Update  $D$  with  $w'$

#### **Algorithm 2: Pseudocode of Authentication Process**

1. Input  $f_p, f_a$
2. Output  $A, A'$
3. Initialize  $Ef_p, Ef_a$
4. Initialize  $D$
5. Get  $Dp', Df''$
6. If ( $Dp' = Ef_p$ ) then
7. Get  $MS_{Fingerprint}$
8. If ( $Df' = Ef_a$ ) then
9. Get  $MS_{Face}$
10.  $MS_{Final} = MS_{Fingerprint} + MS_{Face}$
11. If ( $T < MS_{Final}$ ) then Return  $A$
12. Else Return  $A'$

### **3.4 Watermarking Using Discrete Wavelet Transform**

The cover image is divided into four frequency modules  $FL_1, FM_1, FN_1$ , and  $FO_1$  as shown in Fig. 4 using the discrete wavelet transform (DWT) watermarking process.  $FL_1$  is the smallest frequency module with estimation data. The parallel information is contained in  $FM_1$ . Perpendicular information is stored in  $FN_1$  and crosswise information is stored in  $FO_1$ . Using inverse DWT, the four modules may be reassembled



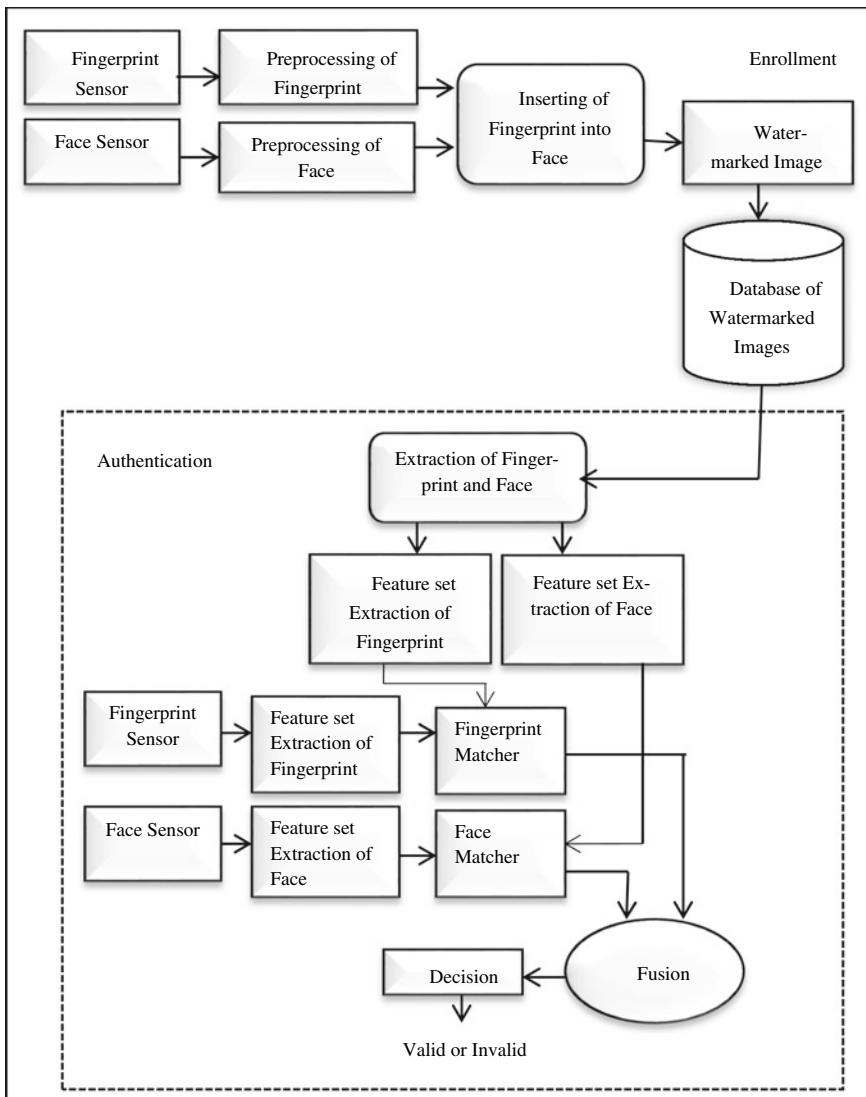
**Fig. 4** Discrete wavelet transform watermarking at level 3

back into the real cover image. The smallest frequency module  $FL_1$  can breakup into another four module, i.e.,  $FL_2$ ,  $FM_2$ ,  $FN_2$ , and  $FO_2$  to represent DWT at level 2. To depict DWT at level3, sub-module  $FL_2$  can again be break up into another four sub-levels, i.e.,  $FL_3$ ,  $FM_3$ ,  $FN_3$ , and  $FO_3$ . This can be expanded up to n-levels. Figure 4 shows DWT watermarking at level 3. The majority of estimation data is kept in the short frequency sub-band, i.e.,  $FL_1$  and another three high-level frequency modules, i.e.,  $FM_1$ ,  $FN_1$ , and  $FO_1$  each includes elaborate image data such as edges, textures, and so on. As a result, the watermark is embedded in the  $FL_1$  components, which is higher security and produces better outcomes than high-level frequency components. The preprocessed images of fingerprint and face are used as watermark and cover

images in the suggested method. The following are the procedure for watermark embedding and extraction in the suggested method:

### Watermark Embedding Process

- (a) Divide the cover face image into four components  $FL_1, FM_1, FN_1, FO_1$  using first-level DWT.



**Fig. 5** Architecture of proposed approach

- (b) )Apply second-level DWT to the  $FL_1$  component to create the  $FL_2$ ,  $FM_2$ ,  $FN_2$ , and  $FO_2$  components.
- (c) )Apply third-level DWT to the  $FL_2$  component to create the  $FL_3$ ,  $FM_3$ ,  $FN_3$ , and  $FO_3$  components.
- (d) )Divide the watermark, fingerprint image into four components using first level DWT:  $WFL_1$ ,  $WFM_1$ ,  $WFN_1$ , and  $WFO_1$ .
- (e) )Apply second-level DWT to the  $WFL_1$  components to create the  $WFL_2$ ,  $WFM_2$ ,  $WFN_2$ , and  $WFO_2$  components.
- (f) )Apply third-level DWT to the  $WFL_1$  components to create the  $WFL_3$ ,  $WFM_3$ ,  $WFN_3$ , and  $WFO_1$  components.
- (g) )Apply the following equation to embed the watermark fingerprint image into the cover face image:

$$W_{wm}FL_3 = FL_3 + b * WFL_3 \quad (1)$$

where  $W_{wm}FL_3$  is a small frequency component of the watermarked image,  $FL_3$  is a small frequency component of the face cover image,  $WFL_3$  is a small frequency component of the watermark fingerprint image, and scaling factor is  $b$ .

- (h) To get the  $FL_2$  components, use inverse DWT at third-level decomposition utilizing the four components  $W_{wm}FL_3$ ,  $FM_3$ ,  $FN_3$ , and  $FO_3$ .
- (i) To get the  $FL_1$  components, use inverse DWT at second-level decomposition utilizing the four components  $FL_2$ ,  $FM_2$ ,  $FN_2$ , and  $FO_2$ .
- (J) To create the watermarked image, use inverse DWT at first-level decomposition utilizing the four components  $FL_1$ ,  $FM_1$ ,  $FN_1$ , and  $FO_1$ .

### Watermark Extraction Process

- (a) Use first-level DWT to separate the cover face image into four components  $FL_1$ ,  $FM_1$ ,  $FN_1$ , and  $FO_1$ .
- (b) Apply second-level DWT to the  $FL_1$  component to create  $FL_2$ ,  $FM_2$ ,  $FN_2$ , and  $FO_2$ .
- (c) Apply third-level DWT to the  $FL_2$  component to create  $FL_3$ ,  $FM_3$ ,  $FN_3$ , and  $FO_3$ .
- (d) Use first-level DWT to separate the watermarked image into four components  $W_{wm}FL_1$ ,  $W_{wm}FM_1$ ,  $W_{wm}FN_1$ , and  $W_{wm}FO_1$ .
- (e) Apply second-level DWT to the  $W_{wm}FL_1$  component to create the  $W_{wm}FL_2$ ,  $W_{wm}FM_2$ ,  $W_{wm}FN_2$ , and  $W_{wm}FO_2$  components.
- (f) Apply third-level DWT to the  $W_{wm}FL_2$  component to create the  $W_{wm}FL_3$ ,  $W_{wm}FM_3$ ,  $W_{wm}FN_3$ , and  $W_{wm}FO_3$  components.
- (g) Using the equation presented below, extract the watermark fingerprint from the watermarked image:

$$E_{tr}W_{wm} = W_{wm}FL_3 - FL_3 \quad (2)$$

where  $E_{tr}W_{wm}$  is a small frequency component of the retrieved watermark,  $W_{wm}FL_3$  is a small frequency component of the watermarked image, and  $FL_3$  is a small frequency component of the cover face image.

- (h) To create the retrieved watermark image of a fingerprint, use reverse DWT at every three levels.
- (i) To create the watermark retrieved cover image of a face, use reverse DWT at every three levels.

### Fusion at Match Score Level

Fusion is the process of combining match scores generated by various biometric matches at the match score level [17]. Let  $MS_{Fingerprint}$  and  $MS_{Face}$  be the fingerprint and face match scores, respectively. Fingerprint and face match scores are not equal because they are not in the same number range. Before combining the scores, normalization is done to convert them into a common domain. The match scores might be converted into a common range [0, 1] by using min–max normalization [18]. The following are the normalized scores provided by the min–max Eq. (3):

$$\begin{aligned} NS_{Fingerprint} &= \frac{MS_{Fingerprint} - Min_{Fingerprint}}{Max_{Fingerprint} - Min_{Fingerprint}}, \\ NS_{Facial} &= \frac{MS_{Facial} - Min_{Facial}}{Max_{Facial} - Min_{Facial}} \end{aligned} \quad (3)$$

where  $[Min_{Fingerprint}, Max_{Fingerprint}]$  are the smallest and greatest fingerprint biometric scores,  $[Min_{Facial}, Max_{Facial}]$  are the smallest and highest face biometric scores, and  $NS_{Fingerprint}$  and  $NS_{Facial}$  are the fingerprint and face biometrics normalized match scores, respectively.

The final match score ( $MS_{Final}$ ) is calculated by combining the normalized fingerprint and face values using the simple sum [18] formula, as shown in Eq. (4):

$$MS_{Final} = NS_{Fingerprint} + NS_{Facial} \quad (4)$$

The decision module receives the final match score  $MS_{Final}$ , which it compares to a predetermined threshold value to determine whether the individual is real or not.

## 4 Experimental Results and Discussion

Some experimental results are shown in this section to demonstrate the robustness and imperceptibility of the proposed watermarking technique. Fingerprint and face data were taken from the test fingerprint image dataset [19], and CVL face images database [20] to apply the suggested method. The image of a face was utilized as the cover image, while the image of a finger was used as a watermark to be inserted in the

cover image. The approach was done in MATLAB and both images were obtained at the same size.

### Mean Square Error (MSE)

It is calculated between cover image  $X(u, v)$  and watermarked image  $X'(u, v)$  using the Eq. (5) given below where  $S$  is an image size. Its value should be kept to a minimum.

$$\text{MSE} = \frac{1}{S} \sum_{u,v} [X(u, v) - X'(u, v)]^2 \quad (5)$$

### Peak Signal to Noise Ratio (PSNR)

It is determined by subtracting the cover image from the watermarked image. The image quality will improve when the PSNR value rises, and vice versa. PSNR is measured in decibels (dB) and should be greater than 30. The following Eq. (6) is used to calculate it:

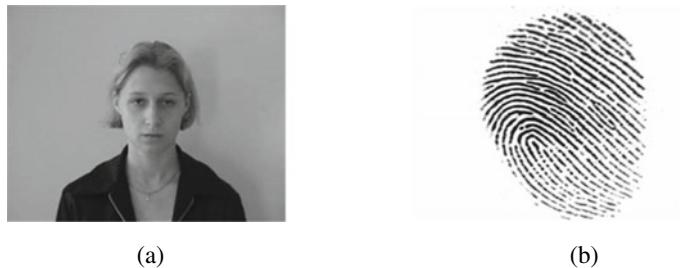
$$\text{PSNR} = \frac{10 * \log_{10}(255 * 255)}{\text{MSE}} \quad (6)$$

### Correlation coefficient (CRC)

It calculates the compatibility of the real watermark  $X(u, v)$  and the extracted watermark  $X'(u, v)$ . It has a value of 0 to 1. It is calculated using the following Eq. (7):

$$\text{CRC} = \frac{\sum_u \sum_v X(u, v) X'(u, v)^2}{\sqrt{\sum_u \sum_v X(u, v)^2} \times \sqrt{\sum_u \sum_v X'(u, v)^2}} \quad (7)$$

Figure 6 exhibits preprocessed images of a face (cover image) and a fingerprint (watermark). At level 3, DWT watermarking was implemented. The watermark is inserted into the cover image using watermark embedding methodology with various scaling factors to obtain watermarked images. The watermark extraction methodology is applied upon watermarked images to extract watermarks. Table 1 shows different face images and watermarked images along with PSNR values. According to this table, the value of PSNR is extremely greater than 30db. The results reveal that the watermarked image is perceptually identical to the original, and that the extracted characteristics are also identical when compared to the watermarked image. For various scaling factor values, PSNR, MSE, and CRC are computed. Watermark embedding and extraction values are shown in Tables 2 and 3. The best output is obtained when scaling factor  $a = 0.007$ . This can be seen in three tables. As a result, both fingerprint and face images are protected using the proposed method.



**Fig. 6** Preprocessed **a** cover image and **b** watermark image

**Table 1** Comparison of watermarked face images with original face images using PSNR value

Scaling factor	Cover Image	Watermark	Water-marked Image	Extracted watermark	PSNR
0.002					120.9802
0.007					147.9120
0.02					104.2643
0.04					70.8906

## 5 Conclusion

A watermarking-based approach to the protection of templates in a multimodal biometric system is suggested in this research. The efficacy of the suggested approach has been demonstrated using face and fingerprint biometric features. Rather than keeping real templates in the database, a watermarked image created by integrating a preprocessed fingerprint image into a preprocessed facial image is preserved in the

**Table 2** Comparison of MSE, PSNR, and CRC values for watermark embedding with DWT watermarking at level 3

Scaling factor	MSE	PSNR	CRC
0.002	04.1030e <sup>-7</sup>	147.9087	1.0000
0.004	10.1120e <sup>-5</sup>	118.1023	1.0000
0.007	03.0023e <sup>-5</sup>	106.9120	1.0000
0.020	03.9820e <sup>-4</sup>	104.2643	0.9987
0.040	08.7670e <sup>-4</sup>	070.8906	0.9860
0.200	00.0170	055.9732	0.9709
0.600	00.1256	026.1023	0.6890
0.800	00.3856	014.8570	0.1452

**Table 3** Comparison of MSE, PSNR, and CRC values for watermark extraction with DWT watermarking at level 3

Scaling factor	MSE	PSNR	CRC
0.002	24.8670	09.7934	1
0.004	24.6927	09.6001	1
0.007	24.4743	09.5989	1
0.020	24.3278	09.9034	1
0.040	21.5975	10.2780	1
0.200	21.9897	12.8067	1
0.600	05.9732	22.8709	1
0.800	00.3786	54.6513	1

database. By concealing fingerprint images into face images, this method ensures that both images are secure. It ensures that watermarked images saved in the database cannot be used to create real biometric templates. The ultimate authentication decision is made based on the fusion of fingerprint and facial matching scores at the match score level. The results reveal that the watermarked image is perceptually identical to the original, and that the extracted characteristics are also identical when compared to the watermarked image. Hence, the suggested method provides security, privacy, and improved recognition accuracy.

## References

1. Vijayakumar, T.: Synthesis of palm print in feature fusion techniques for multimodal biometric recognition system online signature. *J. Innov. Image Process. (JIIP)* **3**, 131–143 (2021). <https://doi.org/10.36548/jiip.2021.2.005>
2. Chin, S.W., Ang, L.-M., Seng, K.P.: A new multimodal biometric system using tripled chaotic watermarking approach. *1*, 1–8 (2008). <https://doi.org/10.1109/itsim.0008.4631557>
3. Ratha, N.K., Connell, H., Bolle, R.M.: Enhancing security and privacy in biometrics-based authentication systems. *IBM Syst. J.* **40**, 21 (2001)
4. Ratha, N.K., Connell, J.H., Bolle, R.M.: An analysis of minutiae matching strength. In: Bigun, J., Smeraldi, F. (eds.) *Audio- and Video-Based Biometric Person Authentication*. AVBPA 2001.

- Lecture Notes in Computer Science, vol. 2091. Springer, Berlin, Heidelberg (2001). [https://doi.org/10.1007/3-540-45344-X\\_32](https://doi.org/10.1007/3-540-45344-X_32)
- 5. Zebbiche, K., Ghouti, L., Khelifi, F., Bouridane, A.: Protecting fingerprint data using watermarking. **6** (2006)
  - 6. Kant, C., Nath, R., Chaudhary, S.: Biometrics security using steganography. **5** (2008)
  - 7. Abdul, W., Nafea, O., Ghouzali, S.: Combining watermarking and hyper-chaotic map to enhance the security of stored biometric templates. Comput. Intell. **15** (2019)
  - 8. Ghouzali, S., Nafea, O., Wadood, A., Hussain, M.: Cancelable multimodal biometrics based on chaotic maps. Appl. Sci. **11**, 8573 (2021). <https://doi.org/10.3390/app11188573>
  - 9. Gautam, A.K., Kapoor, R.: Multi-modal biometric recognition system based on FLSL fusion method and MDLNN classifier. Turkish Journal of Computer and Mathematics Education (TURCOMAT) **12**(12), 241–256 (2021)
  - 10. Zebbiche, K., Khelifi, F., Bouridane, A.: An efficient watermarking technique for the protection of fingerprint images. EURASIP J. Inform. Secur. **2008**, 918601 (2008). <https://doi.org/10.1155/2008/918601>
  - 11. Bousmina, N., Ghouzali, S., Mikram, M., Abdul, W.: DTCWT-DCT watermarking method for multimodal biometric authentication. In: Proceedings of the 2nd International Conference on Networking, Information Systems & Security, pp. 1–7 (2019). <https://doi.org/10.1145/3320326.3320409>
  - 12. Bousmina, N., Ghouzali, S., Mikram, M., Lafkih, M., Nafea, O., Al-Razgan, M., Abdul, W.: Hybrid multimodal biometric template protection. Intell. Autom. Soft Comput. **35**–51 (2021). <https://doi.org/10.32604/iasc.2021.014694>
  - 13. Leghari, M., Memon, S., Dhomeja, L.D., Jalbani, A.H., Chandio, A.A.: Deep feature fusion of fingerprint and online signature for multimodal biometrics. Computers **10**(2), 21 (2021). <https://doi.org/10.3390/computers10020021>
  - 14. Tripathi, M.: Analysis of convolutional neural network based image classification techniques. J. Innov. Image Process. (JIIP) **3**, 100–117 (2021). <https://doi.org/10.36548/jiip.2021.2.003>
  - 15. Colbry, D., Stockman, G., Jain, A.: Detection of anchor points for 3D face verification. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 8 (2005)
  - 16. Sungheetha, A., Sharma, R.: A novel CapsNet based image reconstruction and regression analysis. J. Innov. Image Process. (JIIP) **2**, 156–164 (2020). <https://doi.org/10.36548/jiip.2020.3.006>
  - 17. Manoharan J.S.: A novel user layer cloud security model based on chaotic arnold transformation using fingerprint biometric traits. J. Innov. Image Process. (JIIP) **3**, 36–51 (2021). <https://doi.org/10.36548/jiip.2021.1.004>
  - 18. Jain, A., Nandakumar, K., Ross, A.: Score normalization in multimodal biometric systems. Pattern Recogn. **16** (2005)
  - 19. BiometricsIdealTest. <http://biometrics.idealtest.org/findDownloadDbByMode.do?mode=Fingerprint>. Last accessed 16 Apr 2021
  - 20. Computer Vision Laboratory. <http://www.lrv.fri.uni-lj.si/facedb.html>. Last accessed 16 Apr 2021

# A Big Data Deep Learning Approach for Credit Card Fraud Detection



Kandasamy Illanko, Raha Soleymanzadeh, and Xavier Fernando

**Abstract** Credit card usage, both online and in-person, has increased tremendously in the past few years leading to more credit card frauds and thefts especially during online transactions. For instance, CPA Canada reports that a 34% of respondents have personally experienced fraud, with 18% reporting that a credit card fraud occurred. As these thefts are often anomalous in nature, artificial intelligence techniques, in particular deep learning and neural networks, can be used to detect these fraudulent transactions. This paper studies convolutional neural network (CNN) to analyze the usage data and to predict the occurrence of fraudulent transactions. In addition, SMOTE, random undersampling, and random oversampling technique are used to reduce the false-positive rate. Performance of the proposed model was evaluated using three different sampling methods, as well as without sampling techniques. The results demonstrate the advantage of using sampling methods.

**Keywords** Big data · Deep learning · Fraud detection · Convolutional neural network (CNN)

## 1 Introduction

Credit card transactions are used almost in every occasions these days, from small to big, from ordering pizza to purchasing major appliances. More of these transactions happen online, without a personal identification in place. The ubiquitous nature of

---

This work is supported by Natural Sciences and Engineering Research Council (NSERC) of Canada.

K. Illanko · R. Soleymanzadeh · X. Fernando (✉)

Department of Electrical, Computer, and Biomedical Engineering, Ryerson University,  
Toronto, Canada

e-mail: [fernando@ryerson.ca](mailto:fernando@ryerson.ca)

K. Illanko

e-mail: [killanko@ryerson.ca](mailto:killanko@ryerson.ca)

R. Soleymanzadeh

e-mail: [rsoleymanzadeh@ryerson.ca](mailto:rsoleymanzadeh@ryerson.ca)

these usages leads to number malicious occurrences. Fraudsters have been increasingly used more sophisticated methods to mask fraudulent transactions. Hence, financial institutions invest on advanced techniques to identify fraudulent transactions that typically have a different pattern. Machine learning, in particular, deep learning is profoundly useful here to identify different patterns.

Deep learning neural networks can be broadly classified into convolution neural network (CNN) and its variants, or recurrent neural network (RNN) such as a long short-term memory network (LSTM). The CNN was originally developed for image classification. However, the underlying convolutional filters can also be used on vector data. An LSTM filter, unlike a simple RNN, ‘forgets’ unnecessary data, or not use some of the past data when that data does not contribute to identifying a pattern.

For example, a fraud detection system developed by Zhang et al. [1] for a big bank in China employs deep learning enhanced with an advanced feature engineering method. They claim that their method could determine more fraudulent transactions than the benchmark techniques with low rate of false positives. Roy et al. [2] have used the LSTM model and discovered that the model performance was hugely impacted by size of the network.

This paper uses a CNN to predict a fraudulent credit card transaction. It uses a public dataset available on Kaggle [3] to verify the performance. A major shortcoming of credit card transaction data is that there are very few fraudulent transactions compared to legitimate transactions. This unbalanced phenomenon results in poor training of the neural network. To overcome this limitation, we have used a number of advanced sampling methods in this work.

Rest of the paper is organized as follows. Section 2 reviews relevant literature. In Sect. 3, data description and preprocessing techniques are described. The CNN is discussed in more detail in Sect. 4. An analysis of the model’s results is also shown at the end of this section. The paper concludes in Sect. 5.

## 2 Related Work

Zhang et al. [1] proposed a so-called feature engineering framework with deep learning. They utilized feature variables to predict the fraudulent behavior of credit card transactions. Then various deep learning methods are applied to model transaction behaviors more effectively and provided better detection of fraud. Results from their experiments demonstrate that their proposed methodology is a useful and feasible credit card fraud detection tool.

A support vector machine (SVM), K-nearest neighbor (KNN), artificial neural networks (ANN), and a simple feedforward neural network (FFNN) are investigated in [4]. In order to address the imbalanced dataset problem, the study used an under-sampling and normalization method prior to training the dataset. ReLU is used as the activation function for ANN, and 15 hidden layers were used. ANN was found to be the optimal tool for detecting fraud transactions compared to the other two used machine learning algorithms.

The performance of SVM, ANN, Bayesian networks, hidden Markov model, KNN, fuzzy logic systems, decision trees, and logistic regression techniques in detecting credit card fraud was the focus of a work by Yashvi Jain et al. [5]. Fuzzy and logistic regression algorithms were found to give the lowest overall accuracy compared to other algorithms. SVM, decision trees, and KNN were found to offer a medium level of accuracy. Neural networks, Bayesian networks, and fuzzy systems were shown to have a high detection rate. The naive Bayesian network and the ANN were shown to perform better with respect to all the parameters. Logistic regression and fuzzy logic systems were shown to provide accurate results when using raw and unsampled data.

The work in [6] focuses on network intrusion detection using CNN based on the LeNet-5 implementation. It developed a behavior-based classifier learning model in order to improve the classification accuracy of threat detection and reduce its false-positive rate. Clustering analysis was used to categorize the behavioral characteristics of intrusions from different IP addresses gathered from the National Center for High-performance Computing (NCHC) network nodes.

A deep learning model for credit fraud detection based on convolutional long short-term memory (C-LSTM) has been presented in the paper [7]. Performance of the C-LSTM model was assessed on a German credit and Kaggle dataset. An accuracy of 94% was reported.

Fischer and Krauss [8] used LSTM neural network models to predict out-of-sample movements for the S&P 500 from December 1992 until October 2015. In addition, the performance of the algorithm was compared to memory-free algorithms, including random forest, logistic regression, and neural networks. While LSTM performed better than other algorithms during the financial crisis in 2008, the random forest came out on top over LSTM in general.

Strong financial transaction features that can distinguish between the fraudulent and legitimate classes of transactions were identified in [9]. The study also compares CNN with stacked LSTM (SLSTM), along with a hybrid model combining CNN with LSTM (CNN-LSTM), for credit card fraud detection. They found that learning from short-term sequences in the data is a powerful way to train CNNs, while long-term sequences are a useful way to train LSTM.

This dataset comes from an Indonesian bank, and the predominant class of non-fraud values has been subsampled in four different ratios to create four datasets. The PCA method is used to reduce the number of features by representing them on time scales. According to the results, the accuracy of the classifier is increased as the ratio between the non-fraud values and the fraud values is increased. For training accuracy, SLSTM was the best, CNN-LSTM was in second place, and CNN was the last. The AUC results show that CNN performed best followed by CNN-LSTM and SLSTM. This showed that short-term patterns dominate long-term relationships as far as fraudulent transactions were concerned. However, it was noted that due to the imbalance of the datasets, accuracy is not the only measure used for performance validation purposes.

Abhimanyu et al. [2] evaluated different deep learning techniques. In order to deal with the issue of an imbalanced dataset, the undersampling method was applied. The

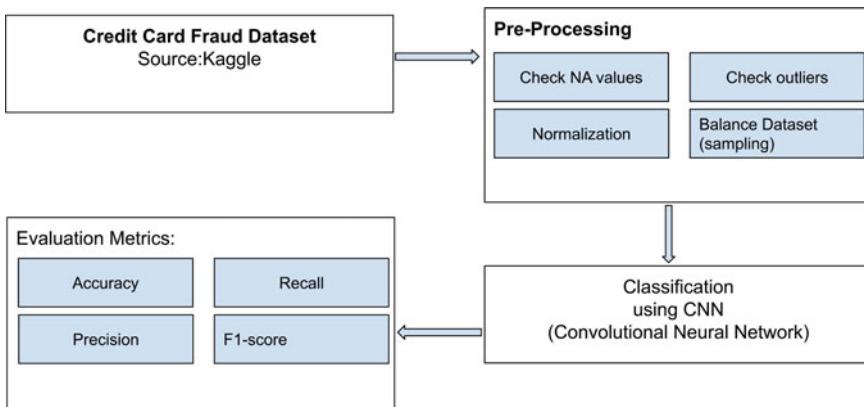
GRU reportedly performed better than all deep learning algorithms, and it was found that network size was the most influential hyperparameter for all algorithms. These authors provide general guidelines on how to evaluate and tune model parameters for credit card fraud monitoring by using neural network skills.

An ensemble model consisting of an LSTM and a GRU was introduced in the paper [10]. A voting mechanism based on ANN was used to make the final decision. To improve the accuracy of the voting decision, rather than using a threshold, the paper presents a new algorithm for training the ANN. Using two real-world datasets, the researchers found the proposed model outperformed the state-of-the-art models in all evaluation criteria. Additionally, the time analysis of the proposed model shows that it is more efficient in terms of real-time performance versus the recent models in the field.

The authors of [11] used the IEEE-CIS fraud detection dataset. Several machine learning and deep learning models as well as a model that uses a bidirectional LSTM and a GRU for advanced classification were developed. These include Naive Base, Voting, Ada Boost, Random Forest, Decision Tree, and Logistic Regression. Techniques for analyzing highly imbalanced datasets, including undersampling, oversampling, and SMOTE were also explored. Model performance is rated using a set of evaluation metrics.

### 3 Methodology

The method used in this paper detects credit card fraud by predicting past transactions based on the knowledge of the ones that were fraudulent. The main deep learning architecture used in this paper is the convolutional neural network (CNN). The work is implemented in Python. The structure of the methodology is presented in Fig. 1.



**Fig. 1** Methodology process

### 3.1 *Data Description*

The credit card fraud detection dataset from Kaggle [3] is the dataset used in this paper. There were 284,807 transactions made by European cardholders on two days in September 2013 as part of this dataset. As would be the case with financial-fraud datasets, this dataset is highly imbalanced with 492 fraudulent transactions out of 284,807 total transactions (0.172%). No missing values were found in the dataset. The dataset has 31 variables. These variables are obtained after applying the principal component analysis (PCA) on all but the time and the financial amount variables.

### 3.2 *Data Preparation*

To overcome the difficulty with the imbalanced dataset, over and under random re-sampling techniques, and synthetic minority oversampling (SMOTE) were used. Among the credit card data, 60% was used for training, while the other 40% was for testing.

## 4 Modeling and Results

In this section, we introduce the model we employed to compare the performance of the different methods. To begin, we examined the CNN model. We present our evaluation metrics and results in the subsequent sections.

### 4.1 *Convolutional Neural Network(CNN) for Credit Card Fraud Detection*

The second layer of CNN created by CONV1D uses ReLU activation, batch normalization, and a dropout layer. Similar to the first layer, the second layer is constructed by dropping 20% of the neurons after the second layer and 10% of the neurons randomly after the first layer. A flattening layer is then used to convert the data into vectors. The last layer was used the Sigmoid activation function for classification.

### 4.2 *Evaluation Metrics*

For credit card fraud detection systems, capturing all fraudulent transactions and reducing false alarms (legitimate transactions misidentified as fraudulent) is the goal of selecting the metric for evaluation. Because of the high-class imbalance in the

datasets, accuracy is not an appropriate metric to evaluate models [12]. A confusion matrix is used in this study to denote non-fraud (legitimate) instances as negatives and fraud instances as positives. True positives are the fraud cases predicted correctly, true negatives are the non-fraud cases predicted correctly, false positives are the non-fraud cases predicted as fraud, and false-negatives are the fraud cases predicted as nonfraud.

### 4.3 Results

Figure 2 shows the result of the used model on performance metrics such as accuracy, precision, recall, and F1\_score, and the respective plot for training and testing is shown in Fig. 3, using random\_under\_sampling method. Figure 4 and 5 represent the model accuracy and model loss during both training and testing, respectively.

**Test Result with under sampling method:**

=====

Accuracy Score: 94.16%

Classification Report:

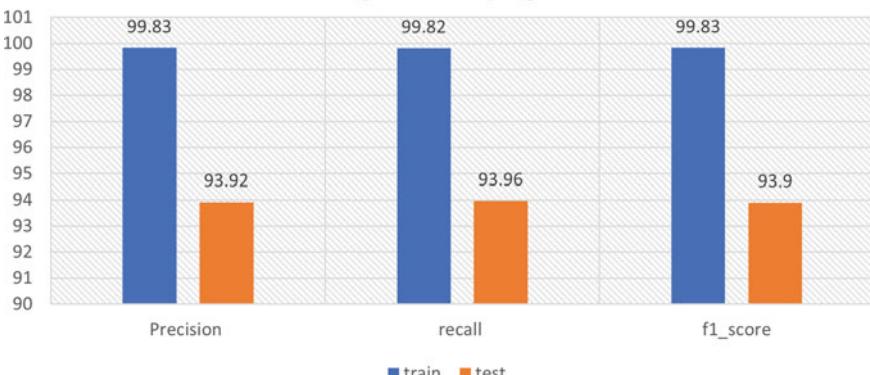
	0	1	accuracy	macro avg	weighted avg
precision	0.903846	0.983871	0.941624	0.943859	0.945077
recall	0.984293	0.901478	0.941624	0.942886	0.941624
f1-score	0.942356	0.940874	0.941624	0.941615	0.941592
support	191.000000	203.000000	0.941624	394.000000	394.000000

Confusion Matrix:

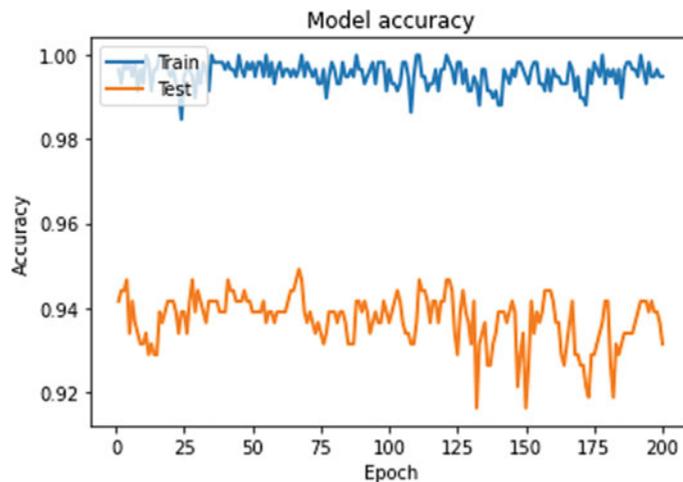
```
[[188  3]
 [ 20 183]]
```

**Fig. 2** Performance metrics and confusion matrix of undersampling method

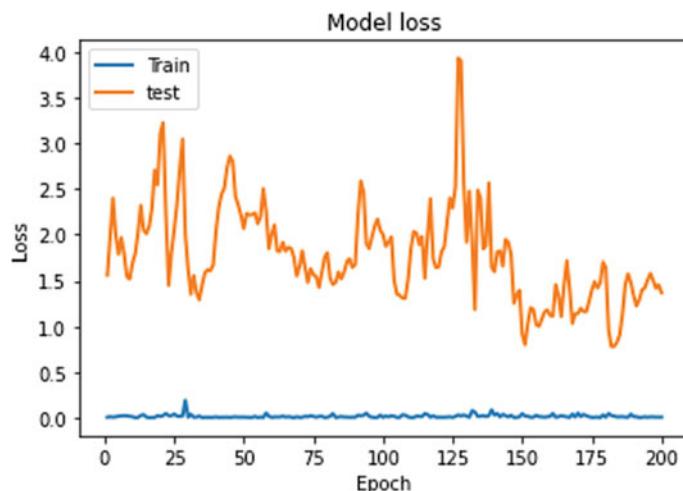
CNN using Under Sampling method



**Fig. 3** Evaluation metrics for CNN using Under\_Sampling Method

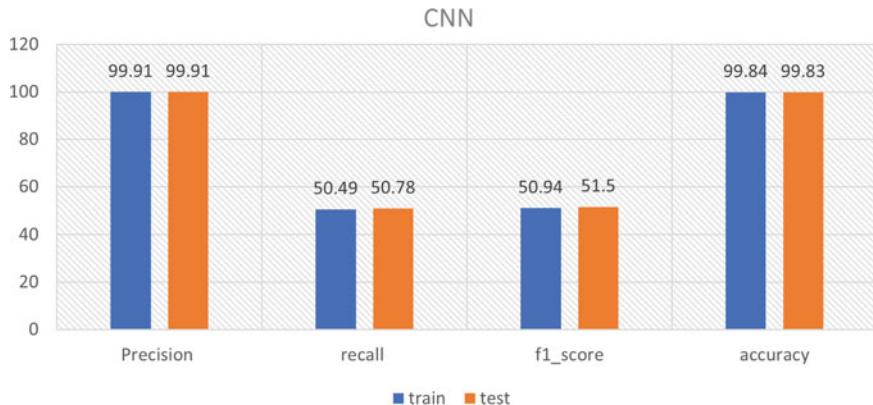


**Fig. 4** Model accuracy



**Fig. 5** Model loss

Regarding the effect of class imbalance on classification performance, we observed the following. Based on the same model, Fig. 6 shows the performance of the CNN without using sampling techniques. Even though the accuracy is high, the precision, recall, and f1-score of the model are low. The CNN model performs better with sampling techniques, and undersampling is the best choice in this environment if training time is an issue since it dramatically reduces the size of the training set. Table 1 is a comparison of the model's performance with different sampling methods.



**Fig. 6** Performance metrics of CNN model without sampling methods

**Table 1** Final results from the proposed algorithms

	Precision (%)	Recall (%)	$f_1$ score (%)
CNN	99.91	50.58	51.5
CNN with undersampling	93.97	93.96	93.9
CNN with oversampling	99.84	99.84	99.84
CNN with SMOTE	99.87	99.87	99.87

## 5 Conclusion

In this paper, convolutional neural networks (CNNs) were used to detect fraudulent credit card transactions. Different sampling methods (SMOTE, random undersampling, and random oversampling) were compared to non-sampling methods. In terms of reducing false-negative and false-positive rates, the oversampling technique performed well. On the other hand, with the undersampling method, the training time was higher. Results showed that the recall and F1-score of the CNN model without the use of sampling methods were 50.58% and 51.5% respectively, while sampling methods (oversampling) were 99.84 and 99.84%. The CNN model performs better when using sampling techniques. In contrast, undersampling is the best option if training time is a concern since it dramatically reduces the size of the training set.

## References

1. Zhang, X., Han, Y., Xu, W., Wang, Q.: Hoba: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture. *Inf. Sci.* **557**, 302–316 (2021)
2. Roy, A., Sun, J., Mahoney, R., Alonzi, L., Adams, S., Beling, P.: Deep learning detecting fraud in credit card transactions. In: 2018 Systems and Information Engineering Design Symposium (SIEDS), pp. 129–134. IEEE (2018)

3. Najadat, H., Altiti, O., Aqouleh, A.A., Younes, M.: Credit card fraud detection based on machine and deep learning. In: 2020 11th International Conference on Information and Communication Systems (ICICS), pp. 204–208. IEEE (2020)
4. Asha, R.B., Suresh Kumar, K.R.: Credit card fraud detection using artificial neural network. *Glob. Trans. Proceed.* **2**(1), 35–41 (2021)
5. Jain, Y., NamrataTiwari, S.D., Jain, S.: A comparative analysis of various credit card fraud detection techniques. *Int. J. Recent Technol. Eng.* **7**(5), 402–407 (2019)
6. Lin, W.-H., Lin, H.-C., Wang, P., Wu, B.-H., Tsai, J.-Y.: Using convolutional neural networks to network intrusion detection for cyber threats. In: 2018 IEEE International Conference on Applied System Invention (ICASI), pp. 1107–1110. IEEE (2018)
7. Arun, G.K., Venkatachalam, K.: Convolutional long short term memory model for credit card detection. In: 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1168–1172. IEEE (2020)
8. Fischer, T., Krauss, C.: Deep learning with long short-term memory networks for financial market predictions. *Eur. J. Oper. Res.* **270**(2), 654–669 (2018)
9. Heryadi, Y., Warnars, H.L.H.S.: Learning temporal representation of transaction amount for fraudulent transaction recognition using CNN, stacked LSTM, and CNN-LSTM. In: 2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom), pp. 84–89. IEEE (2017)
10. Forough, J., Momtazi, S.: Ensemble of deep sequential models for credit card fraud detection. *Appl. Soft Comput.* **99**, 106883 (2021)
11. Najadat, H., Altiti, O., Aqouleh, A.A., Younes, M.: Credit card fraud detection based on machine and deep learning. In: 2020 11th International Conference on Information and Communication Systems (ICICS)
12. Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* **106**, 249–259 (2018)

# A Study of VLC Between Vehicles and Traffic Signal Lights



Jonathan Diller and Xavier Fernando

**Abstract** Visible light communication (VLC) is fast emerging as a viable medium for infrastructure-to-vehicle (I2V) communications. Traffic signal lights equipped with VLC transmitters can effectively communicate with the waiting vehicles to communicate essential information via VLC. However, due to the directional properties of light propagation and the height differences between the vehicle and traffic signal, there are stringent limitations on a feasible range of communications. In this paper, we study this problem using a geometrical approach and proposed for a typical traffic light the optimized attack angle for detecting sensors is  $\approx 16^\circ$  and with source emission profile of  $m = 35$ . New methods for improving performance are also discussed by segmenting the transmitter to create an array of axially misaligned sources focused at different spatial regions.

**Keywords** VLC · Optical wireless · V2I · Field of view · Photodetector

## 1 Introduction

Today in 2021, automobile safety and reducing incidents remain as paramount concern for all modern societies. Vehicle-related fatalities have been identified by the World Health Organization (WHO) as the 8th leading cause of death in modern society [1]. Vehicle light communication (VLC) has been established as a good method for improving overall automotive safety by facilitating a direct communication means for sharing instructions between different automobiles approaching an intersection.

---

Supported by Ryerson University.

---

J. Diller (✉) · X. Fernando  
Ryerson University, Toronto, ON M5B 2K3, Canada  
e-mail: [jonathan.diller@ryerson.ca](mailto:jonathan.diller@ryerson.ca)

X. Fernando  
e-mail: [fernando@ryerson.ca](mailto:fernando@ryerson.ca)

There remain practical problems associated with VLC for I2V applications, such as high noise due to ambient light, adapting for different weather conditions, and finding a receiving mechanism that can be adopted widely by different automobile manufacturers [2].

It remains to be seen if applications like traffic lights can be pragmatically improved by the use of VLC. There is a great opportunity for this method to excel in the field of automobiles due to the ability for the system to exist with very little infrastructure and integrate into existing systems using light. Although currently there may be practical problems with this technology such as lower signal-to-noise ratio (SNR) and limited sensing range, there is still a large benefit in safety and overall traffic efficiency if this can be integrated large scale.

If solutions to some of the challenges discussed below can be addressed, this technology stands to massively increase the data exchange and collaboration between vehicles on the road. Drivers would be still free to make autonomous decisions; however, the impact and knowledge of those decisions could be immediately disseminated throughout the motorist network.

## 2 Objective

The initial objective of this paper is comparing different techniques for integrating VLC systems into existing traffic signals and is comparing the relative strengths & weaknesses of each approach. After comparisons of existing methods, further ideas of receiver and transmitter design are discussed to enhance existing methods.

As a means to review and compare the different techniques, we have established metrics such as SNR, range, and the ability for these systems to be integrated. Improving traffic lights is significant from the standpoint that many cars and areas stand to be positively affected; however, it could be a massive infrastructure integration commitment if a cost-effective solution is not found.

In order to realize the potential of I2V communications, it is important to determine the bottlenecks keeping this approach from being integrated successfully on a large scale. Unfortunately, the traffic management infrastructure is a deeply entrenched system that would take a considerable effort to retrofit to adapt to novel VLC systems. Although that gain is a large improvement with respect to the potential improvements in safety, it would have to be clear in advance that a large improvement could expect to be seen by utilizing a system like VLC. Systems like this would also have to adapt in order to be compliant to the radically changing ability of automobiles to make decisions autonomously.

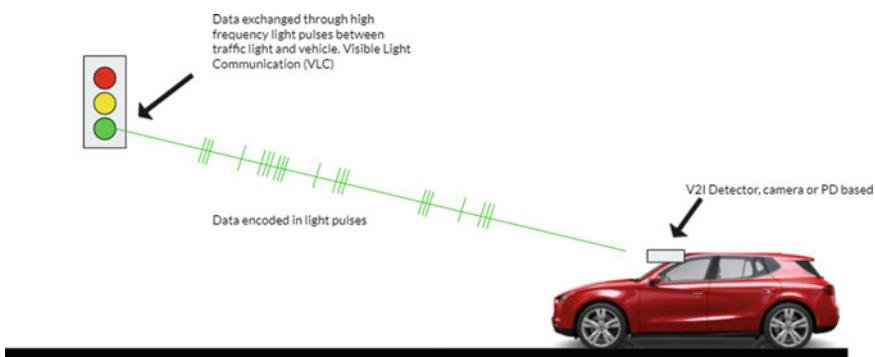
### 3 Review of Literature

#### 3.1 General Review

A broad literature review indicates that there are two widely considered methods for facilitating VLC from the receiving side. The first is using photodiode-based sensors and the second is adapting high-speed cameras. One of the major allures of using this technology is the low cost for integration based on taking advantage of the existing traffic light infrastructure. This limits major innovations to the development of receiving sensors as the large-scale modification of transmitters (traffic lights, street lights, illuminated road signs) would be required costly retrofits to have significant changes in a large scale. A typical communication link between traffic lights and a vehicle is given in Fig. 1.

These two main approaches both have their own advantages that ultimately revolve around maximizing the performance of key figures of merit required for creating a good platform for communication. Before discussing the key differences between the two approaches, it is important to understand what is important in establishing a good receiver and achieving a high-performance communication network. Taken from [2], the major performance metrics are as follows:

1. Robustness to Noise
2. Communication Range
3. Increasing Data Rate
4. Performing Distance Measurements and visible light positioning
5. Enhancing Mobility
6. Developing Parallel VLC
7. Developing heterogeneous short-range communication and VLC networks



**Fig. 1** Typical example of I2V and V2I communications using VLC

This list encompasses a comprehensive set of items to improve upon; however, the first three items appear to be most significant as they are related to fundamental communication issues [1–3].

The first receiving paradigm, photodetectors, relies on establishing a communication link between the transmitter and receiver by selectively removing light from most of other interfering light sources and decoding only the data streaming from desired light transmitters. In order to achieve this, photodetectors are appropriately shrouded such that only light normal to the sensing element is permitted. With shrouding or by creating enclosures the field of view, FOV, of the receiver, that the detectors can be effective over, is optimally controlled [4]. The fundamental trade-off is being that since these sensors are nominally fixed in position within a vehicle narrowing the FOV will eliminate erroneous and unwanted sources of light; however, this also lowers the possibility the sensing element will be correctly directed at the detecting light from the desired transmitter.

The second receiving paradigm being a high-speed camera. High-speed cameras can successfully decode light pulses from traffic lights owing to the fact they are simply sampling static images. In this setup, there is less discrimination from other outside light sources as most incoming light is detected. However, the amount of information that can be conveyed is limited due to the refreshing rate of the camera being intimately linked with the data rate achievable from the transmitted infrastructure. Another issue is blinding the camera due to direct exposure to high-intensity light. These problems can also be addressed by limiting the FOV and ensuring a sufficiently high frame rate of the camera is used. However, with this approach the cost and complexity increases significantly [2].

Other ideas have been established in the literature to further enhance this technology. Once a communication approach is established, information from an infrastructure component such as a traffic light can be relayed between vehicles using both the taillights and headlights of consecutive cars [4, 5]. If the lead car approaching an intersection can establish a sufficient communication link with the infrastructure, that message can be disseminated between trailing vehicles. This helps to remove some of the issues with communication over long distances by leveraging each vehicle as an intermediate signal conditioner and repeater.

Overall, comparing these two approaches, there is a beneficial simplicity in the photodetector approach. It is eliminating the need for additional complexity in the system and promotes the availability to integrate this system to a large number of users. With a high-speed camera-based approach, there is a richness to having the additional information in terms of image processing and sheer ability to increase the overall signal power by attenuating signal outside the FOV.

In conclusion, we may state that the issues are, for photodetectors, increasing the complexity to improve on the major figures of merit while for camera's finding ways to improve performance while lowering cost and ability to be implemented en masse.

### 3.2 Novel Detector Modalities

Below are two novel solutions that have been developed to improve the performance of the photodiode receiving method [1, 3]. These were chosen on the basis that they revolve around novel circuitry and ways of improving the overall signal-to-noise ratios, SNR, of the receiver. Improving this metric stands to greatly improve this technology as seen above the major limiting factor in this technology is range and noise.

There have been many novel approaches to improve photodiode receivers; however, these were chosen on the basis of scalability and complexity, two major hurdles for VLC/V2I.

There are additional creative technologies to improve the performance of these systems such as adding directionality to the FOV of a photodetector by coupling the system with a low-cost camera and mechanical tracking platform to target the overall sensor position using the processed camera image to direct the sensitive photodetective element at the target of interest. There are also good proposals to couple this technology with radio frequency, RF, communication to selectively communicate signals that do not have line of site between transmitting and receiving elements. This hybrid approach is again a good solution for pairing technologies that both thrive in the weakness of the other.

For the two novels approached, [1, 3], both use adaptive gain mechanisms to electronically boost the incoming signal incident on the detector to improve the signal strength. For [1], this adaptive gain control, AGC, is done for the entire sensing area and is meant to increase gain when environmentally there is low light detected, such as when the transmitting signal is far away or not ideally aligned with the target receiver, although this gain mechanism does not discriminate against noise signals it can combat over saturating the sensing element from being washed out when large amounts of sunlight are present in the background while improving the working range of the sensor.

Method [3] relies on a more complex mechanism in analyzing light illuminated on an array of sensors in order to determine where the signal light is mostly present and actively increasing gain in these sensing elements selectively. This additional discrimination against unwanted light sources helps to greatly improve the signal-to-noise ratio for a given input source strength. For transmitter powers of 10 and 100 W, they were able to boost SNR by a factor of  $10^4$  and  $1.5 * 10^4$ , respectively.

These approaches are highly desirable in that novel approaches like this do not stand to significantly increase implementation costs or complicate the form factor of these devices but they greatly improve the performance. With V2I having many pragmatic challenges of needing to function at a high level in a highly varying outdoor environment solutions like this make the technology ever more possible. These approaches are also not mutually exclusive, and the benefits they leverage can be compounded by adding additional aspects such as a physically tracking FOV and local RF communication for transmitting omnidirectionally.

Technologies like this are somewhat competing with existing automotive OEM technology for enhanced driving safety such as Tesla's *Full Self Driving* or GM's *Supercruise*; however, they stand on their own as being able to massively integrate the coherence of traffic and cooperation between vehicles and traffic. For this reason, they stand alone as a means to progress as a technology alongside automakers individual safety improvement systems.

## 4 Proposed Geometrical Analysis

This study is done to determine more low-cost advances that can be made to improve overall device performance. A simple and effective solution used currently being an optical collection device. This is a simple shield that limits the acceptance angle of incoming light incident on the sensor. These typically are used in conjunction with a lens to collect the light not attenuated by the device and focus it onto the detector. The acceptance angle of light permitted to the sensor is typically referred to as the field of view, (FOV).

The trade-off with using a device like this is that a low FOV removes more of the unwanted noise light; however, it also reduces the effective range over which a moving detector is axially aligned with the transmitter. Refer to Fig. 2.

From the figure above, it can be seen that having a low sensor angle increases the  $d_{\min}$  value; however, when the transmitter is close to the detector, such as when a car is stopped at a red light, a shallow sensor angle may be too low in order to view the transmitter (Figs. 3, 4, 5, 6).

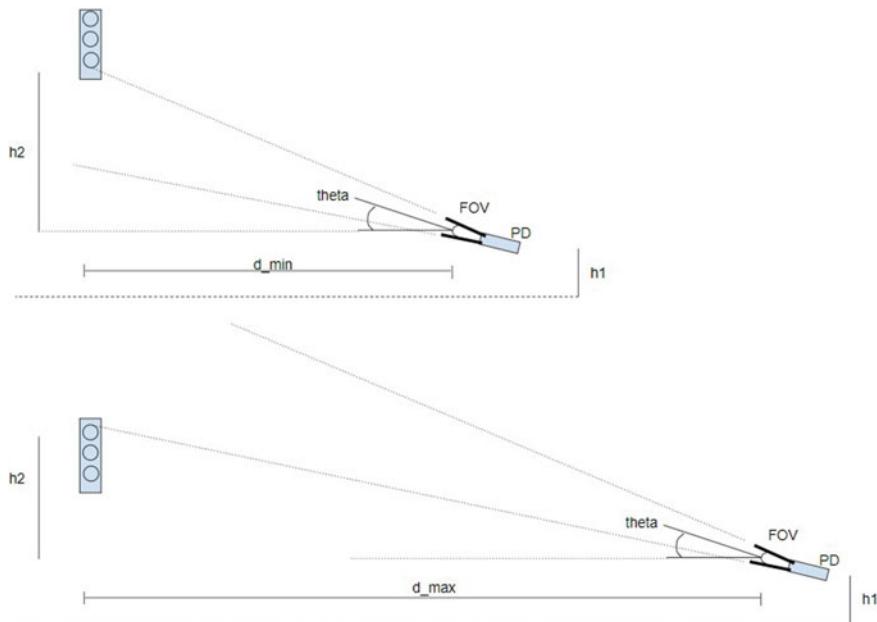
From the figure, the following relationships were obtained solving for distance value range owing to the set of angle values,  $\theta$  and *FOV*.

$$d_{\min} = \frac{h_2 - h_1}{\tan(\theta + \frac{\text{FOV}}{2})}$$

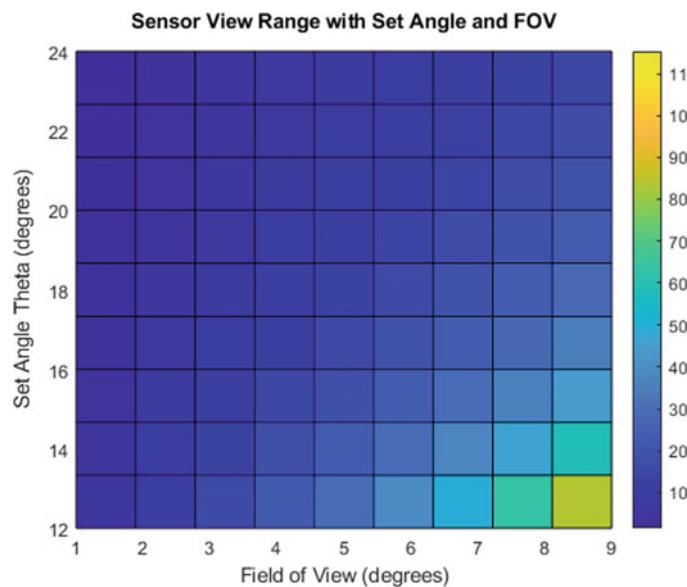
$$d_{\max} = \frac{h_2 - h_1}{\tan(\theta - \frac{\text{FOV}}{2})}$$

The following plot showing the maximum total range of the sensor, ( $d_{\max} - d_{\min}$ ), is given below.

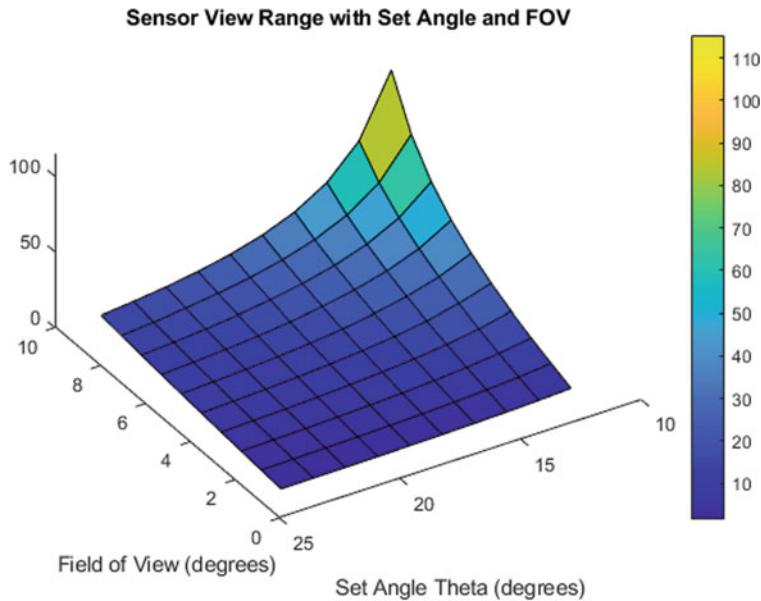
Another consideration with sensor angle is the significance of the  $d_{\min}$  value. In Fig. 2,  $d_{\min}$ , shows the limit of how close a car can be to the traffic light (transmitter). Having a small FOV and low sensor angle will clip off the transmitter light from above meaning that the car would have to be further back and a larger minimum distance to receive the signal. If the FOV is larger, it increases the upper bound on light reaching the transmitter. For typical traffic lights the height of the transmitter,



**Fig. 2** Layout of PD receiver. Theta is the sensor angle of attack with respect to the horizon, FOV is the field of view,  $h_1$  and  $h_2$  are the heights of the receiver and transmitter, respectively

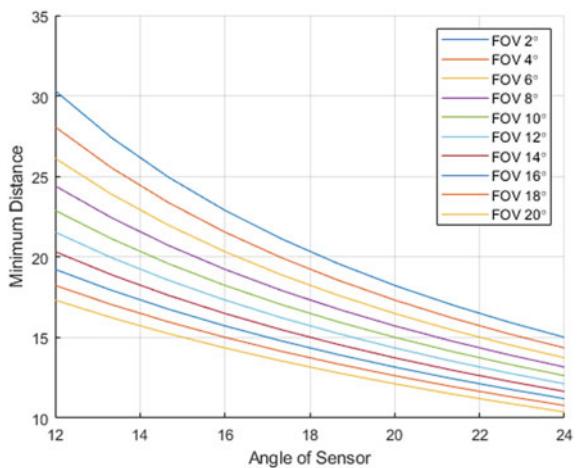


**Fig. 3** FOV versus sensor angle. Color bar values are in meters



**Fig. 4** 3D FOV versus sensor angle. Color bar values are in meters

**Fig. 5** Minimum sensor viewing distance versus sensor angle and FOV

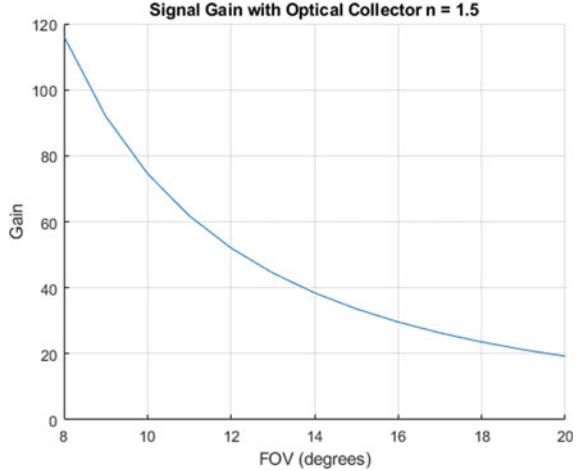


$h$ , and  $d_{\min}$  to be 8 m and 15 m, respectively. The plot below gives the minimum detectable distances visible from a sensor based on the mounting angle,  $\theta$  and  $FOV$ .

From the figure above in order to view at a distance of 15 m, the minimum angle is  $\approx 15^\circ$  and using a FOV of 20.

It can be seen from above that increasing the FOV will decrease the minimum distances and overall effective range of distance over which the sensor is effective;

**Fig. 6** Sensor gain versus FOV



however, there is a problem with arbitrarily increasing the FOV. As discussed above, increasing the FOV allows for unwanted light and increased noise. From [2], a gain equation is given to show the relationship between angle,  $\psi$  and gain achieved,  $g(\psi)$ . In the equation,  $n$  is the refractive index of the collecting lens, taken as 1.5.

$$g(\psi) = \frac{n^2}{\sin^2 FOV}, 0 \leq \psi \leq FOV$$

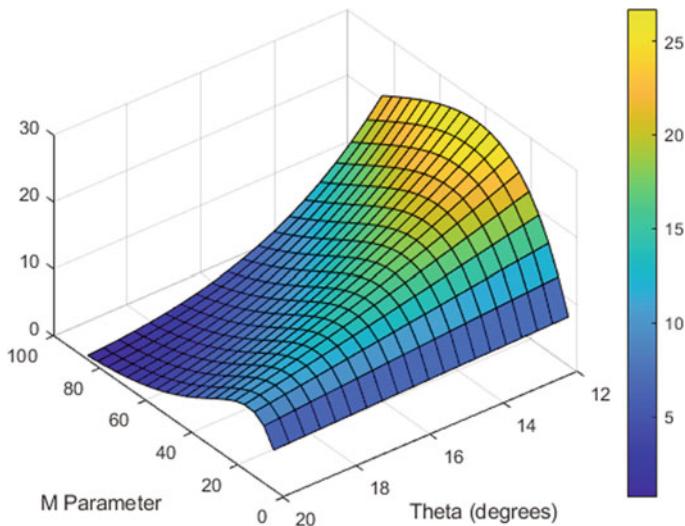
From this figure, it can be seen that keeping a low FOV is advantageous for removing unwanted light and improving SNR.

In addition to the trade-off for optical collection devices, there is a trade-off for the optical transmitter, the traffic light. Assuming the traffic light to be a Lambertian emitter the radiant intensity,  $R_0$  is given below as a function of viewing angle with respect to the emission axis,  $\phi$  and beam profile,  $m$  [2].

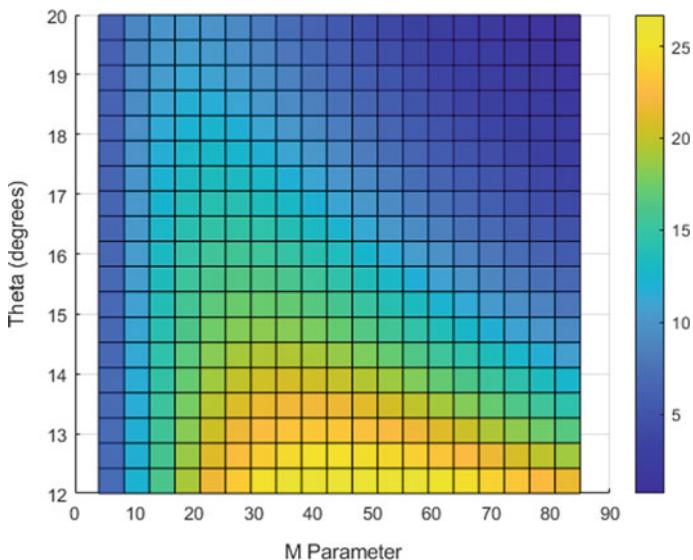
$$R_0(\phi) = \left( \frac{m+1}{2\pi} \right) \cos^m \phi$$

From the equation above, having a high beam parameter,  $m$  ensures a high radiant power for when the viewing angle is close to 0; however, as the misalignment increases, there is a higher drop-off in power. Also, having a low viewing angle decreases the total effective minimum range of the PD and thus usability. Again this problem can be reduced by increasing the FOV of the receiver. However, then the signal gain decreases.

Using the equation above and plotting for varying  $m$  parameter and  $\theta$ , the received intensity is plotted in Figs. 7 and 8, knowing the minimum alimenter angle of the PD should be around  $16^\circ$  and optimal  $m$  parameter of  $\approx 35$



**Fig. 7** Sensor radiant intensity versus *M* parameter and viewing angle



**Fig. 8** 3D sensor radiant intensity versus *M* parameter and viewing angle

#### 4.1 Novel Potential Improvements

With this, there is a potential low-cost benefit to the transmitter by segmenting the output emission angle of the transmitter. If the transmitter light was broken into several bands, each band having a higher  $m$  parameter; however, each having a unique emission angle the radiant intensity could be improved over a greater region. Refer to Fig. 9.

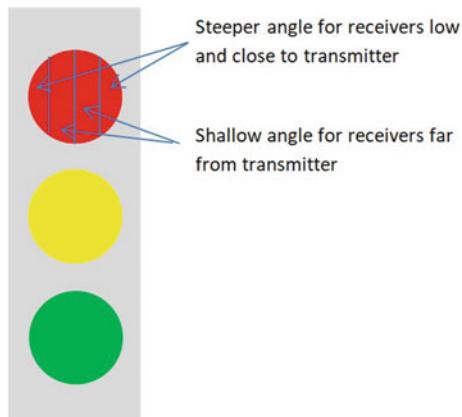
The lower area region at the edges of the transmitter could be directed downward with higher beam parameter, and the center bands could also have a low beam divergence with shallow angle intended for receivers further away. This would help to decouple the trade-off between low FOV/high gain and larger sensor effective range. Each segment of the transmitter could have a unique attack angle to spatially target an area with a high beam parameter.

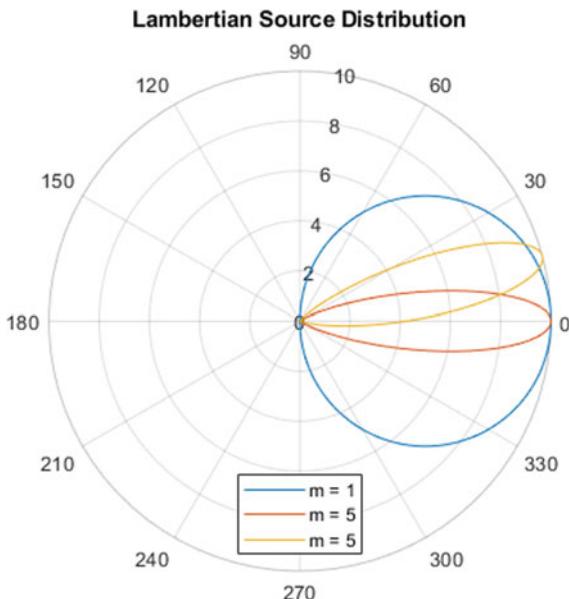
Figure 10 shows this concept with the comparison of two focused and axially misaligned sources. With this, the region approaching an intersection could be effectively be spatially tuned for optical gain by providing sources with higher beam parameter to specific regions.

This idea can also be extended to having an array of detectors on the vehicle to increase the likelihood that the transmitter and receiver are axially aligned, at the same time shrouded with a narrow FOV to maximize the signal gain (Fig. 11).

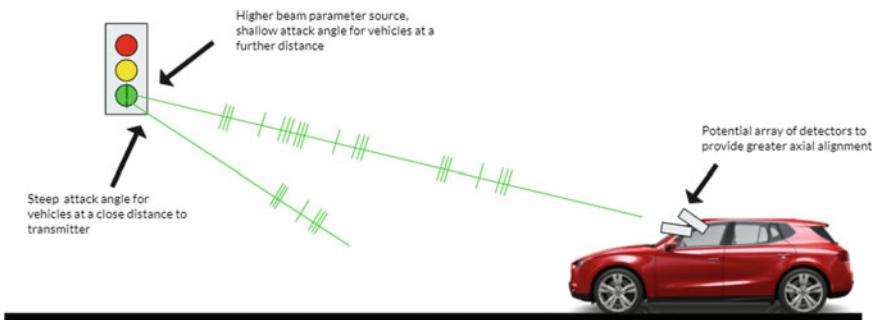
Although this implies increased retrofit costs for large-scale integration, if these types of changes could be made at low cost, they could help improve this technology even further.

**Fig. 9** Traffic light arrangement





**Fig. 10** Comparison of emission profiles with varying beam parameter  $m$ . Two focused profiles, yellow and red curves, have greater emission intensity for angles less than  $15^\circ$



**Fig. 11** Example of array transmitter and receiver for maximized beam parameter and gain at varying vehicle approach distances

## 5 Conclusion

In this paper, we took a geometrical approach considering various transmitter and receiver parameters and studied the performance limitations of VLC link between a traffic light and vehicle. Currently, there are still issues that keep these types of sensors from being integrated large scale. However, a closer look has shown how to improve performance while keeping a high potential for large-scale integration. Keeping this trend seems to only benefit the overall use of this technology and improve the safety for everyone.

Clearly, there exists a trade-off in the receiving range or transmitter and signal-to-noise ratio. Further, trade-offs exist in attempting to shield ambient light sources and restricting over what region transmitters and receivers can create an optical link. By subdividing and splitting the transmitter into multiple sources, performance can be improved by optimizing the transmission profile for the expected region of reception.

In the future technologies, this could combine with OEM and automakers pursuing their own means to drastically improve road safety and reduce fatalities even further. Continuing on this trend of improvement, it is only a matter of time until this can be integrated large scale if the performance becomes practical while the cost remains feasible.

## References

1. Fact Sheet 310-The Top 10 Causes of Death, World Health Org. Geneva, Switzerland (2017, January). Who.int. (2020, January). The Top 10 Causes Of Death. [online] Available at <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>. Accessed 7 Dec 2020
2. Arnon, S.: Optimised optical wireless car-to-traffic-light communication. *Trans. Emerg. Telecommun. Technol.* **25**(6), 660–665 (2014). <https://doi.org/10.1002/ett.2817>
3. Cailean, A.-M., Dimian, M.: Current challenges for visible light communications usage in vehicle applications: a survey. *IEEE Commun. Surv. Tutorials* **19**(4), 2681–2703 (2017)
4. Visible light communications cooperative architecture for the intelligent transportation system. Ieeexplore.ieee.org (2021). [Online]. Available <https://ieeexplore.ieee.org/document/6736001/>. Accessed 27 Oct 2021
5. Visible light communications: application to cooperation between vehicles and road infrastructures. Ieeexplore.ieee.org (2021). [Online]. Available <https://ieeexplore.ieee.org/document/6232225>. Accessed 27 Oct 2021

# A Comprehensive Study of Various DC Faults and Detection Methods in Photovoltaic System



Alaa Hamza Omran, Dalila Mat Said, Siti Maherah Hussin,  
and Sadiq H. Abdulhussain

**Abstract** In the last decade, the growth of solar energy, which is a common form of renewable energy, is getting faster. More than 1.3% of global electricity is supplied by the solar energy. It is estimated that the main source of electricity in 2050 would be solar energy, with a percentage of 11% of the world's power consumption. The PV array does, however, include several parallel PV strings, and each string includes a number of non-parallel modules. Each set, whole sequence, and the module have I–V prosperities of its own, whether at traditional or condition as a fault. If PV modules are interconnected, the total I–V curve is set by interactions between them. Therefore, PV modules function as a series as powerful as the connection of the weakest together. PV arrays damage PV cables and modules and contribute to electrical shock hazards as well as fire hazards. Throughout this paper, a comprehensive study of all kinds of DC faults that might happen at the DC side for the PV system is presented; different protection techniques for PV array faults are clarified. Furthermore, this paper is also discussed and compared in detail various algorithms and techniques used for fault detection and diagnosis methods regarding each type of DC fault, as well as the challenges and analysis of these detection methods are illustrated.

**Keywords** PV system · DC fault · Solar system · Arc fault · Fault detection · I–V characteristic

---

A. H. Omran · D. M. Said (✉) · S. M. Hussin

Centre of Electrical Energy Systems (CEES), School of Electrical Engineering, Universiti Teknologi Malaysia (UTM), Johor Bahru, Malaysia  
e-mail: [Dalila@utm.my](mailto:Dalila@utm.my)

A. H. Omran

University of Information Technology and Communications, Baghdad, Iraq

S. H. Abdulhussain

Department of Computer Engineering, College of Engineering, University of Baghdad, Baghdad, Iraq

## 1 Introduction

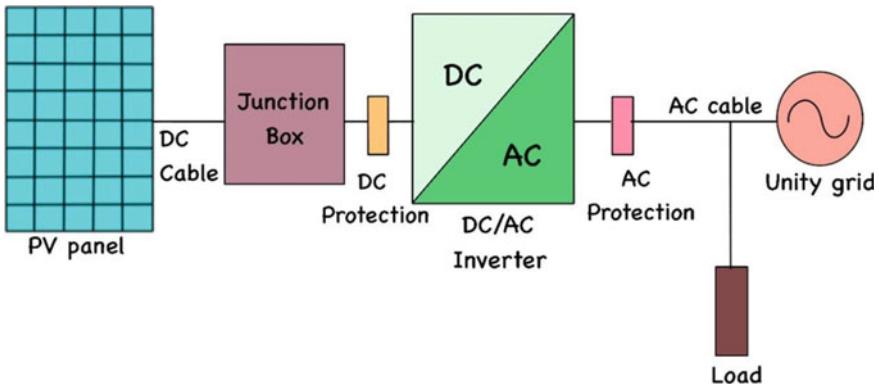
Renewable energy refers to the natural resources used to generate specific energy; these resources are constantly replenished as water, wind, sun rays, etc. [1]. Obtaining a new sustainable power system is significant; solar energy is considered one of the essential sources that are widely used around the world [2, 3]. Today, photovoltaic markets are expanding rapidly according to their advantages, which are the long life of the photovoltaic panel, installation in various locations, and geographical circumstances such as impracticable mountains and positions. Usage on mobile hosts, ease of maintenance, installation as off-grid, and ability to connect to the grid utility that depicted the future as bright for worldwide use of photovoltaic systems [4, 5].

However, the fast rate of progression is essential due to many factors, such as the need for alternative resources to produce electricity other than fossil fuel, concerns about the global climate, reduced photovoltaic costs, and interest in transmitting energy sources to increase the efficiency of the power system [6, 7]. Inverter efficiencies that convert DC current to an AC current through modules are already close to a limit of around 99% [8].

The need for lower cost and higher device performance motivates the investigators to improve the reliability of the PV system. Fault analysis in PV solar arrays is the key task for removing any forms of undesirable and hazardous situations that occur due to fault involvement in PV array operation. They should be found and cleared off quickly. Without precise identification of faults, all the faults in the PV array cannot be grouped and categorized precisely; this can also lead to power losses as well as to fire risks, electrical shocks, and many safety issues [9]. PV systems are exposed to different types of failures; hence, it is important to detect what sort of failures can be obtained in the real system before beginning the device monitoring and fault diagnosis methods. Therefore, the objective of this research is to explain and clarify the modern detection approaches for the diagnosis of DC arc faults in PV systems. The functionalities and limitations of different DC arc models were discussed and compared. Understanding the characteristics and instruments of DC arc faults is critical for the development of an effective detection algorithm. And summarize the various kinds of the DC fault in the PV system associated with the exciting detection methods to be an entrance guide for anyone who wants to have a general understanding of the research over the DC fault in the PV systems. The features and limitations of various protection and detection methods are illustrated, compared, and presented.

## 2 Conceptual Illustration

The rest of the paper is organized, as Sect. 3 presents the Photovoltaic System Structure. All the electrical faults that may occur in the DC side of the PV arrays are recognized and classified in Sect. 4. Sect. 4 illustrates the conventional protection



**Fig. 1** PV system diagram with a local load [11]

methods of the PV array faults; various techniques and algorithms of diagnosing and detecting each type of the DC fault are discussed and compared in detail in Sect. 5. The challenges and analysis of the existing detection method are illustrated in Sect. 6. Finally, the conclusion of this paper is presented in Sect. 7.

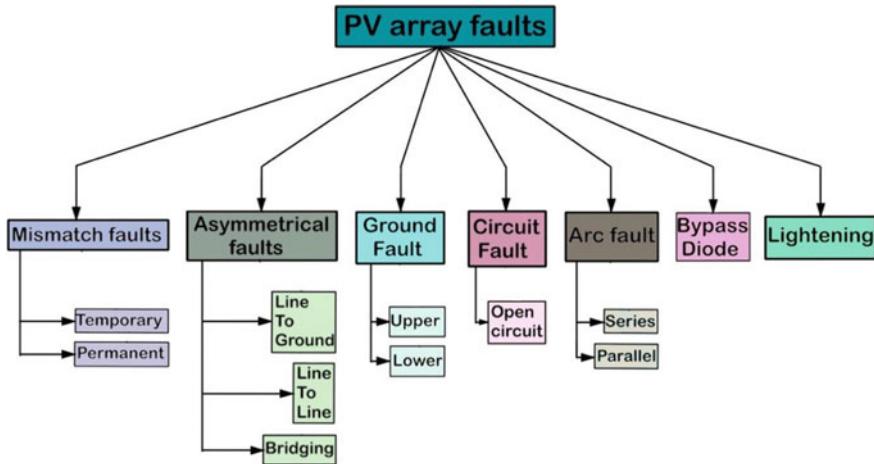
### 3 Photovoltaic System Structure

PV configuration can be classified on the basis of the power levels: commercial, house, residential, and utility scales. These degrees are often arranged on the basis of their relation to the power grid consisting of grid-connected or stand-alone systems [10]. Based on Fig. 1, that clarified the structure of the system, model usually mainly consists of PV modules connected to the DC/AC inverter, usually by a junction box. Usually, blocking diodes are integrated into each solar panel creation. The PV panel is also the combined result of PV cells, which produce electricity when illuminated. The systems also contain the factors that contribute to energy generation in a PV system.

In addition, each cell series is connected to a bypass diode that prevents modules from displaying similar behavior to the receivers and cell heating when partially illuminated.

### 4 Faults of PV Array

For the output power, the power of a photovoltaic array has consisted of the person of the total assembly. As normal, the array output power is extremely near predicted



**Fig. 2** PV array faults

value during the operation period. Some factors reduce the output power of photovoltaic arrays and are called any aspect that minimizes energy production (fault). Faults may be permanent or conditional. These factors occur on PV arrays and will affect their durability and effectiveness destructively [12, 13]. Figure 2 indicates the various forms of faults with PV.

#### 4.1 Mismatch Faults

If the solar cell, module, and array's electrical parameters change from their initial state, the mismatches' faults will occur. The effects of these faults are the losses of the high power and irreversible damages [14]; however, it can be either permanent or temporary [15–22], where these types are discussed in more detail in the next subsections.

##### 4.1.1 Temporary

Temporary mismatch faults, which is the first kind of mismatch faults, it has four different groups that are listed as follows:

1. Partial shading: This is a fault when a transitory shadow is formed on the module (e.g., by the cloud), which results in the formation of the hotspot in the shaded module.
2. Snow covering: This fault occurs due to the weather conditions and geographical location of the system installation.

3. Dust/bird dropping/leaves: Dust, bird dropping, and leaves may be on the module of solar the top.
4. Irradiance distribution: During the daytime, sunlight reaches the PV module with different radiation intensity.

#### 4.1.2 Permanent

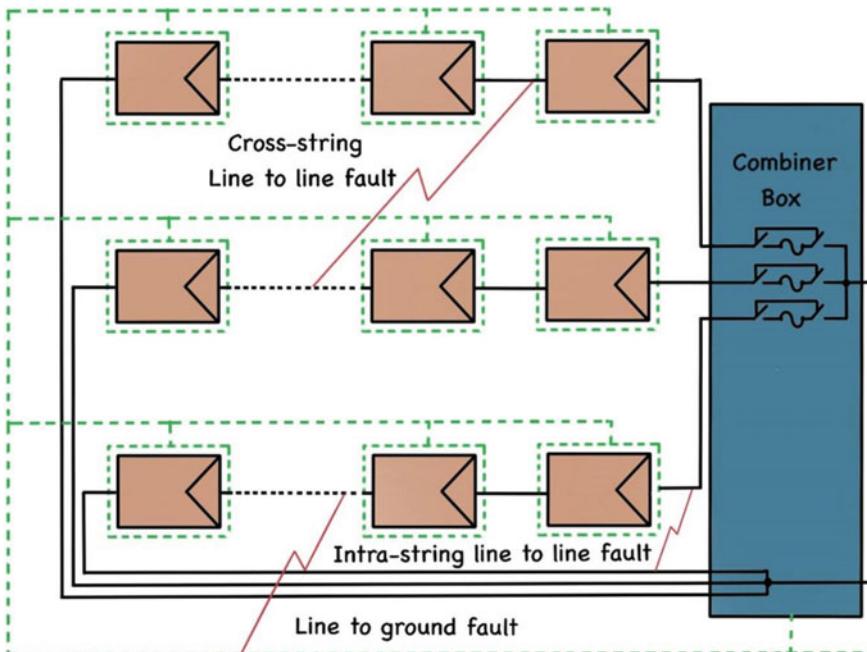
The second type of mismatch faults is the permanent mismatch faults; it can be classified into different groups which are listed as follows:

1. Hot spot: When solar cells operate in the opposite bias and destroy the power instead of producing, the hotspot fault occurs. Factors like diode fault of bypass, shading, and mismatching among electrical properties are the creation of hotspot.
2. Soldering: This type of failure results in careless manual soldering and breaks the cells. Hard shading. The appearance of permanent shadows of solids such as modules buildings minimizes the system output voltage.
3. Microcracks: At any time of cell life, cell cracking can occur. Things like the slicing of wafer, stringing, production of cell, and others that are embedding is the cause of the crack. The installation process is one of the most important of these factors.
4. Degradation of modules: The first cause is the adhesive materials regression between cells and glass. The color of materials converts to yellow from white and a few times to brown resulting in a decreased light arriving in cells of solar and thus minimizing power generation. The second cause is delamination that resulting in gaps occurring among various PV module layers of subsequent lead to adherence losing. Reduction cases in delaminating in power generated because of the reflection of light elevate along with penetration of water inside of the module. Degradation interconnect affects both resistance of series and shunt, while degradation of contact able to increase the resistance of series.
5. Potential induced degradation: This type of fault only occurs in modules of crystalline silicon resulting in decline gradually in the performance of the module. It is attributed to currents stray in most systems as ungrounded PV; modules of PV along + ve or – ve voltage to the ground are PID subjected. The potential included degradation mostly occurs at the negative voltage concerning the potential of the ground. It is enhanced via voltages of high system, temperatures as high, and humidity as high.
6. Degradation of power induced by light: LID is considered as degradation of natural reasoned to the reaction as physical, resulting in PV cell p–n junction. It appears as silicon solar cells losing efficiency and observed as short-circuit current shrinkage and open solar cell circuit voltage.
7. Glass breakage: Solar cell glass can be damaged for some reason, such as poor packaging (transportation), installation, and hail or stone collision.
8. Delamination: When moisture penetrates the solar cell's interior, delamination occurs, and this reduces the active module spaces.

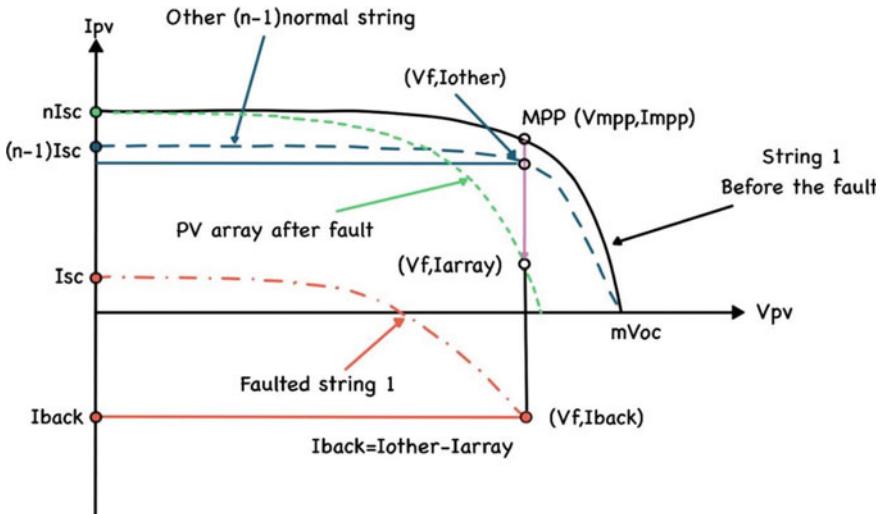
9. Discoloration: If unstable and weak quality materials are used to make cells (ethylene, vinyl, and acetate), the discoloration is occurring. For this and due to the presence of ultraviolet radiation and heat, these materials are starting to turn brown or yellow.
10. Interconnect breakage of bus bars: When the temperature changes repeatedly, the cell interconnections wear out.
11. Defects in the frame: Heavy snowfall in certain areas or falling obstacles on the PV modules can damage the structure.
12. Cell breakage: Because of the large ribbon with a large solder bond, more local stress on the cell is possible, which leads to increase in the possibility of breaking the cell.

## 4.2 Asymmetrical Fault

As illustrated in Fig. 3, the asymmetrical fault might happen within string being the same or between two strings. In a few cases, fault of line to line is named fault of bridging if it occurs between two same order modules from two various strings [23–27]. Generally, the fault of asymmetrical is divided into three parts:



**Fig. 3** Different types of asymmetrical faults in PV

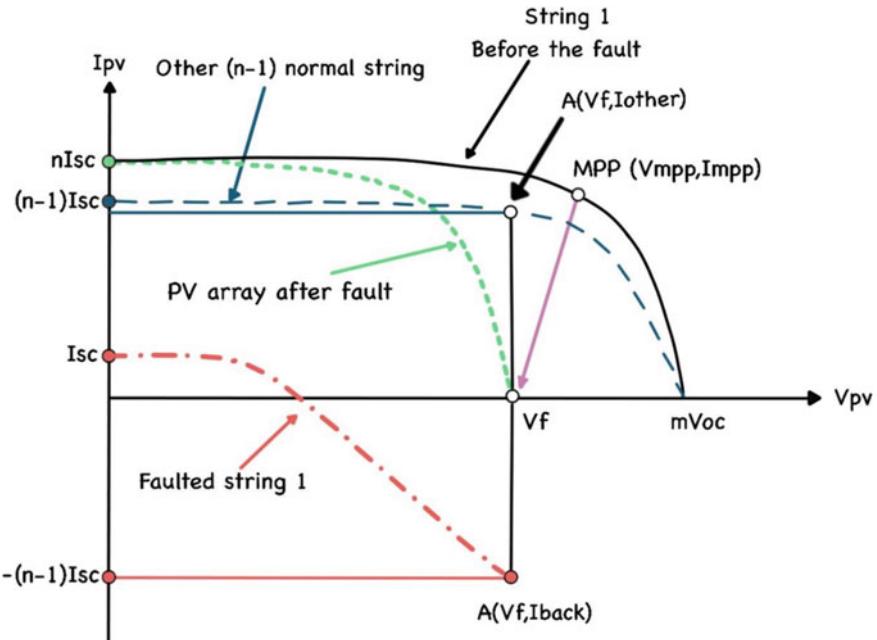


**Fig. 4** Characteristic of I–V for the line-to-line fault

1. The fault of the line to line: It happens if the connection of unintentional between PV array two nodes. The failure of the line to line is a low resistance accidental link established between different potential two points in the electric system or network. In the system of PV, fault of line to line is usually detected as the fault of short circuit among modules of PV or cables of the array with various potential. Figure 4 shows the I–V characteristic of the PV when the line-to-line fault occurs.
2. The fault of the line to the ground: It happens if a single conductor to the ground drops or is in contact with the conductor of neutral.
3. Fault bridging connection resistance between cabling and modules of PV.

#### 4.3 Ground Fault

Typically, the array of PV has various NCC metals exposed or parts of conducting such as frames of the module, racks of mounting, enclosures of metals, panels of distribution, and converters of power and endues appliances chassis. Such conductors are not carrying any current through the usual operation. Nevertheless, there is a hazard of electric shock as the risk of potential from NCC conductors is exposed if the connection of electricity is established between NCC conductors and CCCs because of a fault (i.e., insulation of melting or loss, corrosion, incorrect wiring, and cutoff of wire). Thus, all NCC conductors are connected to earth or ground by a conductor named “conductor of equipment grounding” [28]. The characteristic of the IV during the occurrence of the ground fault is shown in Fig. 5. The ground fault is considered as



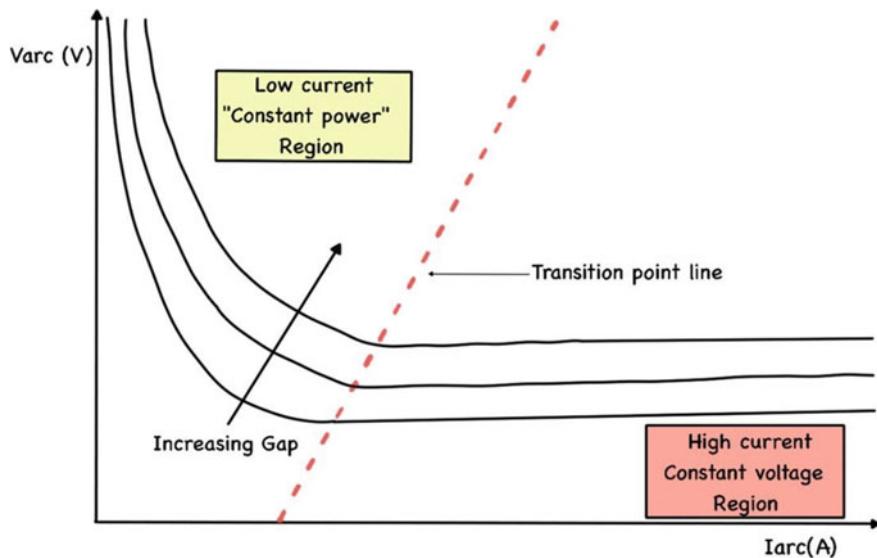
**Fig. 5** Characteristic of I–V in case of ground fault

unintentional of low impedance path through one of the current-carrying conductors and the ground [23–27]. It can be classified as:

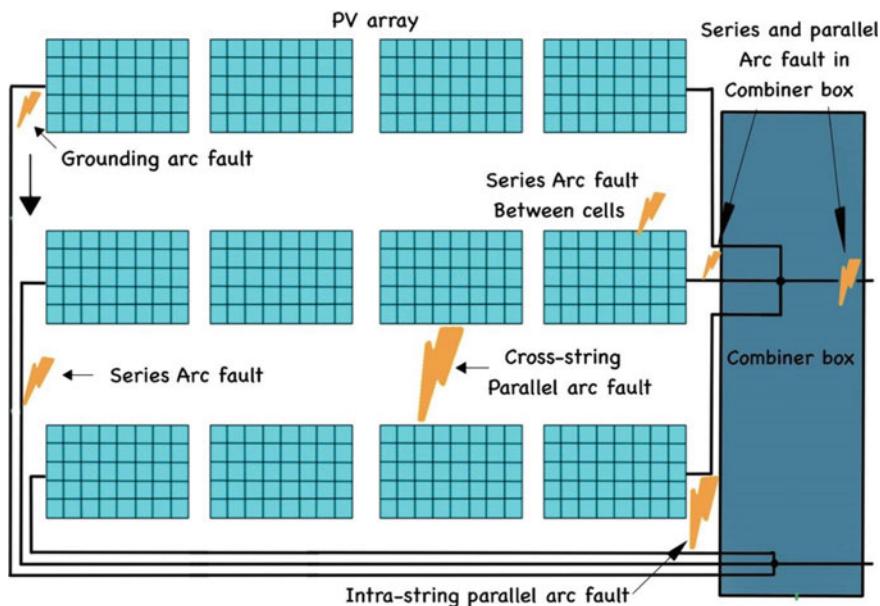
- (i) Upper ground: the path of unintentional low impedance between the last two CCC modules in the ground and the PV string.
- (ii) Lower ground: the path of unintentional low impedance between the second and the third CCC modules in the ground and PV string at large feedback current.

#### 4.4 Arc Fault

As long as systems of modern PV are with structures as a complex with electrical contacts significant number, there are dangerous situations important risks due to electric arc occurrence. Because of PV modules nature and PV design systems, it might be connectors as thousands and a significant cables amount in the system of PV. Each point of connecting is able to create the fault of the arc that increases extraordinarily the fault occurrence possibility of arc, mainly failure of the series arc. Figure 6 shows the characteristic of the V–I of arc and Fig. 7 presents the position of these faults; there are two kinds of arc faults in the system of PV: series and parallel include ground arc fault [29–33].



**Fig. 6** Characteristic of the V–I Arc



**Fig. 7** Example of arc fault

1. Faults of series Arc: solder joints degradation, connections, or wiring inside the box of the junction, screws loosening, or crimping of incorrect might elevate the resistance of the connection. Elevating temperature of operating might lead to thermal stress, resulting in disconnection as complete or accelerated aging.
2. Parallel Arc Fault: faults of the parallel arc may occur from damage to insulation because of aging, mechanical damage, or wildlife as well as last events of series arc fault; i.e., faults as a parallel arc are as follows:
  - (I) Intra-string: fault of parallel arc between two points on the current-carrying the same string conductors (CCC).
  - (II) Cross String: parallel arc fault between two points on the current-carrying the two different strings conductors.
  - (III) Ground: fault of parallel arc between a single point on a current-carrying conductor and another one at ground potential.

#### **4.5 Bypass Diode and Open-Circuit Faults**

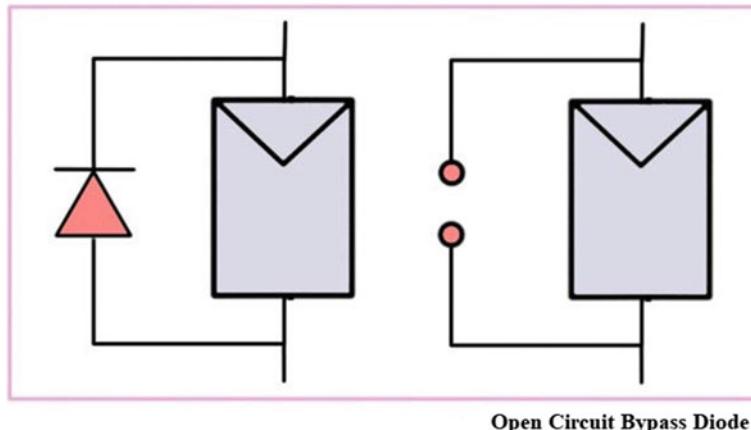
The diodes of Bypass are placed paralleling with solar cells a specific number in the module of PV. Diodes of bypass compensate for the reduction of performance and losses of power as a result of shading the module. Such diodes prevent the phenomenon of bias heating reverse, destruction of the module, and hot spot. In the case of bypass diodes absence, solar cells heat arrive high levels, which cause problems as hotspot that cause burn marks, browning, and fire at worst cases. P-V and V-I PV cell change and drop properties in the presence of the shadow, and when there are no bypass diodes, the voltage of open circuit and array maximum power is reduced sharply [34].

Inverted bypass diode fault in PV arrays comes from connections of the incorrect set via the operator. The fault of the circuit as short originated from the connection as bad among defects of manufactory or solar cells and open-circuit fault that usually occurs as an aftereffect of line-to-line faults. Circuits such as short and open fault in a PV cell are shown in Fig. 8.

The fault of open circuit is considered as a disconnection of accidental at the conductor of normal current-carrying. Such a fault happens if a problem of disconnection appears in one string of PV or more [15]. Problems of disconnection are mostly because of soldering being poor in interconnections of strings.

#### **4.6 Lightening**

Lightning is considered a phenomenon usually observed naturally on earth as well as being visible from the environment as outer. Lightning is often a deadly phenomenon because it may cause people death. In general, lightning is considered as a transient

**Short Circuit Bypass Diode****Fig. 8** Open-circuit and short-circuit bypass diode

with a discharge of high current electric where kilometers length of the path is measured. The air in the cloud between negative and positive charges and between ground and cloud behaves as an insulator. If sufficient opposite charges are building up, the insulating air capacity breaks down. Quick electricity discharge is termed as lightning [35]. This can reduce the PV system reliability and increases system repair and replacement costs. When lightning strikes the system of PV, it will create voltages and current of transient throughout the system. The voltages and current of transient become visible in the terminals of equipment and resulting in failures of dielectric and insulation within components of electronic and electrical solar PV, i.e., inverter, panels of PV, and communications and control equipment [35–38]. In this situation, there are two types of damage: the first one is a direct lightning strike to the discharge, and photovoltaic panels hit few devices as surrounding, and in this case, the panel is destroyed completely. Extreme current lightning is melting structures of semiconductor and panel frame. The second type is more famous, but it is hard to be identified. Due to the presence of overvoltage, the panel damage itself. The semiconductor structure is severely damaged, but it is not visible to the naked eye.

## 5 Fault Diagnosis and Detection Based Different Techniques

In this section, an explanation of the various methods used in the PV systems for fault detection and diagnosis is presented. Algorithms for fault diagnosis are classified into

three categories: methods-based visual, methods-based thermal, and the methods-based electrical, which can be considered as the more common methods as compared to others. Once again, electrical methods are categorized into the following groups:

- (i) Analysis of IV characteristics-based methods
- (ii) Artificial intelligence-based method
- (iii) Signal processing and statistical-based method
- (iv) Power losses-based method
- (v) Voltage and current measurements-based method

All the above detection methods are summarized in Table 1, where a brief description with the method highlighting for each method in each group is presented.

## 6 Challenging and Analyzing of Existing Method

Modern faults are developing with each day; since they are still undetected, they may cause serious problems. Overall, there is a consistent requirement to develop new technologies, and improvements are needed in existing techniques, mostly in the PV cell equivalent model. Rather than a single diode-equivalent model, through which more work is still to be performed, two diode-equivalent models are proposed recently. Plenty of Experimental work performed is a particular method. Perhaps a generic algorithm for fault detection and diagnosis evolves.

1. Present systems are designed to detect approximately up to three faults at the moment [39–67]. A full system can be built, which can detect multiple faults. Most methods are logically sound, and they can be tested experimentally.
2. Cost-effective and easy systems can be built such that small PV systems and power plants can be controlled and supervised using these systems.

The fault detection method or processing time needed to detect a specific fault remains an area of research.

Furthermore, according to recent publications, the features and benchmarks are time of detection, accuracy, reliability, Computational/Simplicity and the cost. These criteria should be considered in designing proposed models.

## 7 Conclusion

This chapter has presented a comprehensive investigation of various types of faults for the DC side which may arise in a PV system. All of the faults are extensively discussed, as are their characteristics and underlying causes, where the definition, description, and the possibility of fault causing for each type of the fault have been illustrated, followed by a discussion of the exciting fault diagnosis and detection; various techniques and algorithms have been used along with a brief explanation,

**Table 1** A summary of the detection methods

Ref. No.	Fault type	Method-based	Method highlights
[39]	Ground fault, short circuiting, Line-to-Line Fault	IV characteristics A technique for fault detection, which can be integrated with Direct-attached storage (DAS)	It proposed a detection of fault method in PV array system, where sensors are required for this method
[40]	Line-to-Line Fault	IV characteristics PSIM software package is used to represent the IV characteristics	It presents a detection of fault method in PV array system, and it is suitable for PV strings
[41]	Line-to-Line Fault, Hot Spots, Arc Fault	IV characteristics Swanson Method	It produced a method of fault detection in PV module system, and it is validated experimentally and suitable for PV array
[42]	Mismatch (Open Circuit and Short Circuit)	IV characteristics Partial diagram of the Sentalis 1000PV solar power plant	It presents a detection of fault method in PV module system; it is validated experimentally and can be considered as low-cost method
[43]	Hotspot fault and Open Circuit fault	IV characteristics Single mode equivalent circuit	It presents a method of hotspot fault diagnosis, and It is validated experimentally and appropriated for medium size array
[44]	Diode Fault	IV characteristics Single mode equivalent circuit	The research presented a diagnosis method of diode fault and module, and this method is appropriate for the both permanent and temporary fault
[45]	Mismatch, Bypass Diode, Ground Fault	IV characteristics Simplified equivalent circuit PV model	This work proposed a method of fault detection for ground and diode and appropriated for medium size array
[47]	Bypass Diode and Shading	IV characteristics One-diode model equivalent circuit PV model	It suggested a diagnosis method of fault module in PV system, and this method is appropriate for the both permanent and temporary fault

(continued)

**Table 1** (continued)

Ref. No.	Fault type	Method-based	Method highlights
[48]	Overheating, power device fault, lesser output voltage and the grid connected voltmeter with incorrect readings	Artificial Intelligence	This paper presents a diagnoses method, where various faults power is detected. This method was tested on a small scale application.
[51]	Constant Energy loss, changing energy loss and total Blackout	AI	The research proposed a diagnoses method, where various faults power is diagnoses which are: overheating, lesser output voltage, and incorrect readings, and this method is rather complex and appropriated for a medium scale
[52]	Broken cell, water infiltration and grounded shading	AI	It presents a detection method of fault detection in PV module system; this method is rather complex, cost-effective and appropriated for a medium scale
[53]	Series Losses	AI	A fault detection method of PV module was presented; it is benefit for the temporary fault detection and considered as medium cost
[54]	Short Circuit	AI	It suggested a fault detection and classification method of diode fault and module; it can be considered as low computation and fast
[55]	Partial shading, uniform shading and increasing in resistance	AI	It demonstrated a fault detection and classification method PV module, where various faults power is diagnoses which are: overheating, lesser output voltage, and incorrect readings, and it can be considered as low computation and fast

(continued)

**Table 1** (continued)

Ref. No.	Fault type	Method-based	Method highlights
[56]	Mismatch, shading degradation charging module and battery module	AI	It presents a fault detection and analysis method of PV module; this method was tested using MATLAB and not validated as experimentally
[58]	Mismatch (open and short circuit), ground fault	Signal processing and statistical	It presents a time domain fault diagnosis method of line-to-line and array fault in PV system; they were tested on the large size of the PV panels; however, they need to a reflectometry equipment
[59]	Mismatch (open and short circuit)	Signal processing and statistical	A fault detection and localization method PV module that is used in power conditioners with the ability of integration, appropriated for a medium-scale PV string, were demonstrated in this research
[61]	Arc Fault, Ground fault	Signal processing and statistical	It proposes a fault detection and localization method PV module, appropriated for a medium-scale PV string
[62, 63]	Arc Fault	Signal processing and statistical	These papers reveal a fault detection method in PV array system and appropriated for PV module with a medium cost
[64]	Arc Fault	Signal processing and statistical	It presents a fault detection method in PV array system and appropriated for PV module
[65]	Constant energy loss, short time energy loss	Power loss analysis	It proposes a fault detection and diagnosis method using MATLAB, where a comparison of electrical parameters is used in the detection process

(continued)

**Table 1** (continued)

Ref. No.	Fault type	Method-based	Method highlights
[66]	Line-to-Line and open circuit	Voltage and current measurements	The research proposes a fault detection method in PV using learning model
[67]	Mismatch (open and short circuit)	Voltage and current measurements	It presents a fault detection method in PV using Arduino board to detect the mismatch fault
[68]	Detect areas with imperfections and power generation	Voltage and current measurements	It demonstrates a new system that uses IoT to transfer the signal in order to process it

and a method highlighting for each algorithm and techniques is presented. These methods have to offer a direct response to the fault with high accuracy in a fast way, where many fault detection and diagnosis methods address this concern. Finally, it is believed that this research will aid in understanding all of the types of failures that may influence PV system quality and cause destructive issues including such fires or even electrical shock hazards, and that it will serve as a guide for everyone who seeks to acquire a general understanding of the study on DC faults in PV systems.

**Acknowledgements** The authors would like to express the appreciation to the Ministry of Higher Education Malaysia (MOHE), the support of the sponsors [Vot Number = Q.J130000.3551.07G53 and R.J130000.7351.4J347] and also to the Universiti Teknologi Malaysia (UTM) for providing the best education and research facilities to achieve the aims and goals in research studies and works.

## References

- King, D.L., Boyson, W.E., Kratochvil, J.A.: Analysis of factors influencing the annual energy production of photovoltaic systems. In: Conference Record of the Twenty-Ninth IEEE Photovoltaic Specialists Conference, vol. 2002. pp. 1356–1361 (2002)
- Frish, J.-R.: New renewable energy resources: (a guide to the future). *Appl Sol Energy* **33**(5), 25–35 (1997)
- Samet, H., Asl, D.K., Ghanbari, T., Omran, A.H.: Optimal number and location of the required measurement units for fault detection of PV arrays. In: 2018 IEEE International Conference on Environment and Electrical Engineering and 2018 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe), pp. 1–5 (2018)
- Omran, A.H., Raheem, I., Hussein, A.D.: Minimizing the losses of PV panel generation by designing an intelligent controller based on FPGA. *Int. J. Eng. Technol.* **7**(4), 4846–4849 (2018)
- Shukla, P.N., Khare, A.: Solar photovoltaic energy: the state-of-art. *Int. J. Electr. Electron. Comput. Eng.* **3**(2), 91 (2014)
- Despotou, E., Gammal, A.E., Fontaine, B., et al.: Global Market Outlook for Photovoltaics Until 2014. The European Photovoltaic Industry Association, Brussels, Belgium (2010)

7. Omran, A.H., Abid, Y.M., Ahmed, A.S., Kadhim, H., Jwad, R.: Maximizing the power of solar cells by using intelligent solar tracking system based on FPGA. In: 2018 Advances in Science and Engineering Technology International Conferences (ASET) [Internet], pp. 1–5. IEEE (2018). Available from <https://ieeexplore.ieee.org/document/8376786/>
8. Chiradeja, P., Ramakumar, R.: An approach to quantify the technical benefits of distributed generation. *IEEE Trans. Energy Convers.* **19**(4), 764–773 (2004)
9. King, D.L., Kratochvil, J.A., Boyson, W.E.: Photovoltaic array performance model. United States. Department of Energy (2004)
10. Refaat, A., Kalas, A., Daoud, A., Bendary, F.: A control methodology of three phase grid connected PV system. In: Power Systems Conference (Clemson University USA), vol. 2013 (2013)
11. Brooks, B.: The bakersfield fire: a lesson in ground-fault protection. *SolarPro Mag.* **62** (2011)
12. Zhao, Y., Lehman, B., de Palma, J.-F., Mosesian, J., Lyons, R.: Fault analysis in solar PV arrays under: low irradiance conditions and reverse connections. In: 2011 37th IEEE Photovoltaic Specialists Conference, pp. 2000–2005 (2011)
13. AbdulMawjood, K., Refaat, S.S., Morsi, W.G.: Detection and prediction of faults in photovoltaic arrays: a review. In: 2018 IEEE 12th International Conference on Compatibility, Power Electronics and Power Engineering (CPE-POWERENG 2018) [Internet]. IEEE, pp. 1–8 (2018). Available from <https://ieeexplore.ieee.org/document/8372609/>
14. Whaley, C.: Best practices in photovoltaic system operations and maintenance. No. NREL/TP-7A40-67553. National Renewable Energy Lab. (NREL), Golden, CO (United States) (2016)
15. Pillai, D.S., Rajasekar, N.: A comprehensive review on protection challenges and fault diagnosis in PV systems. *Renew. Sustain. Energy Rev.* **91**, 18–40 (2018)
16. Madeti, S.R., Singh, S.N.: A comprehensive study on different types of faults and detection techniques for solar photovoltaic system. *Sol. Energy.* **158**, 161–185 (2017)
17. Triki-Lahiani, A., Abdelghani, A.B.-B., Slama-Belkhodja, I.: Fault detection and monitoring systems for photovoltaic installations: a review. *Renew. Sustain. Energy Rev.* **82**, 2680–2692 (2018)
18. Jain, P., Xu, J.-X., Panda, S.K., Poon, J., Spanos, C., Sanders, S.R.: Fault diagnosis via PV panel-integrated power electronics. In: 2016 IEEE 17th Workshop on Control and Modeling for Power Electronics (COMPEL). pp. 1–6 (2016)
19. Heidari, N., Gwamuri, J., Townsend, T., Pearce, J.M.: Impact of snow and ground interference on photovoltaic electric system performance. *IEEE J. Photovoltaics.* **5**(6), 1680–1685 (2015)
20. Nguyen, X.H.: Matlab/Simulink based modeling to study effect of partial shadow on solar photovoltaic array. *Environ. Syst. Res.* **4**(1), 20 (2015)
21. Hu, Y., Zhang, J., Cao, W., Wu, J., Tian, G.Y., Finney, S.J., et al.: Online two-section PV array fault diagnosis with optimized voltage sensor locations. *IEEE Trans. Ind. Electron.* **62**(11), 7237–7246 (2015)
22. Sabbaghpur Arani, M., Hejazi, M.A.: The comprehensive study of electrical faults in PV arrays. *J Electr. Comput. Eng.* **2016** (2016)
23. Zhao, Y., De Palma, J.-F., Mosesian, J., Lyons, R., Lehman, B.: Line–line fault analysis and protection challenges in solar photovoltaic arrays. *IEEE Trans. Ind. Electron.* **60**(9), 3784–3795 (2012)
24. Brooks, B.: The ground-fault protection blind spot: A safety concern for larger photovoltaic systems in the United States. A solar ABCs White paper (2012)
25. Bower, W.I., Wiles, J.C.: Analysis of grounded and ungrounded photovoltaic systems. In: Proceedings of 1994 IEEE 1st World Conference on Photovoltaic Energy Conversion-WCPEC (A Joint Conference of PVSC, PVSEC and PSEC), pp. 809–812 (1994)
26. Eskandari, A., Mohammadreza, A., Jafar M., Amir N.: A weighted ensemble learning-based autonomous fault diagnosis method for photovoltaic systems using genetic algorithm. Available at SSRN 4058819
27. Ball, G., Brooks, B., Flicker, J., Johnson, J., Rosenthal, A., Wiles, J.C., et al.: Inverter ground-fault detection ‘blind spot’ and mitigation methods. Solar American Board for Codes and Standards (2013)

28. Alam, M.K., Khan, F., Johnson, J., Flicker, J.: A comprehensive review of catastrophic faults in PV arrays: types, detection, and mitigation techniques. *IEEE J. Photovoltaics* [Internet]. **5**(3), 982–997 (2015, May). Available from <http://ieeexplore.ieee.org/document/7045450/>
29. Lu, S., Phung, B.T., Zhang, D.: A comprehensive review on DC arc faults and their diagnosis methods in photovoltaic systems. *Renew. Sustain. Energy Rev.* [Internet]. **89**, 88–98 (2018, June). Available from <https://linkinghub.elsevier.com/retrieve/pii/S1364032118300996>
30. Zhu, L., Ji, S., Liu, Y.: Generation and developing process of low voltage series DC arc. *IEEE Trans. Plasma Sci.* [Internet]. **42**(10), 2718–2719 (2014 October). Available from <http://ieeexplore.ieee.org/document/6841042/>
31. Uriarte, F.M., Gattozzi, A.L., Herbst, J.D., Estes, H.B., Hotz, T.J., Kwasinski, A., et al.: A DC arc model for series faults in low voltage microgrids. *IEEE Trans. Smart Grid* [Internet]. **3**(4), 2063–2070 (2012, December). Available from <http://ieeexplore.ieee.org/document/6305496/>
32. Smith, D.: Arc flash hazards on photovoltaic arrays. Color State Univ Fort Collins, Color USA (2013)
33. Klement, K.: DC arc flash studies for solar photovoltaic systems: challenges and recommendations. *IEEE Trans. Ind. Appl.* [Internet]. **51**(5), 4239–4244 (2015, September). Available from <http://ieeexplore.ieee.org/document/7094947/>
34. Hamada, T., Nakamoto, K., Nanno, I., Fujii, M., Oke, S., Ishikura, N.: Characteristics of failure Schottky barrier diode and PN junction diode for bypass diode using induced lightning surge test. In: 2018 7th International Conference on Renewable Energy Research and Applications (ICRERA), pp. 482–486 (2018)
35. Ahmad, N.I., Ab-Kadir, M.Z.A., Izadi, M., Azis, N., Radzi, M.A.M., Zaini, N.H., et al.: Lightning protection on photovoltaic systems: a review on current and recommended practices. *Renew. Sustain. Energy Rev.* **82**, 1611–1619 (2018)
36. Méndez, Y., Acosta, I., Rodriguez, J.C., Ramirez, J., Bermúdez, J., Martinez, M.: Effects of the PV-generator's terminals connection to ground on electromagnetic transients caused by lightning in utility scale PV-plants. In: 2016 33rd International Conference on Lightning Protection (ICLP), pp. 1–8 (2016)
37. Hernandez, J.C., Vidal, P.G., Jurado, F.: Lightning and surge protection in photovoltaic installations. *IEEE Trans. Power Deliv.* **23**(4), 1961–1971 (2008)
38. Benesova, Z., Haller, R., Birk, J., Zahlmann, P.: Overvoltages in photovoltaic systems induced by lightning strikes. In: 2012 International Conference on Lightning Protection (ICLP), pp. 1–6 (2012)
39. Stellbogen, D.: Use of PV circuit simulation for fault detection in PV array fields. In: Conference Record of the Twenty Third IEEE Photovoltaic Specialists Conference—1993 (Cat No93CH3283-9) [Internet], pp. 1302–1307. IEEE. Available from <http://ieeexplore.ieee.org/document/346931/>
40. Chao, K.-H., Ho, S.-H., Wang, M.-H.: Modeling and fault diagnosis of a photovoltaic system. *Electr Power Syst Res.* **78**(1), 97–105 (2008)
41. Kaplanis, S., Kaplani, E.: Energy performance and degradation over 20 years performance of BP c-Si PV modules. *Simul. Model. Pract. Theor.* **19**(4), 1201–1211 (2011)
42. Gokmen, N., Karatepe, E., Celik, B., Silvestre, S.: Simple diagnostic approach for determining of faulted PV modules in string based PV arrays. *Sol Energy*. **86**(11), 3364–3377 (2012)
43. Tina, G.M., Cosentino, F., Ventura, C.: Monitoring and diagnostics of photovoltaic power plants. In: *Renewable Energy in the Service of Mankind*, vol. II, pp. 505–516. Springer, Berlin (2016)
44. Chine, W., Mellit, A., Pavan, A.M., Lugh, V.: Fault diagnosis in photovoltaic arrays. In: 2015 International Conference on Clean Electrical Power (ICCEP), pp. 67–72 (2015)
45. Fezzani, A., Mohammed, I.H., Drid, S., Chrifi-alaoui, L.: Modeling and analysis of the photovoltaic array faults. In: 2015 3rd International Conference on Control, Engineering & Information Technology (CEIT), pp. 1–9 (2015)
46. Hachana, O., Tina, G.M., Hemsas, K.E.: PV array fault diagnostic technique for BIPV systems. *Energy Build.* **126**, 263–274 (2016)

47. Wu Y, Lan Q, Sun Y. Application of BP neural network fault diagnosis in solar photovoltaic system. In: 2009 International conference on Mechatronics and Automation, pp. 2581–2585 (2009)
48. Coleman, A., Zalewski, J.: Intelligent fault detection and diagnostics in solar plants. In: Proceedings of the 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems [Internet], pp. 948–953. IEEE (2011). Available from <http://ieeexplore.ieee.org/document/6072914/>
49. Ducange, P., Fazzolari, M., Lazzerini, B., Marcelloni, F.: An intelligent system for detecting faults in photovoltaic fields. In: 2011 11th International Conference on Intelligent Systems Design and Applications [Internet], pp. 1341–1346. IEEE (2011). Available from <http://ieeexplore.ieee.org/document/6121846/>
50. Spataru, S., Sera, D., Kerekes, T., Teodorescu, R.: Detection of increased series losses in PV arrays using fuzzy inference systems. In: 2012 38th IEEE Photovoltaic Specialists Conference, pp. 464–469 (2012)
51. Rezgui, W., Mouss, L.-H., Mouss, N.K., Mouss, M.D., Benbouzid, M.: A smart algorithm for the diagnosis of short-circuit faults in a photovoltaic generator. In: 2014 First International Conference on Green Energy ICGE 2014, pp. 139–43 (2014)
52. Spataru, S., Sera, D., Kerekes, T., Teodorescu, R.: Diagnostic method for photovoltaic systems based on light I-V measurements. *Sol Energy* **119**, 29–44 (2015)
53. Mohamed, A.H., Nassar, A.M.: New algorithm for fault diagnosis of photovoltaic energy systems. *Int. J. Comput. Appl.* **114**(9)
54. Schirone, L., Califano, F.P., Pastena, M.: Fault detection in a photovoltaic plant by time domain reflectometry. *Prog. Photovoltaics Res. Appl.* **2**(1), 35–44 (1994)
55. Takashima, T., Yamaguchi, J., Otani, K., Kato, K., Ishida, M.: Experimental studies of failure detection methods in PV module strings. In: 2006 IEEE 4th World Conference on Photovoltaic Energy Conference, pp. 2227–2230 (2006)
56. Takashima, T., Yamaguchi, J., Ishida, M.: Disconnection detection using earth capacitance measurement in photovoltaic module string. *Prog. Photovoltaics Res. Appl.* **16**(8), 669–677 (2008)
57. Johnson, J.D., Kuszmaul, S.S., Bower, W.I., Schoenwald, D.A.: Using PV module and line frequency response data to create robust arc fault detectors
58. Johnson, J., Pahl, B., Luebke, C., Pier, T., Miller, T., Strauch, J., et al.: Photovoltaic DC arc fault detector testing at Sandia National Laboratories. In: 2011 37th IEEE Photovoltaic Specialists Conference, pp. 3614–3619 (2011)
59. Dini, D.A., Brazis, P.W., Yen, K.-H.: Development of arc-fault circuit-interrupter requirements for photovoltaic systems. In: 2011 37th IEEE Photovoltaic Specialists Conference, pp. 1790–1794 (2011)
60. Chouder, A., Silvestre, S.: Automatic supervision and fault detection of PV systems based on power losses analysis. *Energy Convers. Manag.* **51**(10), 1929–1937 (2010)
61. Zhao, Y., Ball, R., Mosesian, J., de Palma, J.-F., Lehman, B.: Graph-based semi-supervised learning for fault detection and classification in solar photovoltaic arrays. *IEEE Trans. Power Electron.* **30**(5), 2848–2858 (2014)
62. Mahendran, M., Anandharaj, V., Vijayavel, K., Winston, D.P.: Permanent mismatch fault identification of photovoltaic cells using arduino. *ICTACT Journal on Microelectronics*. **1**(2), 79–82 (2015)
63. Shakya, S.: A self monitoring and analyzing system for solar power station using IoT and data mining algorithms. *J. Soft Comput. Paradigm* **3**(2), 96–109 (2021)
64. Wang, W., Liu, A.C.-F., Chung, H.S.-H., Lau RW-H, Zhang J, Lo AW-L. Fault diagnosis of photovoltaic panels using dynamic current–voltage characteristics. *IEEE Trans Power Electron.* **31**(2), 1588–1599 (2015)
65. Chao, K.-H., Chen, C.-T., Wang, M.-H., Wu, C.-F.: A novel fault diagnosis method based-on modified neural networks for photovoltaic systems. In: International Conference in Swarm Intelligence, pp. 531–539 (2010)

66. Syafaruddin, Karatepe, E., Hiyama, T.: Controlling of artificial neural network for fault diagnosis of photovoltaic array. In: 2011 16th International Conference on Intelligent System Applications to Power Systems [Internet], pp. 1–6. IEEE (2011). Available from <http://ieeexplore.ieee.org/document/6082219/>
67. Schirone, L., Califano, F.P., Moschella, U., Rocca, U.: Fault finding in a 1 MW photovoltaic plant by reflectometry. In: Proceedings of 1994 IEEE 1st World Conference on Photovoltaic Energy Conversion-WCPEC (A Joint Conference of PVSC, PVSEC and PSEC), pp. 846–849 (1994)
68. Takashima, T., Yamaguchi, J., Otani, K., Oozeki, T., Kato, K., Ishida, M.: Experimental studies of fault location in PV module strings. Sol. Energy Mater. Sol. Cells **93**(6–7), 1079–1082 (2009)

# Forecasting and Seasonal Analysis of Air Quality Index using Machine Learning Models during COVID-19 Pandemic



Priyanka Harjule, Basant Agarwal, Ashish Burdak, Satvik Gupta,  
Saurav Singh, and Shivdeep Singh

**Abstract** In many cities, air pollution prevention and control have become a necessity. Industries and excessive automotive traffic in cities pollute the air to unacceptable levels, which have a harmful impact on human health. To defend individuals from health risks, forecasting, predicting and regulating air pollution are critical. There has been substantial research into predicting when inadequate air quality would occur. However, most studies are constrained by a lack of panel data, hence making it impossible to account for the factors, including seasonal behaviour. This paper presents methods of predicting air quality using machine learning techniques and forecasting the air pollution levels to take precautionary measures to minimise air pollution. This paper analyses concentrations of major pollutants in metropolitan areas (PM2.5, PM10, NO<sub>2</sub>, BTX, O<sub>3</sub>, CO, SO<sub>2</sub>, and NH<sub>3</sub>), contributing to air pollution. This paper also analysed the effect of lockdown due to the pandemic COVID-19 on the AQI at different places at Jaipur. It includes three locations of Jaipur from where the data are gathered; then, the paper will suggest the solutions that can be implemented to achieve the most desirable results, which will lead to lower levels of pollution at a lower cost.

---

P. Harjule (✉)

Department of Mathematics, Malaviya National Institute of Technology, Jaipur, India  
e-mail: [priyanka.maths@mnit.ac.in](mailto:priyanka.maths@mnit.ac.in)

B. Agarwal · A. Burdak · S. Gupta · S. Singh · S. Singh

Department of Computer Science and Engineering, Indian Institute of Information Technology  
Kota, Kota, India  
e-mail: [basant.cse@iiitkota.ac.in](mailto:basant.cse@iiitkota.ac.in)

A. Burdak

e-mail: [2018kucp1054@iiitkota.ac.in](mailto:2018kucp1054@iiitkota.ac.in)

S. Gupta

e-mail: [2018kucp1095@iiitkota.ac.in](mailto:2018kucp1095@iiitkota.ac.in)

S. Singh

e-mail: [2018kucp1028@iiitkota.ac.in](mailto:2018kucp1028@iiitkota.ac.in)

S. Singh

e-mail: [2018kucp1048@iiitkota.ac.in](mailto:2018kucp1048@iiitkota.ac.in)

**Keywords** Time series forecasting · Machine learning · Air quality index · Seasonal variation · COVID-19

## 1 Introduction

The most difficult challenges, such as air pollution and its influence on adjacent inhabitants, must be handled first and foremost. Air pollutants are discharged into the atmosphere from a variety of origins, including both natural and artificial action [1][1]. The primary step in mitigating air pollution is to monitor and evaluate the ambient air quality. The scientific community, policymakers, regulators and most importantly, the general people are frequently presented with large amounts of monitoring data that do not express air quality status clearly. So to solve this problem, a new concept of air quality index (AQI) is introduced in which all contaminants are given equal weight in calculating the AQI value [3]. The quality rating for each contaminant is computed using observed and standard values. AQI is a statistical value used to communicate the air quality in a given area or town on an hourly or daily basis [4]. The principal goal of AQI is to preserve public health particularly that of vulnerable persons like the elderly, children and asthmatics [5]. In acknowledgement of this concern, many researchers have spent decades examining and creating various models and methodologies for air quality analysis and evaluation. Numerous researches have been undertaken to investigate air pollution trends and their potentially dangerous repercussions. One recent study found that pollutant concentrations in Delhi are far higher than the legal levels [6]. Another study [7] looked into the consequences of vehicular air pollution on human health. These findings indicate that air quality monitoring and regulation will be required in the future.

A convenient AQI forecast would reveal the anticipated air quality trend and empower the government to take more effective and efficient restorative measures. In 2018–2019, the air quality in Jaipur deteriorated at an alarming rate due to rising levels of contaminants in the ambient air. As a consequence, it has been decided to concentrate the research on air quality in the Jaipur region.

Specific objectives of the research

- Prediction of AQI using supervised machine learning models
- Forecast AQI and some major pollutants in Jaipur using SARIMA
- Seasonal analysis of AQI in Jaipur
- Lockdown analysis.

This paper has incorporated machine learning models like linear regression, random forest regression, gradient boost regression, XGBoost regressor, decision tree and adaptive boosting. These regressors were employed to predict the AQI. Based on the root mean square error (RMSE), mean absolute error (MAE) and their R<sup>2</sup> value, a comparative study is performed in order to figure out the best fitting model. In general, XGBoost regressor comes out to be the best model for the location

taken. Later, we use SARIMA to forecast the AQI and some major pollutants which played a crucial role in AQI.

An effort is undertaken to quantify AQI fluctuations on an annual and seasonal basis for the Jaipur region over a four-year period. This research was administered during the summer, monsoon, post-monsoon and winter seasons. In addition, the year-by-year frequency of incidence of AQI in each category for all three sites is examined, providing an in-depth examination of trends during the research period. This type of research is unprecedented, which takes account of the seasonal variables such as temperature( $^{\circ}\text{C}/^{\circ}\text{F}$ ), visibility (km or miles), pressure, humidity (%), wind speed (km/h, mph, knots or m/s), fog days, rain days, thunder days, sun hour, snow amount (cm), rain amount (mm or inches), UV index and incorporated it with AQI variables such as particulate matter (PM)2.5, PM10, BP, NO, Nox, NO<sub>2</sub>, NH<sub>3</sub>, SO<sub>2</sub>, CO, ozone, benzene, xylene, MP-xylene and produced significant results.

Now coming to the next aspect about which the paper intensely discusses the effects of COVID-19 on AQI levels, the coronavirus (also known as COVID-19) pandemic has wreaked havoc across India, including Rajasthan. Every state is battling to treat sick people and stop the spread of the virus [8]. Transportation, industry, power plants, construction activities, biomass and waste burning, road dust resuspension and residential activities are the primary contributors to air pollution. Furthermore, certain activities such as the operation of diesel generator sets, restaurants, landfill fires and other similar activities contribute to air pollution.

People's movement has been restricted during the lockdown, with the exception of vital services. Various institutional operations, such as educational and hotel services, were also halted [9]. Consequently, data analysis and comparison of data for the period before the enforcement of limitations indicated that air quality has improved in numerous towns and cities across Rajasthan [10, 11]. This paper analyses the trend related to AQI in the pre- and post-lockdown period.

This paper focuses on Jaipur's pollution levels before and after the first stage of lockdown. Pollution levels are examined in the same months of 2019–2020 to see whether there is a difference. The paper centres around the three locations of Jaipur for analysis viz., Shastri Nagar, Adarsh Nagar and Police Commissionerate. The main reason to take these locations is that we have the data of the pollutants of only these three locations at Jaipur, provided by the national air quality index portal by the Central Pollution Control Board, Ministry of Environment, Forests and Climate Change [12].

## 2 Background

The air quality index is a method for effectively communicating the condition of air quality to people in simple words. It converts complex data on numerous contaminants' air quality into a single number (index value), nomenclature and colour [13] (Fig. 1).

**Fig. 1** Breakpoints of AQI scale (0–500) ref. [13]

AIR QUALITY INDEX (AQI)	CATEGORY
0-50	Good
51-100	Satisfactory
101-200	Moderate
201-300	Poor
301-400	Very Poor
401-500	Severe

The classifications are based on air pollution concentrations in the environment and their potential health effects (known as health breakpoints) [14]. For eight pollutants (PM10, PM2.5, SO<sub>2</sub>, CO, NO<sub>2</sub>, O<sub>3</sub>, NH<sub>3</sub> and Pb) for which short-term (up to 24-h) national ambient air quality standards are mandated, AQ sub-indices and health breakpoints have been developed [15].

### 3 Related Work

There have been several articles that have used machine learning for air pollution prediction in the past. More recent studies have centred on advanced statistical learning algorithms for evaluating air quality and predicting pollution levels. Neural networks have been employed by Garcia et al. [16] and Park et al. [17] to create models for forecasting the occurrence of individual pollutants, such as particles less than 10 microns (PM10). To train their models, Park et al. [17] employed an artificial neural network (ANN).

Zheng and et al. [18] attempted to forecast the reading of an air quality monitoring station for the following 48 h using current meteorological data, weather forecasts and air quality data from the station and other stations within 100 kms. They exercised a mix of machine learning and deep learning algorithms, including a temporal predictor based on linear regression and a spatial predictor based on neural networks. Ozgur Kisi and Kulwinder Singh used statistical approaches such as multivariate adaptive regression spline, least-square support vector regression and M5 model tree models to create a model to determine pollution concentrations [19].

For AQI category prediction, Yu [20] suggested RAQ, a random forest technique. After that, Yi [21] used deep neural networks to predict AQI categories. For forecasting AQI levels, Veljanovska and Dimoski [22] used several settings to surpass k-nearest neighbour (KNN), decision tree and SVM. Altogether these are state-of-the-art research, but any detailed study which considers Jaipur as the principal significant point has not been performed recently. Because Jaipur is home to about 4 million people and pollution trends have changed drastically in past few years, a

need arises to perform a detailed analysis of Jaipur AQI levels whilst incorporating all the new advancements in machine learning modelling and new data.

Now, more about the lockdown. A statewide lockdown was imposed in India due to COVID-19, from March 24th to April 14th, and was then extended to May 3rd. Because of the lockdown, pollution levels in 88 cities around the country dropped drastically, as the analysis performed by Sharma et al. [9] clearly depicts it. Studies conducted by Srivastava et al. [11] and Kumari et al. [23] also reported improvements in the air quality over the period of lockdown. We wanted to quantify the measures of improvements for Jaipur and discover the impact of vehicles on pollution levels so AQI analysis also performed.

## 4 Experimental Settings

### 4.1 Study Area

The city of Jaipur was chosen for the study, and its geographical regions are depicted in Fig. 2. In the North Indian region, Jaipur is one of the highly polluted cities. In Jaipur, the levels of airborne particulate matter (PM2.5 and PM10) are incredibly high. PM is regarded as one of the most hazardous contaminants to human health [9]. At the moment, there are three air monitoring stations in Jaipur, each in a different location [13].

**Fig. 2** Location of Adarsh Nagar, Shastri Nagar and Police Commissionerate at Jaipur



## 4.2 Data Source and Dataset Used

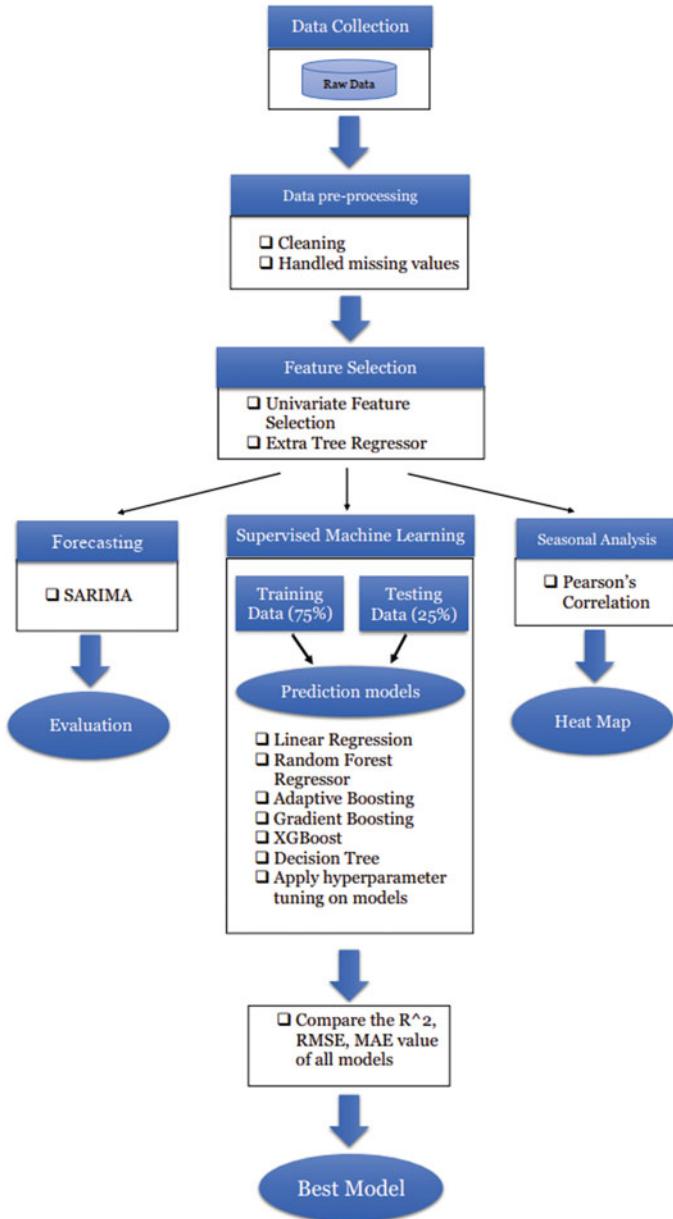
**Data Collection:** The research considered air quality and weather data for analysis. For pollution data, the national air quality index Website by the Central Pollution Control Board (CPCB), Ministry of Environment, Forests and Climate Change [12] was referred to, and Web scrapers were deployed to fetch the desired data. Air quality data involve all the pollutants which are being monitored at all the three sites taken up as a subject for this research. Weather data are collected using an API provided by the world weather online Website [24], which is considered one of the most reliable and trusted databases for weather-related information. Data from 1st-Dec-2017 to 2nd-May-2021 are considered for this paper. This period was carefully chosen to keep in mind the availability of data for all the parameters. Also, Jaipur had seen a significant spike in the vehicle and factory count since late 2017; thus, data before 2017 were not essential for this research.

**Data Pre-processing:** The data collected for all attributes from different sources had some missing values, which were impossible to retrieve because of non-availability on any source on the Internet. The dataset was combined and redistributed into separate sets for further analysis, including seasonal analysis and AQI analysis during pandemic. This paper analyses the pollution level changes based on the research paper's goals, so data were carefully divided according to seasons, dates and many other required categories. During the analysis, we will be keeping in mind the various aspects of the weather like humidity, precipitation, temperature and other related variables. To manage the missing data accurately, without losing the seasonality information, the missing data were replaced with the most relevant value observed around the date where the data were missing.

## 5 Proposed Methodology

### 5.1 Process Flow in the Proposed Work

As shown in Fig. 3, first, data have been selected from different sources, then data cleaning and data pre-processing is performed. Later, feature selection techniques are employed in order to reduce the number of non-contributing feature(s) in predicting target variables and to avoid chance of overfitting. Further, supervised machine learning models were applied, and  $R^2$ , RMSE and MAE values were compared in order to find the best fitting model to calculate the AQI. In addition, SARIMA was applied for forecasting of AQI. Also, seasonal analysis using Pearson correlation was performed, and results were analysed using heatmaps.



**Fig. 3** Workflow for estimation of air quality index

## 5.2 Feature Selection

The process of picking a subset of beginning characteristics that include significant information for predicting output data is known as feature selection. Feature extraction is used in the situation of redundant data [25]. Feature extraction entails choosing the best input parameters from a given input dataset. The resulting reduced dataset is used for further investigation [26].

As per our initial dataset, we have 28 features that are—particulate matter (PM)2.5, PM10, BP, NO, Nox, NO2, NH3, SO2, CO, ozone, benzene, xylene, MP-xylene, eth-xylene, SunHour, UVIndex, HealthIndex, WindChillC, Wind Gust km/h, Cloud-Cover, humidity, precipMM, pressure, tempC, wind direction degree, windspeed km/h and one target feature—AQI, totalling 28 features.

A model having extraneous features can damage the model; including more features make the model more complex and that may lead to overfitting [27]. Therefore, it becomes imperative to eliminate irrelevant features before training.

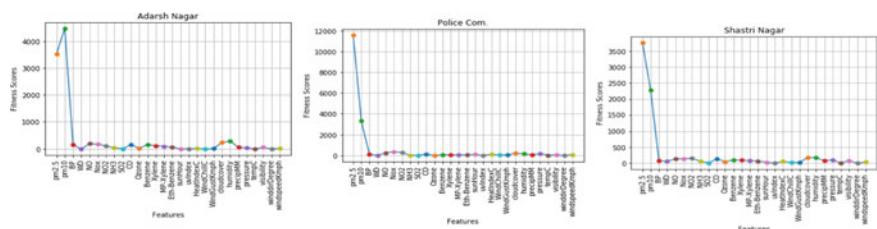
### Univariate Feature Selection

Univariate feature selection yields two primary features at Adarsh Nagar, Shashtri Nagar and Police Commissionerate, which are particulate matter PM2.5 and PM10, whilst other features have almost identical feature scores.

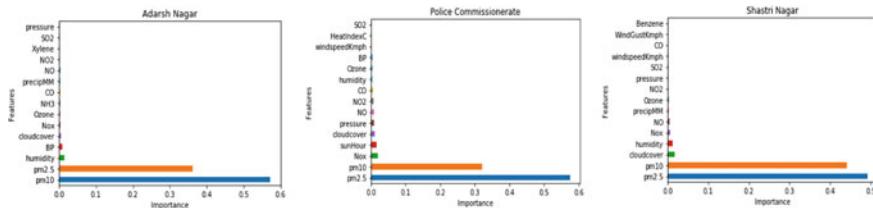
### Extra Tree Regressor

After analysing the features mentioned above using feature selection techniques, i.e. univariate and feature selection using extra tree regressor, 15 principal features are found, different for different neighbourhoods as discussed below.

For Adarsh Nagar—PM2.5, PM10, BP, NO, Nox, NO2, CO, benzene, xylene, MP-xylene, eth-benzene, CloudCover, humidity, precipMM, visibility and one target feature AQI making it to total 16 features to work upon as shown in Figs. 4 and 5. Whilst for Police Commissionerate—PM2.5, PM10, BP, NO, Nox, NO2, CO, benzene, eth-benzene, SunHour, HeatIndex, CloudCover, humidity, pressure, wind-speed and one target feature as previously mentioned, total 16 features to work upon as shown in Figs. 4 and 5. And for Shastri Nagar—PM2.5, PM10, BP, NO, Nox,



**Fig. 4** Feature scores versus features using univariate feature selection for Adarsh Nagar, Police Commissionerate and Shastri Nagar



**Fig. 5** Features versus their importance using extra tree regressor for Adarsh Nagar, Police Commissionerate and Shastri Nagar

NO<sub>2</sub>, CO, benzene, xylene, MP-xylene, CloudCover, humidity, precipMM, pressure, visibility and one target feature as previously mentioned, a total of 16 features to work upon as shown in Figs. 4 and 5.

### 5.3 SARIMA

Seasonal autoregressive integrated moving average (SARIMA) or simply seasonal ARIMA is a variant of ARIMA that explicitly handles univariate time series data with a seasonal component. First, ‘S’ in the word SARIMA refers to seasonality. This is the thing that makes it different from the standard ARIMA version. Further discussed in the seasonality analysis, it is clear that seasons significantly influence air pollutants’ levels, and thus AQI involves a significant seasonal movement. So to forecast the AQI for the future, it is essential to keep hold of the seasonality interference to get better and more accurate results. In this case, SARIMAX has used additive decomposition to take care of the seasonal movement of data.

The autoregressive (AR) relates to making predictions based on lagged values of our target variable. Let us take an example; we might estimate tomorrow’s AQI using today’s, yesterday’s and the day before yesterday’s AQI data. It makes predictions based on three lagged data to be an AR(3) model. Now, ‘I’ means integrated. It implies that rather than taking the raw target numbers, their difference is found. For example, an AQI model will forecast tomorrow’s AQI, and an AQI prediction model might aim to anticipate tomorrow’s change in AQI (i.e. tomorrow’s AQI minus today’s AQI). We need this because many time series have a trend, causing the raw data to be non-stationary. By subtracting the difference, our Y variable becomes stationary.

Then finally for moving average (MA) lagged prediction errors are fed into a moving average model as inputs. It is not a directly observable parameter like the others (and it is not fixed because it varies with the other parameters in the model). At a high level, feeding the model’s mistakes back to itself helps it get closer to the correct answer (the actual Y values) [28].

## 6 Results

### 6.1 AQI Prediction using Supervised Machine Learning Models

In order to predict the air quality index, we first conducted feature selection which was presented beforehand; now, using the latest set of features, we will predict the air quality index. This paper has incorporated machine learning models like linear regression, random forest regression, gradient boost regression, XGBoost regressor, decision tree, adaptive boosting. These regressors were utilised to predict the AQI.

For each of the models mentioned above, hyperparameter tuning (random search and grid search) has been performed to increase each model's precision individually. On the basis of root mean square error (RMSE), mean absolute error (MAE) and their R<sup>2</sup> value, a comparative study is performed in order to figure out the most suitable fitting model.

In all the algorithms below, whenever any algorithm uses any random state, it is assigned randomly to 42 to stop the minor changes in accuracy at every run, as the randomness of accuracy is halted due to this, so it also helps in performing better hyperparameter tuning. The dataset is partitioned into 75% for training and 25% for testing; all the graphs obtained below are on 25% testing data.

#### 1. Linear Regression

Here, we got 93.8% accuracy for Adarsh Nagar, 93.5% for Police Commissionerate and 93.2% for Shastri Nagar, as shown in Table 1 on default parameters. Figure 6 shows the difference between the actual values and predicted values for all three locations with the 25% testing data using the linear regression model.

#### 2. Decision Tree

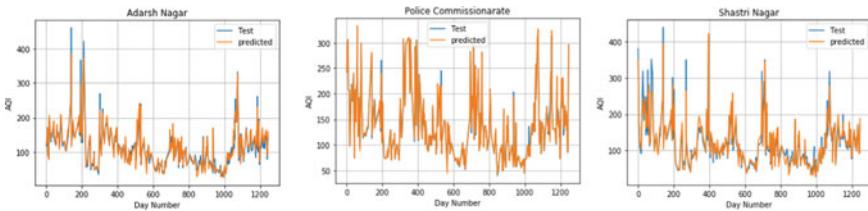
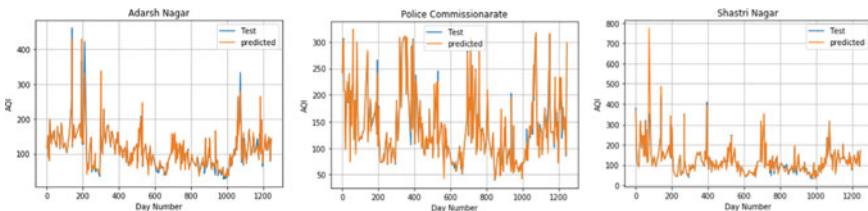
In decision tree, the results were obtained as shown in Table 1 with the value of *min\_samples\_split* as 4, *max\_depth* as 11, *min\_samples\_leaf* as 1, *max\_features* as 'auto' and other parameters were default for Adarsh Nagar, which results in 96.88% accuracy. For Police Commissionerate, 98.34% accuracy was the best found with the value of *min\_samples\_split* as 6, *max\_depth* as 10, *min\_samples\_leaf* as 2, *max\_features* as 'auto'. At last, for Shastri Nagar, 85.7% accuracy is the best obtained with default parameters. Figure 7 shows the difference between the actual values and predicted values.

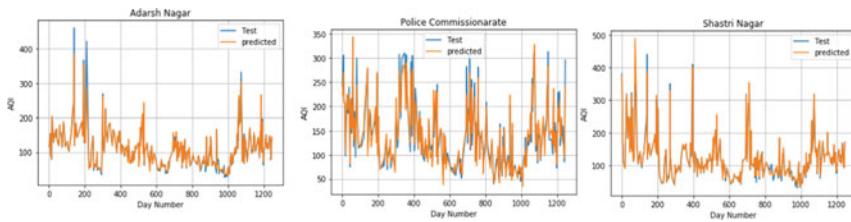
#### 3. Random Forest

In random forest, the results were obtained as shown in Table 1 with the value of *n\_estimators* as 1300, *max\_depth* as 90 and other parameters were default for Adarsh Nagar, which comes to be 96.7% accuracy. For Police Commissionerate, 98.86% accuracy was the best found, with the same values of hyperparameters as of Adarsh Nagar. At last, for Shastri Nagar, accuracy comes to be 98% with the values of *n\_estimators* as 800, *max\_depth* as 100 and other parameters as default. Figure 8 shows the difference between the actual values and predicted values.

**Table 1** Value of RMSE, MAE and  $R^2$  of all three sites, with and without hyperparameter tuning

Location	Model	RMSE	MAE	$R^2$	
				Training	Testing
Adarsh Nagar	Linear Regression	13.5883	9.7259	0.919260	0.938275
	Decision tree	9.6674	3.7404	0.995443	0.968757
	Random forest	9.9280	2.9073	0.995015	0.967007
	AdaBoost	12.0656	8.6528	0.958865	0.951334
	Gradient boosting	6.0103	3.0862	0.999690	0.987923
	XGBoost	6.1204	3.2078	0.999941	0.987478
Police Commissionerate	Linear Regression	16.5827	12.4566	0.941260	0.934969
	Decision tree	8.3714	2.9726	0.996090	0.983427
	Random forest	6.9518	2.3873	0.997454	0.988571
	AdaBoost	14.4894	10.3825	0.961580	0.950352
	Gradient boosting	6.5540	3.1931	0.999942	0.989842
	XGBoost	6.1289	3.3132	0.999999	0.991117
Shastri Nagar	Linear Regression	17.0264	11.2731	0.925347	0.932492
	Decision tree	24.7806	4.0424	1.000000	0.857002
	Random forest	9.2676	2.5952	0.990552	0.980000
	AdaBoost	18.4651	16.1773	0.930116	0.920603
	Gradient boosting	8.7657	3.2968	1.000000	0.982107
	XGBoost	4.8124	3.1459	0.999929	0.994607

**Fig. 6** Time series plot of predicted and actual AQI versus day number using linear regression**Fig. 7** Time series plot of predicted and actual AQI versus day number using decision tree



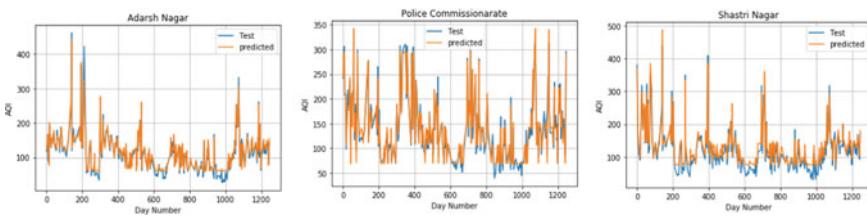
**Fig. 8** Time series plot of predicted and actual AQI versus day number using random forest

#### 4. Adaptive Boosting (AdaBoost)

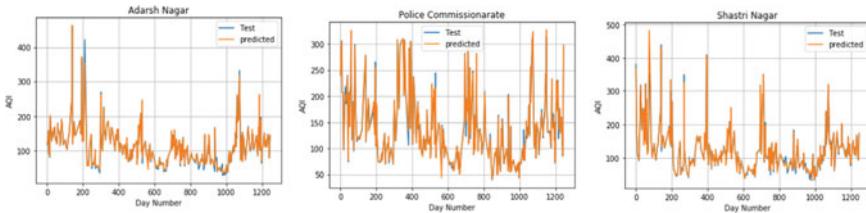
In adaptive boosting, the results were obtained as shown in Table 1 with the value of *n\_estimators* as 57, *loss* as ‘linear’, *learning rate* as 0.3582 and other parameters as the default for Adarsh Nagar, which results in 95.13% accuracy. For Police Commisionerate, 95.04% accuracy was the best found with the value of *n\_estimators* as 69, *loss* as ‘exponential’ and *learning rate* as 0.0531. At last, for Shastri Nagar, accuracy comes to be 92.06% with the values of *n\_estimators* as 68, *loss* as ‘square’, *learning rate* as 0.5403 and other parameters as default. Figure 9 shows the difference between the actual values and predicted values.

#### 5. Gradient Boosting

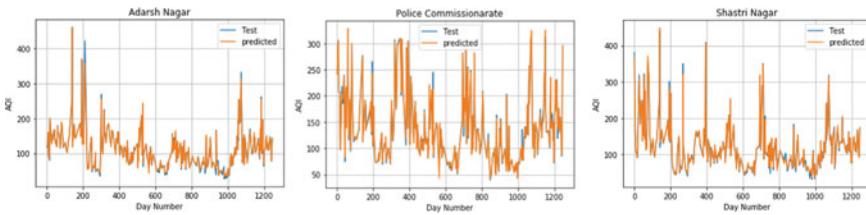
In gradient boosting, the results were obtained as shown in Table 1 with the value of *n\_estimators* as 4500, *max\_depth* as 3, *learning rate* as 0.01 and other parameters were default for Adarsh Nagar, which results in 98.79% accuracy. For Police Commisionerate, 98.98% accuracy was the best found with the value of *n\_estimators* as 800, *max\_depth* as 3 and *learning rate* as 0.1. At last, for Shastri Nagar, accuracy comes to be 98.21% with the values of *n\_estimators* as 2500, *max\_depth* as 3, *learning rate* as 0.1 and other parameters as default. Figure 10 shows the difference between the actual values and predicted values.



**Fig. 9** Time series plot of predicted and actual AQI versus day number using adaptive boosting



**Fig. 10** Time series plot of predicted and actual AQI versus day number using gradient boosting



**Fig. 11** Time series plot of predicted and actual AQI versus day number using XGBoost

## 6. XGBoost

In XGBoost, the results were obtained as shown in Table 1 with the value of *n\_estimators* as 2000, *max\_depth* as 3, *learning rate* as 0.05 and other parameters were default for Adarsh Nagar, which results in 98.75% accuracy. For Police Commissionerate, 99.11% accuracy was the best found with the value of *n\_estimators* as 2500, *max\_depth* as 3 and *learning rate* as 0.1. At last, for Shastri Nagar, accuracy comes to be 99.46% with the values of *n\_estimators* as 1000, *max\_depth* as 3, *learning rate* as 0.1 and other parameters as default. Figure 11 shows the difference between the actual values and predicted values.

**Comparison:** Table 1 shows all the observations combined for all three sites. RMSE, MAE and R<sup>2</sup> values of all the models for easy comparison, detailed inference can be found below for each site separately.

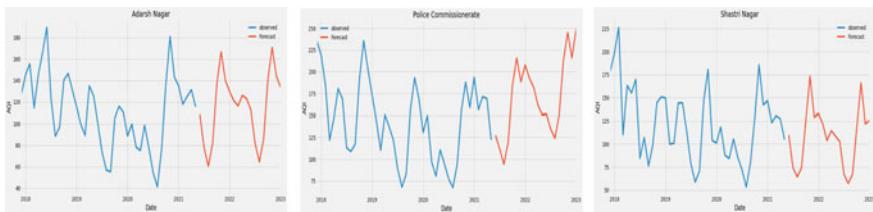
1. **Adarsh Nagar:** As we can see in Table 1, XGBoost and gradient boost regressor are the best fitting models for Adarsh Nagar's location as they have the highest accuracy (98.75% and 98.79%, respectively) amongst all other models. Whilst linear regression and AdaBoost did not perform well compared to others. Root mean squared error and mean absolute error were also least in XGBoost and gradient boost regressor.
2. **Police Commissionerate:** As we can see in Table 1, XGBoost regressor is the best fitting model for the location Police Commissionerate with accuracy.
3. **Shastri Nagar:** As we can see in Table 1, XGBoost is the best fitting model for the location Shastri Nagar with an accuracy of 99.46%; the decision tree performed very poorly, AdaBoost and linear regression did not perform well compared to other models whilst the others performed average concerning other models.

## 6.2 Forecasting using SARIMA

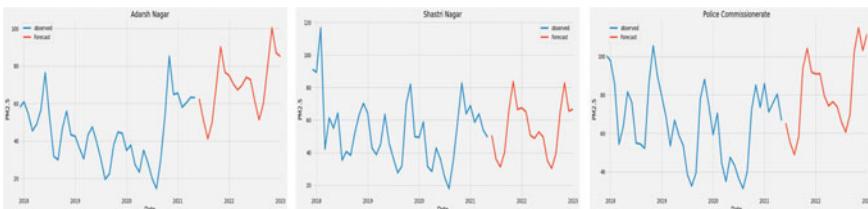
This paper uses SARIMA as the data have seasonality variations which is best tackled by the SARIMA model. As discussed earlier in the article, ‘S’ in the SARIMA means seasonality; in this model, seasonality factor is integrated in the ARIMA model.

In this section, the concentration of particulate matter (PM2.5 and PM10) and the value of the air quality index in Adarsh Nagar, Police Commissionerate and Shastri Nagar is being analysed during the period Dec-2017 to May-2021. The values of AQI and particulate matter from 2019–05–01 to 2021–05–02 are predicted after training the model. Figures 12, 13 and 14 show the previous data from Dec-2017 until May-2021. The Y-axis represents the concentration of PM2.5, whilst the X-axis represents the date.

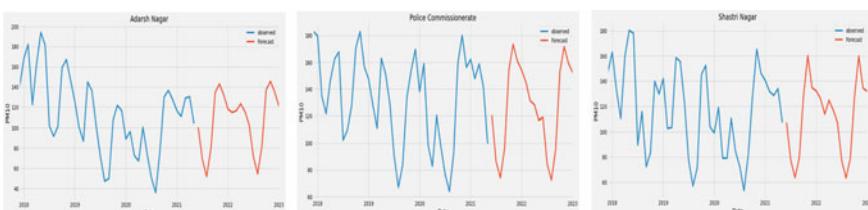
For Adarsh Nagar, mean squared error comes out to be of 554.33, and the AQI is forecasted as shown in Fig. 12. In 2021–22, AQI reaches a maximum of 166.716 in



**Fig. 12** Observed and forecasted value of AQI for the next 1.5 years



**Fig. 13** Observed and forecasted value of PM2.5 for the next 1.5 years



**Fig. 14** Observed and forecasted value of PM10 for the next 1.5 years

the month of November and during the year 2022–23, its maximum value reaches 170.750 which is greater than last year's air quality index. In Police Commissionerate, mean squared error comes out to be 686.09, and AQI is forecasted as shown in Fig. 12. During 2021–2022, AQI reaches a maximum of 215.621 in the month of November and on the 1st day of 2023, its maximum value reaches 250. For Shastri Nagar, mean squared error is 979.78, and AQI is forecasted as shown in Fig. 12. During 2021–2022, AQI reaches a maximum of 173.083 and its maximum value reaches 125.633 in the beginning of 2023.

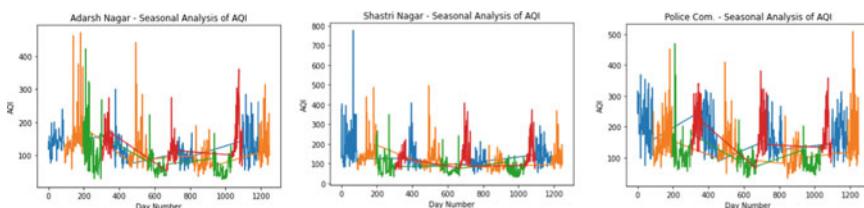
### 6.3 Seasonal Analysis

The seasons are divided into four types: summer, monsoon, post-monsoon and winter. In Fig. 15 below, the winter season falls from December to February, represented by blue, summer from March to mid-June represented by orange, rainfall from mid-June to September depicted by green and post-monsoon from October to December represented by red.

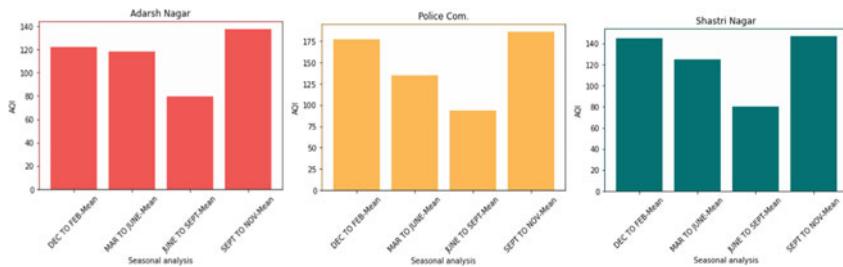
The AQI values have been plotted based on the above specified pre-decided seasons. The following graph given in Fig. 15 depicts the AQI on the Y-axis and days (starting from Dec-2017) on the X-axis. According to the seasonal analysis of all three sites, a clear trend in AQI change was observed in the season. A deeper analysis, correlation analysis and several other observations have been done (Figs. 16, 17, 18 and 19).

The seasonal study from Fig. 16 on the variation of levels of pollutants and AQI suggests that the AQI is at its worst in the post-monsoon season; this is primarily due to the fact that Diwali celebrations happens during this period. AQI is at its most profound levels during the monsoon season; that is because of the rainfall, which diminishes the amount of particulate matter and several other water-soluble pollutants. AQI levels at all three sites follow the same pattern when it comes to seasonal variations.

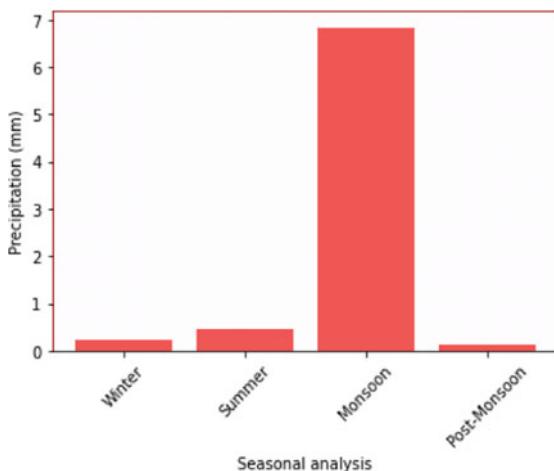
$$\text{AQI\_MONSOON} < \text{AQI\_SUMMERS} < \text{AQI\_WINTERS} < \text{AQI\_POST-MONSOON}.$$



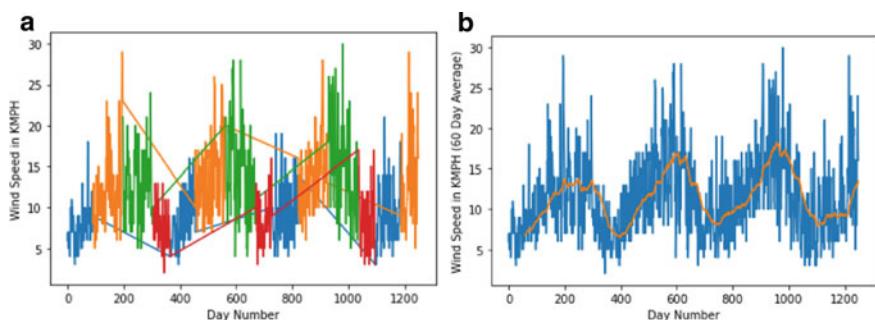
**Fig. 15** Season-wise time series (day number versus AQI plot) for Adarsh Nagar, Shastri Nagar and Police Commissionerate



**Fig. 16** Mean values of AQI (season wise) for Adarsh Nagar, Shastri Nagar and Police Commissionerate

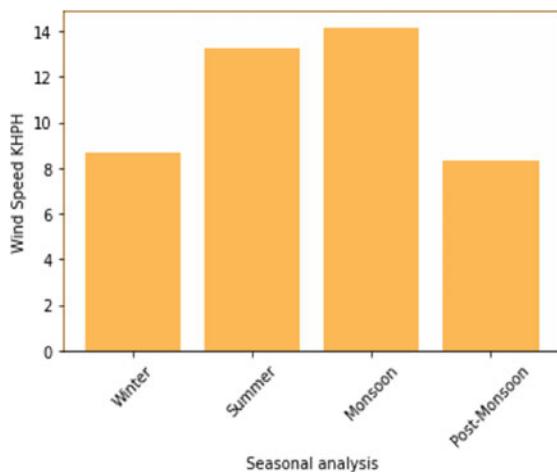


**Fig. 17** Precipitation in Jaipur at Adarsh Nagar, Shastri Nagar and Police Commissionerate



**Fig. 18** **a** Time series day number versus wind speed. **b.** Time series day number versus wind speed (60 Day rolling average)

**Fig. 19** Wind speed in kmph versus seasons at Jaipur



To support the fact that rainfall is the most prominent factor for the drastic reduction in AQI values during monsoon season, we have taken season-wise average precipitation levels, as can be seen in Fig. 17, and the precipitation levels of all four seasons show favourable behaviour. More rainfall follows up to lower AQI levels.

The post-monsoon season receives the least rainfall and suffers from maximum average AQI, and winters receive a little more rainfall, and AQI levels in winters are a little less than that of the post-monsoon season. Furthermore, summers endure a little more rainfall than winters, and AQI is second-best in summers, and for monsoon, AQI is least when precipitation is maximum.

Just like precipitation levels, wind speed is also suspected to influence the AQI levels on the seasonal scale. Since the wind speed observations were almost similar for all three sites, we will be discussing the Adarsh Nagar site here, and the results are depicted in Fig. 18a, b and Fig. 19.

Figure 18 shows the 60 days moving average of the wind speeds, which is used to show better and smoother curves. Similar observations can be drawn from Fig. 19, which holds mean values season wise

According to the observations from Fig. 19, higher average wind speeds relate to lower AQI values (Fig. 17) at the season scale. However, when observed on the day scale, wind speed does not affect the AQI quite much. On a day scale, Pearson correlation implies a negative correlation of -0.35 to -0.48 between AQI and wind speed, based on the site location.

Detailed season wise—Site wise correlation interpretation can be found in subsequent analysis. These correlation matrices use 30 data columns, including pollutants, weather, COVID-related data and other climatic variables as well. For simplicity and representation purposes, only essential features (according to the feature importance analysis done during research) are discussed below in the given heatmaps.

**Verdicts of Seasonal Analysis:** Whilst doing a comparable correlation analysis on the complete dataset, it was observed that AQI is highly correlated with PM2.5

and PM10. AQI has also been discovered considerably correlated with atmospheric humidity and percentage cloud cover. Monsoon season is fundamentally responsible for this observation. Correlation analysis also confirmed that NOx, NO2, NH3 and NOx are also amongst the other primary pollutants. The conclusions match with the standards used by the Indian Government.

Seasonal analysis of Police Commissionerate insinuates that AQI is highly correlated with PM2.5 and PM10 concentrations, and PM2.5 is more contributing than PM10. This interpretation also suggests that AQI is considerably correlated with atmospheric humidity irrespective of the season for this specific area. It also strengthened that NOx, NO2 and CO are also amongst the other primary pollutants. However, we do not have NH3 as a major pollutant here.

Seasonal analysis of the Adarsh Nagar area suggests similar insights about AQI, just like Police Com. All other observations are similar to Police Com. site except the fact that in this site, NH3 has a significant effect on the AQI. Also, benzene, xylene and carbon-based pollutants are significant in Adarsh Nagar.

For Shastri Nagar, even though AQI shows a similar pattern, other pollutants are not much correlated with AQI. Just particulate matter and humidity seem to be affecting air quality in Shastri Nagar.

Non-pollutant parameters like humidity are studied along with AQI just to see any possible relationship between air quality and these parameters.

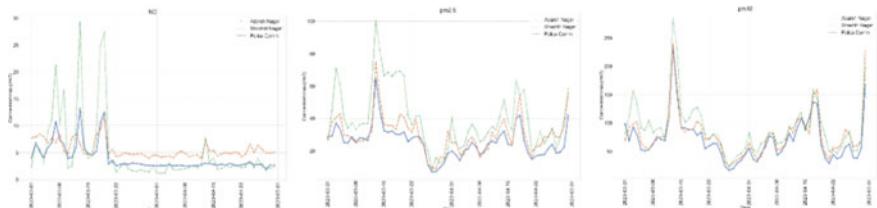
## 6.4 Lockdown Analysis

COVID-19 is a major disaster that caused havoc all over the world. The first case of COVID-19 was seen in China in December 2019 [8]. Gradually, all other nations get affected by this deadly virus. Considering the spread of the virus, the Government of India issued an advisory to screen travellers coming from other countries. Soon, the Indian Prime Minister announced Janta curfew on 22 Mar and soon a complete lockdown from 24 March 2020 to 14 April 2020 (for 21 days). This decision impacted major cities like Jaipur, Delhi and Mumbai, where many workers started moving from one place to another.

During the lockdown period, only the essential vehicles were moving. It caused less emission of harmful gases in the air [29]. The effect of lockdown on the air quality has been addressed by comparing the concentration of gases during the lockdown phase to the pre- and post-lockdown phase.

In Police Commissionerate, the concentration of PM2.5 before the lockdown was higher than Adarsh Nagar and Shashtri Nagar, whilst in the lockdown period, the same trend is followed in all three locations. A similar trend is observed in the air quality index, which is correlated with the PM2.5 and other gases (Fig. 20).

Another particulate matter PM10, which was higher in pre-lockdown period drops in the lockdown phase. The same trend is followed in other cities too, and changes observed were also related to PM2.5.



**Fig. 20** Change in concentration of PM2.5, PM10, NO, NO<sub>2</sub>, NOx and CO in the lockdown phase at Adarsh Nagar, Shashtri Nagar and Police Commissionerate

The concentration of NO, NO<sub>2</sub> and NOx starts decreasing when the lockdown was imposed. Most of the NOx comes from the transport sector, as the transport facilities were restricted, causing NOx to decrease during the lockdown period. Similarly, NO<sub>2</sub> also decreases during the lockdown period. The concentration of NO, NO<sub>2</sub> at Police Commissionerate was higher before the lockdown period than other locations; this is because Police Commissionerate falls nearby many roads, and Ajmer-Jaipur Road also passes nearby causing the location vulnerable to gases emitted from the vehicles moving through highways. Carbon monoxide (CO) was also higher before the lockdown period, but during the lockdown, its concentration decreased. When the complete lockdown ended, concentration of air pollutants started increasing due to more movement of vehicles and thus more pollution in the air. As AQI is dependent on the concentration of these gases, AQI also started increasing when the government provided some relaxation after the Mar–Apr lockdown period.

## 7 Discussion

This research suggests that amongst various supervised machine learning models, XGBoost regressor performs the most reliable out of the models, this article covers for all three locations, and gradient boosting also works considerably well. Using these models, AQI is predicted when inputs of various pollutants are available.

Later, SARIMA is being used to forecast the AQI and some significant pollutants. So that pre-counter measures can be taken in order to reduce air pollution. In Adarsh Nagar, the AQI will increase in upcoming years and in the Police Commissionerate area, which will be more prominent. In the case of Shastri Nagar, the air quality index will be decreasing in upcoming years, as seen in the results of SARIMA.

It is observed that the air quality index (AQI) in the lockdown period was relatively much lower than before the lockdown was imposed, and since the lockdown period ended, AQI started increasing and has been increasing ever since. The precautions taken by the Indian Government resulted in less emission of polluting gases and particulate matter in the lockdown period. Results show that the emission of harmful gases from vehicles and industries decides the air quality, and hence, it is crucial to take necessary measures to reduce air pollution.

Seasonal analysis done in this research explored the effects on the air quality due to precipitation and wind speed. The research concludes that both wind speed and precipitation are inversely related to the AQI, and the effect is considerably enormous. The seasonal analysis also analysed the data by breaking it down into four seasons, namely: winter, summer, monsoon and post-monsoon. For every season, Pearson correlation analysis was done to check for any possible season-specific result. It was followed by the weekday and holidays analysis, which was targeted to analyse the level of AQI based on weekdays and holidays. It concluded that Jaipur sees the lowest AQI level on Tuesdays, and median AQI values on holidays were relatively lower than median AQI values on regular days.

## 8 Conclusion

AQI is an essential factor in determining the air quality, and this analysis shows that the air quality had improved during the lockdown period and caused less emission of harmful gases in this period. This research shows that pollution will increase in the future. Whilst the pollution decreased during the lockdown, it started increasing later. The leading cause for the degradation of air quality in Jaipur is transportation. Considering this, it is suggested that electric vehicles should be adopted and promoted by the government. New policies which incentivise the adoption of e-vehicle will tremendously help in maintaining healthy air in Jaipur. Furthermore, incorporating an advanced public transport network in highly populated regions of Jaipur will also play a key role in curbing high levels of AQI.

## References

1. Al-Salem, S.M., Bouhamrah, W.S.: Ambient concentrations of benzene and other VOCs at typical industrial sites in Kuwait and their cancer risk assessment. *Res. J. Chem. Environ.* **10**, 4246 (2006)
2. Gupta, M.C., Ghose, A.K.M.: The effect of coal smoke pollutants on the leaf epidermal architecture in Solanummolengena variety pusapurble long. *J. Environ. Pollut.* **41**(4), 315–321 (1986)
3. EPA (Environmental Protection Agency), Measuring Air Quality: The Pollutant Standards Index. EPA 451/K-94-001 (1994)
4. Thom, G.C., Ott, W.R.: A proposed uniform air pollution index. *Atmos. Environ.* **10**, 261–264 (1976)
5. Rizwan, S.A., Nongkynrih, B., Gupta, S.K.: Air pollution in Delhi: its magnitude and effects on health. *Indian J Community Med* **38**(1), 4–8 (2013)
6. Nagpure, Gurjar, B., Martel, J.: "Human health risks in national capital territory of Delhi due to air pollution". *Atmos. Pollut. Res.*, **5**(3), 371–380, (2014)
7. Aggarwal, P., Jain, S.: Impact of air pollutants from surface transport sources on human health: a modeling and epidemiological approach. *Environ. Int.* **83**, 146–157 (2015)

8. WHO (2020) World Health Organization. Novel coronavirus (2019-nCoV). [Online]. Available: [http://www.euro.who.int/en/health-topics/health-emergencies/novel-coronavirus-2019ncov\\_old](http://www.euro.who.int/en/health-topics/health-emergencies/novel-coronavirus-2019ncov_old). Accessed 02 Apr 2020
9. Sharma, S., Zhang, M., Gao, J., Zhang, H., Kota, S.H.: Effect of restricted emissions during COVID-19 on air quality in India. *Sci. Total Environ.* **728**, 138878 (2020)
10. Navinya, C., Patidar, G., Phuleria, H.C.: Examining effects of the COVID-19 national lockdown on ambient air quality across urban India. *Aerosol Air Qual. Res.* (2020). <https://doi.org/10.4209/aaqr.2020.05.0256>
11. Srivastava, S., Kumar, A., Bauddh, K., Gautam, A.S., Kumar, S.: 21-day lockdown in India dramatically reduced air pollution indices in Lucknow and New Delhi, India. *Bull. Environ. Contam. Toxicol.* (2020)
12. Central Pollution Control Board, Ministry of Environment, Forests and Climate Change [Online]. Available: <http://www.cpcb.nic.in/>
13. AQI Table. [Online] (2021). Available: [https://app.cpcbccr.com/AQI\\_India/](https://app.cpcbccr.com/AQI_India/)
14. National Air Quality Index: (2015). [Online]. Available: [https://app.cpcbccr.com/CCR\\_docs/FINAL\\_REPORT\\_AQI\\_pdf](https://app.cpcbccr.com/CCR_docs/FINAL_REPORT_AQI_pdf)
15. AQI Standards (2009). [Online]. Available: [https://cpeb.nic.in/uploads/National\\_Ambient\\_Air\\_Quality\\_Standards.pdf](https://cpeb.nic.in/uploads/National_Ambient_Air_Quality_Standards.pdf)
16. Garcia, J.M., Teodoro, F., Cerdeira, R., Coelho, R.M., Kumar, P., Carvalho, M.G.: Developing a methodology to predict PM10 concentrations in Urban areas using generalized linear models. *Environ. Technol.* **37**, 2316–2325 (2016)
17. Park, S., Kim, M., Kim, M., Namgung, H.-G., Kim, K.-T., Cho, K.H., H, K., Kwon, S.-B.: Predicting PM10 concentration in seoul metropolitan subway stations using artificial neural network (ANN). *J. Hazard. Mater.* **341**, 75–82 (2018)
18. Zheng, Y., et al.: “Forecasting fine-grained air quality based on big data.” In KDD ‘15, Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
19. Kisi, O., Kulwinder, S., Kirti, S.: Modeling of air pollutants using least square support vector regression, multivariate adaptive regression spline, and M5 model tree models. *Air Qual. Atmos. Health* **10**(7), 873–883 (2017)
20. Yu, R., Yang, Y., Yang, L., Han, G., Move, O.A.: RAQ a random forest approach for predicting air quality in urban sensing systems. *Sensors* **16**, 86 (2016)
21. Yi, X., Zhang, J., Wang, Z., Li, T., Zheng, Y.: Deep distributed fusion network for air quality prediction. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, London, UK (2018)
22. Veljanovska, K., Dimoski, A.: Air quality index prediction using simple machine learning algorithms. *Int. J. Emerg. Trends Technol. Comput. Sci.* **7**, 25–30 (2018)
23. Kumari, P., Toshniwal, D.: Impact of lockdown measures during COVID-19 on air quality a case study of India. *Int. J. Environ. Health Res.* 1–8 (2020)
24. World Weather Online [Online]. Available: <https://www.worldweatheronline.com/>
25. Wolf, L., Shashua, A.: Feature selection for unsupervised and supervised inference: the emergence of sparsity in a weight-based approach, *The. J. Mach. Learn. Res.* **6**, 1855–1887 (2005)
26. Zhao, Z., Liu, H.: Spectral feature selection for supervised and unsupervised learning. In Proceedings of the 24th international conference on Machine learning, ACM, pp. 1151–1157 (2007)
27. Paris, G., Robilliard, D., Fonlupt, C.: Exploring overfitting in genetic programming. In Artificial Evolution, International Conference, Evolution Artificielle, Ea 2003, Marseilles, France, October. DBLP, pp. 267–277 (2004)
28. Khandelwal, I., Adhikari, R., Verma, G.: Time series forecasting using hybrid ARIMA and ANN models based on DWT decomposition. *Procedia Comput. Sci.* **48** (2014). <https://doi.org/10.1016/j.procs.2015.04.167>
29. Mahato, S., Pal, S., Ghosh, K.G.: Effect of lockdown amid COVID-19 pandemic on air quality of the megacity Delhi, India. *Sci. Total Environ.* 139086 (2020)

# Accurate Segmentation of Lung Nodule using Adaptive Weights as Feature for Recurrent Neural Network



R. Janefer Beula and A. Boyed Wesley

**Abstract** This paper proposes a lung nodule segmentation algorithm that uses adaptive weights as a feature for the recurrent neural network. The algorithm initially detects the lung parenchyma from which the background region is minimized. However, the boundaries of the obtained nodule candidate region are not accurate. To optimize the boundary, few layers of boundary pixels are used to estimate the adaptive weights. The low-level adaptive weights are estimated using the signal formed by the inner and outer boundary layer pixel in zigzag order. The low-level adaptive weights are clustered into two clusters using the fuzzy C means (FCM) clustering algorithm. These clustered weights are used to estimate the exact boundaries of the nodule regions. The high-level adaptive weights are estimated throughout the nodule region are trained using a recurrent neural network that can eliminate false positives. To avoid the errors caused due to adaption, the signal is mirrored, and the features estimated on the second half are used for training the model. The evaluation was done with the LIDC datasets using the metrics such as Hausdorff distance, probability rand index (PRI), accuracy, recall, and precision. The scheme provides accuracy, recall, and precision of 94.08%, 89.3%, and 94.1%, respectively.

**Keywords** Lung nodule segmentation · Recurrent neural network · Fuzzy means clustering

## 1 Introduction

A statistics by global cancer (2015) tells that every year around 19.5% of deaths are due to cancer-related diseases. To prevent death due to lung cancer, it is necessary to provide proper treatment at its early stage. Therefore, the exact diagnosis of lung nodule is essential at its early stage. Detection of lung nodule at its early stage has difficulty such as small size and low contrast. The presence of low contrast nodules

---

R. J. Beula (✉) · A. B. Wesley

Department of Computer Science, Nesamony Memorial Christian College Affiliated To  
Manonmaniam Sundaranar University, Tirunelveli, India  
e-mail: [janeferbeula@gmail.com](mailto:janeferbeula@gmail.com)

leads to the wrong estimation of nodule boundaries, i.e. the nodule may be segmented to the background, or the non-nodule region may be segmented to the foreground. It is still challenging to segment the nodules that lie close to the lung wall (Juxta vascular) or blood vessels (Juxta pleura) since the intensity of the lung wall or blood vessels lies close to the lung nodules. The contribution of the paper is as follows:

- (i) The proposed algorithm initially pre-processes the image and eliminates the background region.
- (ii) The estimated nodule region does not have exact boundaries and also contain false positives. To estimate the accurate boundaries, few inner and outer layers are used to estimate adaptive weights estimated in zigzag order. The weights are clustered into two parts using the FCM clustering algorithm that represents the nodule and non-nodule boundary pixels accurately.
- (iii) The false positives are further minimized using high-level adaptive weights that are trained using the recurrent neural network algorithm.

The paper is arranged as follows. Section 2 shows a few of the related works, and the proposed system is provided in Sect. 3. The experimental results and the conclusion are provided in Sects. 4 and 5, respectively.

## 2 Related Works

Schemes such as morphological processing [1], energy optimization [2], region growing [3], and statistical learning methods [4] are used to extract the region of interest (ROI) from the background. For the classification of nodules and non-nodules, the features are extracted from the scan images. The commonly used feature extraction algorithms include spherical harmonics [5], Fourier shape feature [6], histogram of oriented gradients (HOG) [7], local binary pattern (LBP) [[8]], and grayscale co-occurrence matrix (GLCM)[9]. Suchetha et al. [10] proposed a novel deep learning architecture to classify segmented fundus images. Xie et al. [11] proposed a knowledge-based collaborative deep learning. This approach can classify the malignant and benign nodules. The authors Wu et al. [12] used a dual branch network along with coarse to fine segmentation. This approach also focuses on the boundaries but the computational complexity is higher.

The authors Cao et al. [13] used a two-stage CNN. In the first stage, CNN architecture is used whilst in the second stage, UNet is used. Ensemble learning is used for the reduction of false positives. Deep learning with multiple strategies was proposed [14], where the authors used both X-ray and CT images for validation. However, the model works well in CT images but the performance is less for X-ray images.

The authors Manoharan et al. [15] proposed a graph-cut algorithm that can detect the soft tissues in the lung nodules. The authors further [16] used the probability of malignancy calculation. The algorithm uses a neural classifier and nodule selection for the reduction of false positives. The algorithm was tested on a limited number of images. The authors Sathis et al. [17] proposed an adaptive shape-based interactive

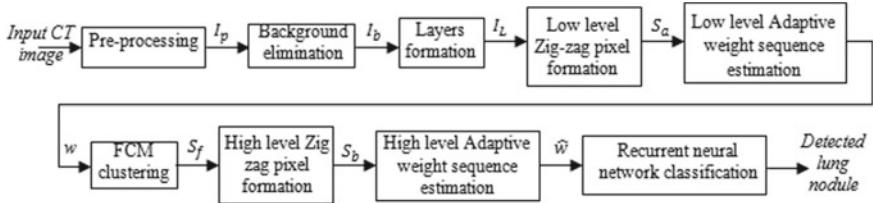
scheme for segmenting the nodules. This approach improves the speed and energy function of the watershed segmentation approach. The authors only focussed on the energy function and speed. The convolution neural networks [18] were analyzed by the authors Neha et al. where different types of neural networks like AlexNets, ResNet 50, and GoogLeNet. The networks ResNet 50 and GoogLeNet provide better results compared to AlexNet. The computational complexity is found to be higher. The author Ivan et al. [19] proposed a transfer learning with RetinaNet for the detection and classification of lung nodules from the thorax CT scan images. The authors Ying Su et al. [20] used faster R-CNN for the detection of a lung nodule. The authors used two models, namely VGG16 and ZF models. The ZF model shows a better performance than the VGG16 model. The scheme provides a detection accuracy of 91.2%. However, the scheme fails to eliminate the false positives that lie close to the lung wall. The usage of the cascaded network increases the complexity of the algorithm. Transfer learning [20] is also used to detect the lung nodule, where two training strategies such as freezing layers and fine-tuning are used. The shape and statistical-based features are also used for detecting the lung nodules. The lung locale was isolated using the histogram of the grayscale, and the refining was performed using the morphological operations. The scheme has a lower accuracy, since the shape of nodules and non-nodules is similar in several cases. The optimal fuzzy model [21] was used to detect the lung nodule where the authors used the FCM for segmentation and selecting the features. The authors Baker et al. [22] used a three-stage algorithm for detecting the cancerous lung nodule. A modified LoG algorithm is used to enhance the suspicious regions. The potential cancerous nodule is detected in the second stage, and in the third stage, five texture features are detected. The algorithm is evaluated using only 60 cases.

The performance of the existing scheme is low if the nodule is at the early stage or if the nodules lie close to the lung nodules or blood vessels. Therefore, the proposed work aims to segment the accurate detection of nodule boundaries and estimate of best adaptive features that can highly differentiate the nodule from the non-nodule regions. The next section shows the proposed algorithm.

### 3 Proposed Algorithm

The block diagram of the proposed lung nodule segmentation algorithm is depicted in Fig. 1. The proposed algorithm has the process such as pre-processing, background elimination, layers formation, low-level zigzag pixel formation, low-level adaptive weight sequence estimation, FCM clustering, high-level zigzag pixel formation, high-level adaptive weight sequence estimation, and recurrent neural network classification. The proposed algorithm has stages such as

- (i) Pre-processing and background elimination that aims to eliminate the background regions other than the nodule candidate.



**Fig. 1** Block diagram of proposed lung nodule segmentation algorithm

- (ii) Low-level weight sequence estimation and FCM clustering that aims to find the exact boundaries of the nodule candidates.
- (iii) High-level weight sequence feature estimation that different the nodule and non-nodule candidates.
- (iv) The recurrent neural network classifier to classify the nodule candidate from the non-nodules.

### 3.1 Pre-processing

The pre-processing includes the process such as scaling to a size of  $U \times V$ , median filtering followed by adaptive histogram equalization. The median filtering is used to remove the noise that gets induced during image acquisition, and adaptive histogram equalization is used to remove the non-uniformities in illumination. Let  $I_p$  be the pre-processed CT image.

### 3.2 Background Elimination

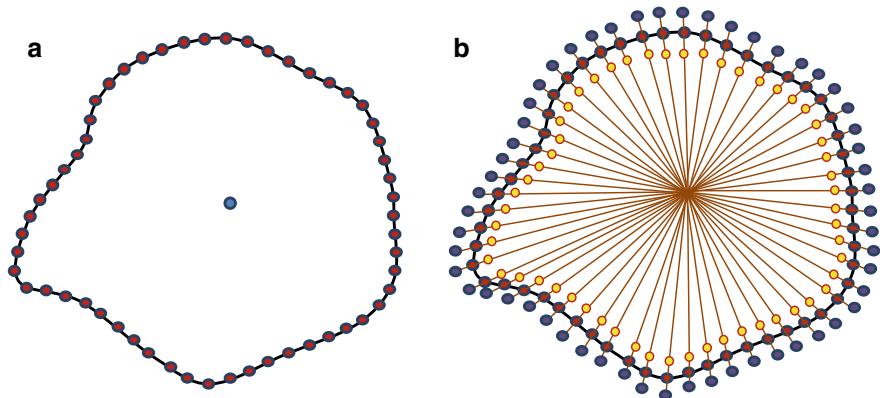
The background elimination process aims to eliminate the background region other than the nodule candidate regions. The background elimination was performed using a multi-level thresholding algorithm. This includes the segmentation of lung parenchyma followed by elimination of the non-nodule region. Let  $I_b$  be the image in which the background is eliminated, which contains the nodule candidate regions. Let  $I_b$  be a single nodule candidate region.

### 3.3 Layers Formation

The boundaries of the nodule candidate region do not contain accurate boundaries. It may also contain the pixels of the non-nodule region at the boundaries. It may also miss the boundaries of the nodule region to its background. Therefore,  $N$  number

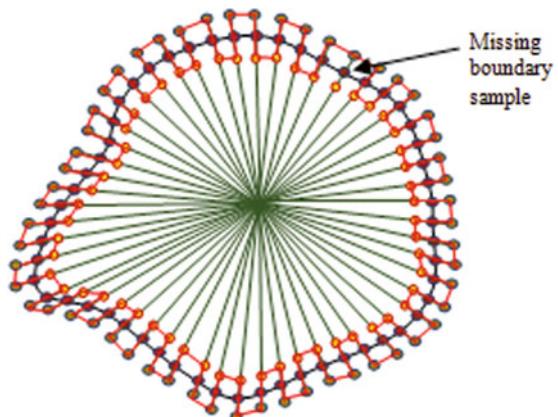
of inner and outer layers are chosen on the region  $I_b$  for the exact estimation of boundary pixels. Figure 2 depicts an example for nodule candidate and inner, outer layer formation. Let the boundary layer pixels be represented as  $\text{HF } l_b$ . The  $N$  number of inner and outer layers pixels be represented as  $l_{i,n}$  and  $l_{o,n}$ , where  $n = 1, 2, \dots, N$ . Therefore, the layer for processing is given by

$$l_L = \{l_{i,N}, l_{i,N-1}, \dots, l_{i,2}, l_{i,1}, l_b, l_{o,1}, l_{o,2}, \dots, l_{o,N}\} \quad (1)$$



**Fig. 2** **a** Estimation of boundary layers, **a** nodule candidate, **b** inner and outer layer formation  $N = 1$

**Fig. 3** low-level zigzag pixel formation



### 3.4 Low-level Zigzag Pixel Formation

The low-level zigzag pixel formation represents that the arrangement of pixels at its layers  $\{l_{i,N}, l_{i,N-1}, \dots, l_{i,2}, l_{i,1}, l_b, l_{o,1}, l_{o,2}, \dots, l_{o,N}\}$  in a zigzag fashion as depicted in Fig. 3. Let the pixels that are arranged in zigzag order be represented as  $S_a$ . An example of low-level zigzag pixel formation is depicted in Fig. 4.

### 3.5 Low-level Adaptive Weight Sequence Estimation

From the signal  $S_a$ , the reference signal  $x(l)$  is estimated by filtering  $S_a(l)$  using the moving average filter given by

$$x(l) = \frac{1}{N} \sum_{i=0}^{N-1} S_a(l-i) \quad (2)$$

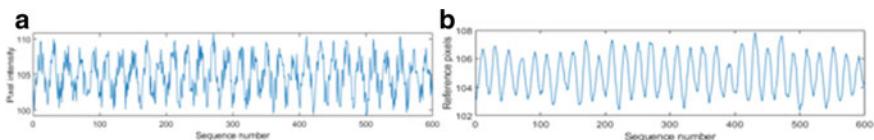
The signal  $S_a(l)$  and  $x(l)$  at the boundaries can be represented as shown in Fig. 4a, b, respectively. The signal  $S_a(l)$  and  $x(l)$  are mirrored using the equations  $S_A(l) = [S_a(l), S_a(-l)]$  and  $X(l) = [x(l), x(-l)]$ . The mirrored signal  $S_A(l)$  and  $X(l)$  be represented as shown in Fig. 5a, b. The adaptive weight feature is estimated using the equation,

$$w_k(l+1) = w_k(l) + \delta \times e(l) \times X(l) \quad (3)$$

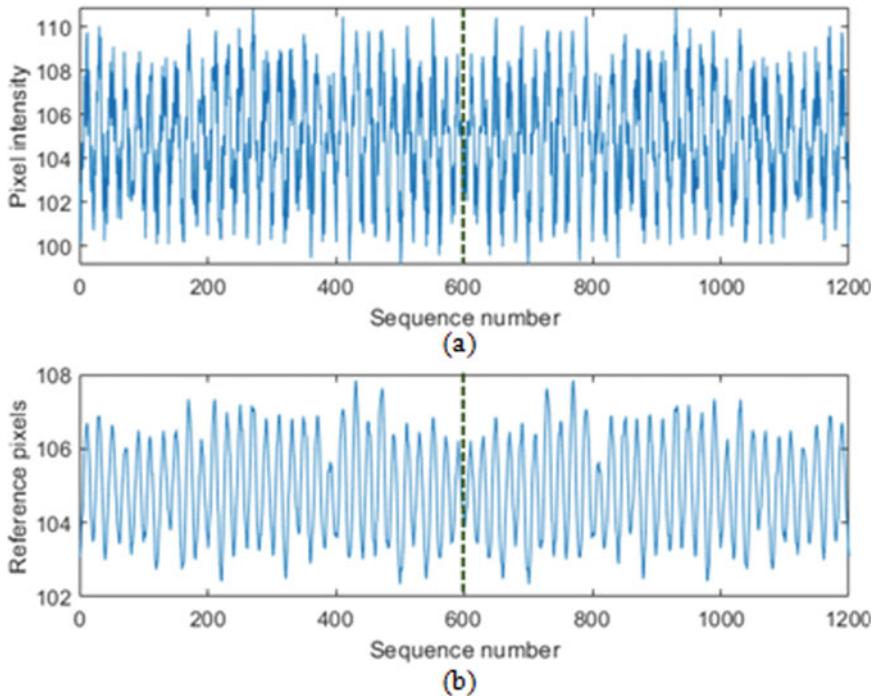
$$e(l) = S_A(l) - y(l) \quad (4)$$

$$y(l) = \sum_{k=1}^L w_k(l) X(l) \quad (5)$$

where  $L$  is the order for adaptive feature estimation,  $\delta$  is the step-size, and  $k = 1, 2, \dots, L$ . For one sample,  $L$  number of weights are estimated from which the cumulative weight is estimated as



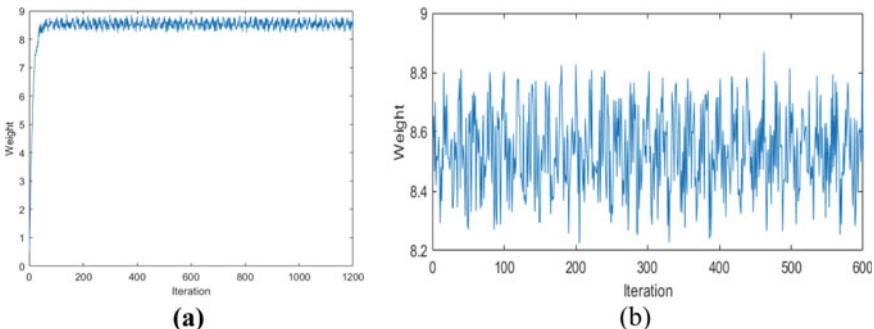
**Fig. 4** **a** Pixel sequence formed along with the layers in zigzag order  $S_a(l)$ , **b** reference signal  $x(l)$  estimated by filtering of moving average filter



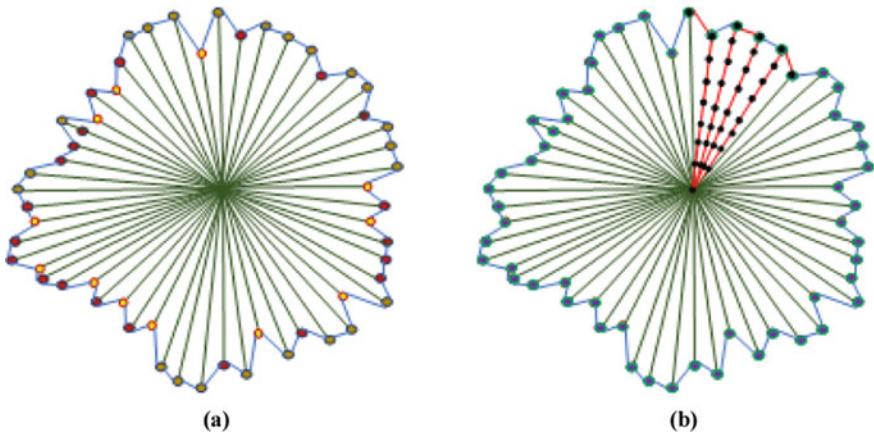
**Fig. 5** **a** Mirrored signal  $S_A(l)$ , **b** mirrored signal  $X(l)$

$$W(l) = \frac{1}{L} \sum_{k=1}^L w_k(l) \quad (6)$$

The adaptive weight feature  $W(l)$  is represented as shown in Fig. 6a. The second half of the feature gives the actual weight feature  $w(l)$  as depicted in Fig. 6b



**Fig. 6** **a** Low-level weights  $W(l)$ , **b** low-level weights  $w(l)$  for the second half



**Fig. 7** **a** Output of FCM clustering, **b** high-level zigzag pixel formation

### 3.6 FCM Clustering

The k-means algorithm works on the principle of clustering the distance between the training data and centroid with  $k$  number of clusters. In FCM clustering to partition data into the number of clusters with a minimum similarity between different clusters, k-means clustering can be best suited with clusters of more than 2. Since we aim to cluster the boundaries as a nodule and non-nodule region where there is fuzziness in the intensity, we have used the FCM algorithm for clustering. The weights  $w(l)$  that are obtained on the second half cycle are applied for the FCM clustering algorithm that can cluster the weights into two clusters. One cluster contains the weights that correspond to the pixels that belong to the background whilst another cluster contains the weights that correspond to the pixels that belong to the foreground. This gives an accurate estimation of boundaries for the nodule candidate. Based on the classified pixels, the exact boundary is formed as depicted in Fig. 7a. Let  $S_f$  represents boundary-optimized nodule.

### 3.7 High-level Zigzag Pixel Formation and High-level Adaptive Weight Feature Estimation

The high-level zigzag pixels are applied on the boundary-optimized nodule **candidates**. The difference between the low-level zigzag pixels formation and high-level pixel formation is that the  $M$  number of layers are formed inside the boundaries estimated by the FCM clustering algorithm that covers the complete nodule candidate region as shown in Fig. 7b. Let  $S_b$  represents the high-level zigzag pixel sequence formed throughout the nodule candidate and  $\hat{w}$  represent the high-level

adaptive weight feature. The high-level adaptive weight feature estimation process is the same as that of low-level adaptive weight feature estimation.

### 3.8 Recurrent Neural Network

The detected nodule candidate is classified as nodule/ non-nodule using the recurrent neural network. Let  $y_l(t)$  be the activation of processing element (PE)  $l$  at time  $t$  and  $g$  represent the transformation function, then the activation can be expressed as

$$y_l(t) = g \left( \sum_i \int_{t'}^t m_{li}(t' - t) \cdot y_i(t') dt' \right) \quad (7)$$

where  $m(\cdot)$  represents the temporal kernel. Let  $W_{li}$  represents the output of the previous time step of PE  $i$  on the current PE  $l$ , then the activation can be reformulated as

$$y_l(t) = g \left( \sum_i W_{li} y_i(t-1) \right) \quad (8)$$

For a sigmoid function

$$y_l(t) = \frac{1}{1 + e^{-(\sum_i W_{li} x_i + b_l)}} \quad (9)$$

The term  $b_l$  represents the bias of the processing element  $l$  where  $x_i$  is the input vector  $i$ th component.

The RNN has two important properties such as (i) the updation of hidden state can be done in a complicated way due to the non-linear dynamics and (ii) a lot of information about the past can be efficiently stored due to the distributed hidden states. The forget gate in an LSTM with a forward pass has the activation vector

$$F_t = g(U_f h_{t-1} + W_f x_t + b_f) \quad (10)$$

The activation vector of the input/update gate is

$$I_t = g(U_i h_{t-1} + W_i x_t + b_i) \quad (11)$$

The activation vector of the output gate is

$$I_t = g(U_i h_{t-1} + W_i x_t + b_l) \quad (11)$$

The activation vector of the output gate is

$$O_t = g(U_o h_{t-1} + W_o x_t + b_o) \quad (12)$$

The activation vector at the cell input is

$$\hat{C}_t = J(U_c h_{t-1} + W_c x_t + b_c) \quad (13)$$

The cell state vector is given by

$$C_t = f_t \odot C_{t-1} + i_t \odot \hat{C}_t \quad (14)$$

$\odot$  denotes the element-wise Hadamard product.

The hidden state vector of LSTM is

$$h_t = O_t \odot J(C_t) \quad (15)$$

Here,  $g$  is the sigmoid function and  $J$  is the hyperbolic tangent function. The architecture of the recurrent neural network is depicted in Fig. 8.

The high-level adaptive weight features  $\hat{w}$  estimated on the nodule candidates of train images are trained using the model. In the testing phase, the adaptive weight features  $\hat{w}$  on the nodule candidates of the test images are classified to detect the actual nodules.

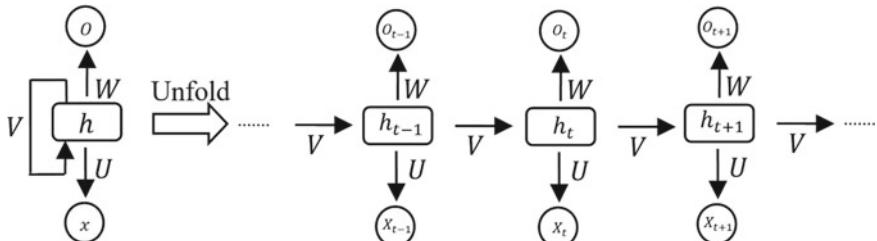
### Algorithm

**Input:** CT lung image  $I$ .

**Output:** Classification result.

**Step 1:** Pre-process the image  $I$ , to obtain the pre-processed image  $I_p$ .

**Step 2:** Eliminate the background regions of the image  $I_p$  by applying the multi-level thresholding algorithm to obtain the nodules  $I_b$ .



**Fig. 8** Architecture of recurrent neural network model

**Step 3:** Eliminate  $N$  number of the inner and outer layer  $l_{i,n}$  and  $l_{o,n}$  for the image  $l_b$ .

**Step 4:** Form the low-level zigzag pixels on the boundary layer pixels  $l_{i,n}$ ,  $l_{o,n}$ , and  $l_b$ .

**Step 5:** Estimate the low-level adaptive weight sequence  $w(l)$ .

**Step 6:** Apply FCM clustering on the weights, to estimate the accurate boundaries.

**Step 7:** Estimate the high-level adaptive weight features on the high-level zigzag pixels.

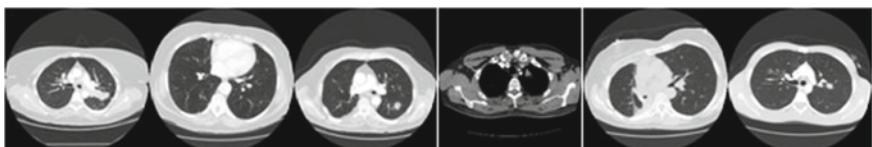
**Step 8:** Apply the high-level zigzag pixels to the trained recurrent neural network to classify the nodule candidate as nodule or non-nodule.

## 4 Experimental Results

The evaluation of the proposed algorithm was evaluated using the LIDC datasets using the MATLAB tool. Figure 9 depicts a few sample images from the LIDC dataset. The LIDC-IDRI dataset contains 1018 low-dose lungs, CTs from lung patients. The datasets contain the types of nodules such as juxta pleura, juxta vascular, and solitary nodules with the ground truth results. The datasets also contain information such as volume and location of nodule. The recurrent neural network uses the LSTM with 100 hidden units, a fully connected Softmax layer. The training was performed with a maximum number of epochs of 20 and an initial learning rate of 0.0001, for the accurate detection of candidate boundaries, the low-level zigzag pixel is formed with 5 inner layers and 5 outer layers. For the estimation of adaptive weight features, the algorithm uses the initial weight  $w_k(l) = 0.5$  and step-size  $\delta = 0.5$ . For training the model, 70% of the data is used for testing, and the remaining 30% is used for testing. The RNN uses 674 images for training and the remaining 290 images for testing.

The performance of the proposed algorithm was evaluated using the metrics such as Hausdorff distance, probability rand index, accuracy, recall, and precision. The Hausdorff distance and probability rand index are used to validate the exactness of segmentation at its boundaries which are expressed by the relations

$$\text{HD} = \max(d(g, s), d(s, g)) \quad (16)$$



**Fig. 9** Sample images from LIDC dataset

where  $d(g, s) = \max_{a \in g} \min_{b \in s} |a - b|$ .

$$\text{PRI} = \frac{S_s + S_g}{S_s + S_g + D_s + D_g} \quad (17)$$

where  $S_s + S_g$  and  $D_s + D_g$  represent the similarity and difference between ground truth and segmented result. The metrics accuracy, recall, and precision are used to measure the performance of the classification algorithm in eliminating false negatives which are expressed by

$$\text{accuracy} = \frac{t_n + t_p}{t_p + t_n + f_n + f_p} \quad (18)$$

$$\text{recall} = \frac{t_p}{t_n + t_p} \quad (19)$$

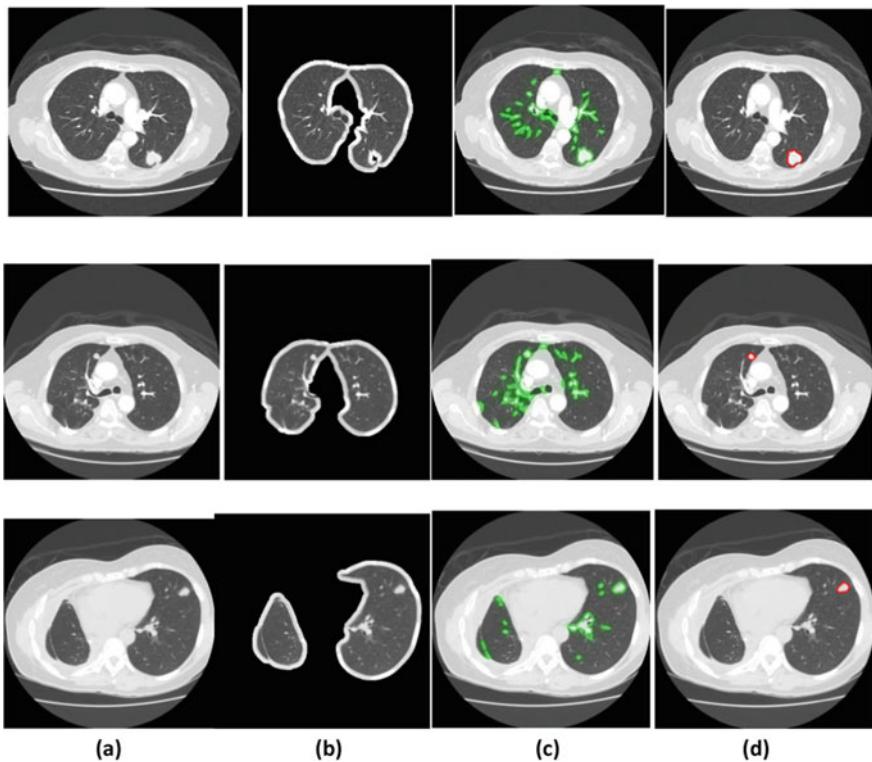
$$\text{precision} = \frac{t_p}{f_p + t_p} \quad (20)$$

Here, tn, tp, fp, and fn are the true negative, true positive, false positive, and false negatives, respectively.

Figure 10 depicts the experimental results of the proposed algorithm where Fig. 10a shows the input CT images and Fig. 10b depicts the lung parenchyma detected by background elimination. Figure 10c shows the nodule candidates detected after background elimination. This contains the actual nodule candidates along with the false positives. The false positives are eliminated by the recurrent neural network classifier as depicted in Fig. 10d. Figure 11 shows the confusion matrix obtained on the classified result of the test images.

The comparison was done with the traditional methods such as deep learning [10], two-stage CNN [13], and DBN [12]. Table 1 depicts the comparison of accuracy, recall, precision, and AUC comparison of the proposed method with the traditional methods. The proposed scheme provides an accuracy, recall, precision, and AUC of 94.08%, 89.3%, 94.1%, and 96.21%, respectively, which is higher than the traditional methods as depicted in Fig. 12. The proposed scheme has an improvement in accuracy, recall, precision, and AUC of 0.96%, 1.13%, 4.34%, and 0.08% than the DBN scheme.

Figure 12 shows the ROC comparison of the proposed method with the traditional scheme. The proposed method has a higher AUC than the traditional schemes such as DBN, two-stage CNN, and deep learning. The proposed method has an AUC of 96.21%. This shows that the proposed scheme can better differentiate the nodule from the non-nodules than other recent schemes. As the step-size  $\delta$  increases from 0.1 to 1, the PRI gradually increases and reaches a maximum of 0.941 at  $\delta = 0.5$  and gradually reduces. Similarly, the HD values gradually decrease and reach a minimum value of 1.79 at  $\delta = 0.5$  and gradually increase (Fig. 13).



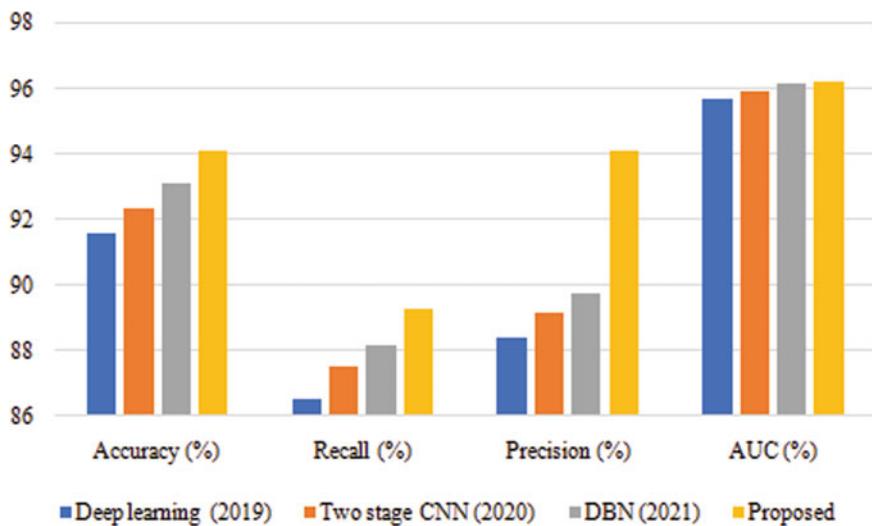
**Fig. 10** Experimental results **a** input CT image, **b** segmented lung parenchyma, **c** segmented nodule candidates **d** classified nodules

**Fig. 11** Confusion matrix obtained by the proposed scheme

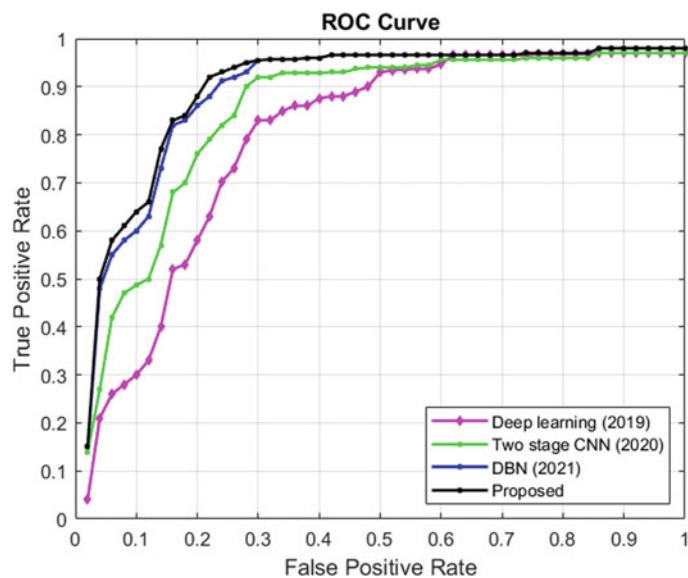
		Predicted	
		Nodule	Non-nodule
Actual	Nodule	243 ( $t_p$ )	2 ( $f_n$ )
	Non-nodule	15 ( $f_p$ )	29 ( $t_n$ )

**Table 1** Performance comparison of the proposed method with the traditional methods

Schemes	Accuracy (%)	Recall (%)	Precision (%)	AUC (%)
Deep learning (2019)	91.6	86.52	88.4	95.7
Two-stage CNN (2020)	92.34	87.54	89.17	95.92
DBN (2021)	93.12	88.17	89.76	96.13
Proposed	94.08	89.3	94.1	96.21



**Fig. 12** Graphical comparison of accuracy, recall, precision, and AUC with the traditional schemes (.)

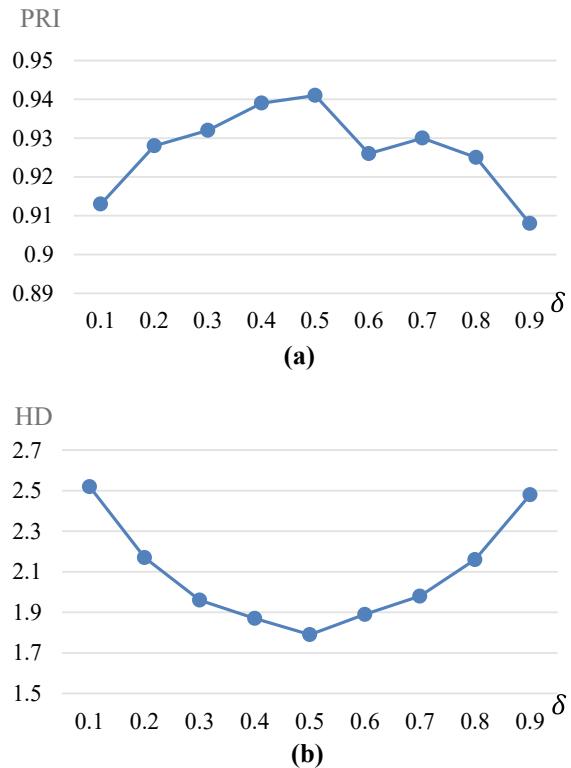


**Fig. 13** ROC curve for the proposed scheme

**Table 2** Comparison of PRI and HD for the proposed method

Metric	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
PRI	0.913	0.928	0.932	0.939	0.941	0.926	0.93	0.925	0.908
HD	2.52	2.17	1.96	1.87	1.79	1.89	1.98	2.16	2.48

Table 2 depicts the comparison of PRI and HD for different step-size  $\delta$ . The PRI and HD of the proposed method are 0.941 and 1.79, respectively, which is better at step-size  $\delta = 0.5$  as depicted in Fig. 14. Higher the PRI and lower the HD values indicate that the segmentation of nodules region by FCM algorithm closely matches with the ground truth results provided by the databases. The proposed algorithm shows a better segmentation result for the step-size  $\delta = 0.5$ .

**Fig. 14** Variation of PRI and HD with step-size  $\delta$  **a** PRI, **b** HD

## 5 Conclusion

This paper proposes a lung nodule segmentation algorithm that uses adaptive weights for the exact detection of boundaries and classification. The algorithm initially pre-processes the image and eliminates the background. The boundaries of the detected lung nodule candidates are then optimized using the FCM algorithm by estimating the low-level adaptive weights along with the boundary layers in zigzag order. The high-level adaptive weights are extracted throughout the lung nodule candidate and are classified in a pre-trained recurrent neural network. The validation was done using the LIDC dataset using the metrics such as Hausdorff distance, probability rand index, accuracy, recall, and precision. The proposed approach provides a Hausdorff distance and PRI of 1.79 and 0.941, respectively. The accuracy, recall, and precision in nodule/non-nodule classification are estimated as 94.08%, 89.3%, and 94.1%, respectively.

## References

1. Diciotti, S., Lombardo, S., Falchini, M., Picozzi, G., Mascalchi, M.: Automated segmentation refinement of small lung nodules in CT scans by local shape analysis. *IEEE Trans. Biomed. Eng.* **58**, 3418–3428 (2011). <https://doi.org/10.1109/TBME.2011.2167621>
2. Farag, A.A., El Munim, H.E.A., Graham, J.H., Farag, A.A.: A novel approach for lung nodules segmentation in chest CT using level sets. *IEEE Trans. Image Process.* **22**, 5202–5213 (2013). <https://doi.org/10.1109/TIP.2013.2282899>
3. Song, J., Yang, C., Fan, L., Wang, K., Yang, F., Liu, S., Tian, J.: Lung lesion extraction using a toboggan based growing automatic segmentation approach. *IEEE Trans. Med. Imaging.* **35**, 337–353 (2016). <https://doi.org/10.1109/TMI.2015.2474119>
4. Wu, D., Lu, L., Bi, J., Shinagawa, Y., Boyer, K., Krishnan, A., Salganicoff, M.: Stratified learning of local anatomical context for lung nodules in CT images. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 2791–2798. IEEE, San Francisco, CA, USA (2010). <https://doi.org/10.1109/CVPR.2010.5540008>
5. El-Bazl, A., Vanbogaert, E., Gimel'jarb, G., El-Ghar, A.: A novel shape-based diagnostic approach for early diagnosis of lung nodules. 4
6. Xie, Y., Zhang, J., Liu, S., Cai, W., Xia, Y.: Lung nodule classification by jointly using visual descriptors and deep features. In: Müller, H., Kelm, B.M., Arbel, T., Cai, W., Cardoso, M.J., Langs, G., Menze, B., Metaxas, D., Montillo, A., Wells III, W.M., Zhang, S., Chung, A.C.S., Jenkinson, M., Ribbens, A. (eds.) *Medical Computer Vision and Bayesian and Graphical Models for Biomedical Imaging*. pp. 116–125. Springer International Publishing, Cham (2017). [https://doi.org/10.1007/978-3-319-61188-4\\_11](https://doi.org/10.1007/978-3-319-61188-4_11)
7. Chen, S., Qin, J., Ji, X., Lei, B., Wang, T., Ni, D., Cheng, J.-Z.: Automatic scoring of multiple semantic attributes with multi-task feature leverage: a study on pulmonary nodules in CT images. *IEEE Trans. Med. Imaging.* **36**, 802–814 (2017). <https://doi.org/10.1109/TMI.2016.2629462>
8. Srensen, L., Shaker, S.B., de Bruijne, M.: Quantitative analysis of pulmonary emphysema using local binary patterns. *IEEE Trans. Med. Imaging.* **29**, 559–569 (2010). <https://doi.org/10.1109/TMI.2009.2038575>
9. Dhara, A.K., Mukhopadhyay, S., Dutta, A., Garg, M., Khandelwal, N.: A Combination of shape and texture features for classification of pulmonary nodules in lung CT images. *J. Digit. Imaging.* **29**, 466–475 (2016). <https://doi.org/10.1007/s10278-015-9857-6>

10. Das, S., Kharbanda, K., M, S., Raman, R., D, E.D.: Deep learning architecture based on segmented fundus image features for classification of diabetic retinopathy. *Biomed. Signal Process. Control.* **68**, 102600 (2021). <https://doi.org/10.1016/j.bspc.2021.102600>
11. Xie, Y., Xia, Y., Zhang, J., Song, Y., Feng, D., Fulham, M., Cai, W.: Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest CT. *IEEE Trans. Med. Imaging.* **38**, 991–1004 (2019). <https://doi.org/10.1109/TMI.2018.2876510>
12. Wu, Z., Zhou, Q., Wang, F.: Coarse-to-fine lung nodule segmentation in CT images with image enhancement and dual-branch network. *IEEE Access.* **9**, 7255–7262 (2021). <https://doi.org/10.1109/ACCESS.2021.3049379>
13. Cao, H., Liu, H., Song, E., Ma, G., Jin, R., Xu, X., Liu, T., Hung, C.-C.: A two-stage convolutional neural networks for lung nodule detection. *IEEE J. Biomed. Health Inform.* 1–1 (2020). <https://doi.org/10.1109/JBHI.2019.2963720>
14. Nasrullah, N., Sang, J., Alam, M.S., Mateen, M., Cai, B., Hu, H.: automated lung nodule detection and classification using deep learning combined with multiple strategies. *Sensors* **19**, 3722 (2019). <https://doi.org/10.3390/s19173722>
15. Dr. Manoharan, S., Sathish: Improved version of graph-cut algorithm for CT images of lung cancer with clinical property condition. *J. Artif. Intell. Capsule Netw.*, **2**, 201–206 (2020). <https://doi.org/10.36548/jaicn.2020.4.002>
16. Dr. Manoharan, S., Sathish: Early diagnosis of lung cancer with probability of malignancy calculation and automatic segmentation of lung CT scan images. *J. Innov. Image Process.*, **2**, 175–186 (2020). <https://doi.org/10.36548/jiip.2020.4.002>
17. Sathish: Adaptive shape based interactive approach to segmentation for nodule in lung CT scans. *J. Soft Comput. Paradigm.*, **2**, 216–225 (2020). <https://doi.org/10.36548/jscp.2020.4.003>
18. Tripathi, M.: Analysis of convolutional neural network based image classification techniques. *J. Innov. Image Process.*, **3**, 100–117 (2021). <https://doi.org/10.36548/jiip.2021.2.003>
19. Harsono, I.W., Liawatimena, S., Cenggoro, T.W.: Lung nodule detection and classification from Thorax CT-scan using RetinaNet with transfer learning. *J. King Saud Univ.—Comput. Inf. Sci.* S1319157820303335 (2020). <https://doi.org/10.1016/j.jksuci.2020.03.013>
20. Su, Y., Li, D., Chen, X.: Lung nodule detection based on faster R-CNN Framework. *Comput. Methods Programs Biomed.* **200**, 105866 (2021). <https://doi.org/10.1016/j.cmpb.2020.105866>
21. Veronica, B.K.J.: An effective neural network model for lung nodule detection in CT images with optimal fuzzy model. *Multimed. Tools Appl.* **79**, 14291–14311 (2020). <https://doi.org/10.1007/s11042-020-08618-x>
22. Baker, A.A., Ghadi, Y.: Cancerous lung nodule detection in computed tomography images. *TELKOMNIKA Telecommun. Comput. Electron. Control.*, **18**, 2432 (2020). <https://doi.org/10.12928/telkomnika.v18i5.15523>

# Review on Segmentation of Facial Bone Surface from Craniofacial CT Images



Jithy Varghese and J. S. Saleema

**Abstract** Three-dimensional (3D) representation of facial bone surface is needed in the virtual surgical planning for orthognathic surgery. Segmentation of facial bone surface from computed tomography images is first step in developing the 3D model. With the advent in the computer vision techniques, various automatic and semi-segmentation algorithms were developed in the recent years for segmentation of facial bone surface from craniofacial computed tomography images. In the proposed paper, the various segmentation techniques for extracting bone surface from 3D CT images for corrective jaw surgery available in the literature have been discussed. By reviewing all the methods available in the literature, it is found that each method has its own merits and demerits.

**Keywords** Segmentation · Orthognathic surgery · Mandible segmentation · Craniofacial image analysis · Computed tomography · Deep learning · Machine learning

## 1 Introduction

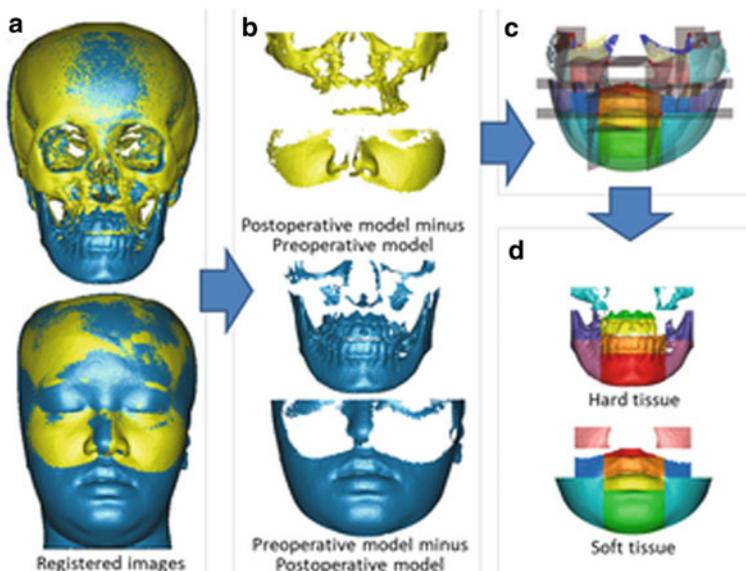
The mouth and maxillofacial regions form soft and hard anatomical tissues of the mouth, jaws, face, and skull [1]. The hard tissue is the maxilla, the mandible, and the teeth. The soft tissues consist of four muscles that are used to chew the food (Fig. 1). Muscles monitor the movement of the mandible and teeth for chewing [2]. The failure of any of these muscles leads to problem in chewing. Also to treat patients with oral cancer, the surgeons will be forced to remove a portion of any of the hard tissues in the facial area to avoid the growth of the tumor. This surgery is on the basis of virtual surgical planning (VSP) to perform the process accurately. Apart from this,

---

J. Varghese (✉) · J. S. Saleema

Department of Computer Science, Christ Deemed To Be University, Bangalore, Karnataka, India  
e-mail: [Jithy.varghese@res.christuniversity.in](mailto:Jithy.varghese@res.christuniversity.in)

J. Varghese  
Christ Academy Institute for Advanced Studies, Bangalore, Karnataka, India



**Fig. 1** Soft and hard tissues in oral and maxillofacial region [3]

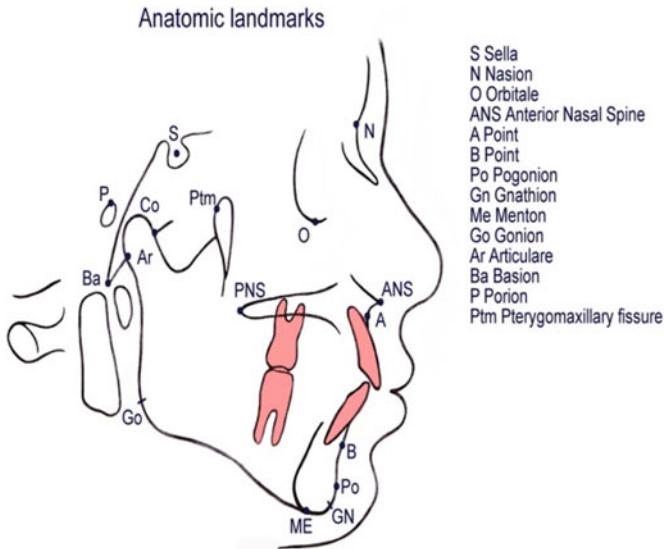
in dental implant surgery as well as orthognathic surgery also precise extraction of bones in the facial area is required for the 3D virtual surgical planning.

Oral and maxillofacial surgeons perform orthognathic surgery to correct misalignment of the jaw, skeletal, and dental abnormalities. After the surgery, the esthetic look of the patient improves. The 3D imaging techniques like computed tomography (CT) and magnetic resonance imaging (MRI) are used widely for cephalometric analysis of face in the orthognathic surgery. CT is typically used in dental medicine for the study of bone structures such as maxilla and teeth while MRI is used to study muscles.

Cephalometric for Orthognathic Surgery (COGS) explains the landmarks of orthognathic surgery and the steps used. They are called cephalometric standards used for orthognathic surgery. The morphological diversity between various ethnic and racial groups causes variations in cephalometric standards; therefore, standards defined for one population group do not apply to all. For orthognathic surgery in the Indian population, a standard cephalometric hardwood norm has not been established.

Segmentation is a preparatory step for virtual orthognathic surgical planning. The result of the segmentation can then be used to obtain further diagnostic insights by plotting the hard tissue landmarks (Fig. 2) and representing it as 3D model.

The computer-aided radiological task to analyze various imaging modalities is fulfilled through the key fundamental step of the delineation of regions of interest using segmentation algorithms. The segmentation aims at a 3D representation of hard and soft tissues, which may be used for virtual preparation in orthodontics and orthographic surgery. The segmentation of image algorithms is based on either the theory of discontinuity or similarity. The theory of discontinuity is used for the



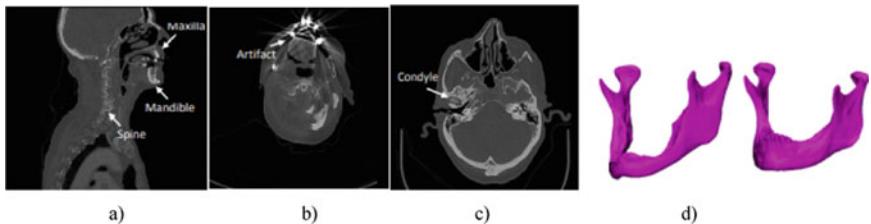
**Fig. 2** Anatomic landmarks used in the orthognathic surgery [Burstone CJ, James RB, Legan HL-1978]

extraction of areas that vary in intensity, color, texture, or other image statistics. The concept of similarity is used for grouping pixels based on common characteristics.

Segmentation of a medical image is a challenging task because there is no universal segmentation algorithm process that works with all forms of anatomy. The study addresses prior research on state-of-the-art issues in oral or maxillary facial images with soft and hard tissue segmentation. The various issues involved in the segmentation of facial bone surface are discussed in this section [66]. The anatomy of bone structures in the facial region is complex. Finding the exact boundaries and localizing the various bones is a very challenging task. When X-ray beams pass through structures with high density like teeth and metal implants, it leads to attenuation errors in cone beam computed tomography images and computed tomography images; this further leads to noise and artifacts in the scan image [4]. The tomography process thinnens the various bones in the facial surface which reduces the contrast. Shape of the various bones varies among the individuals [5]. Figure 3 shows various challenges in the field of facial bone segmentation.

## 2 Public Database

There are four publically available datasets of head and neck CT images. They are (1) Public Domain Database for Computational Anatomy (PDDCA) (<http://www.imagegenglab.com/newsite/pddca/>) [6, 65], PDDCA consists of open-access CT images



**Fig. 3** Challenges in facial bone segmentation. **a** Various bone-structured organs in the H&N scans. **b** Metal artifacts. **c** Low intensity. **d** Large variation in mandibles between patients [65]

of 48 patients used for cancer research from Radiation Therapy Oncology Group (RTOG) (2) The Cancer Imaging Archive (TCIA) [7, 65] provides Head–Neck Cetuximab [8, 65], database which is manually annotated at the mandible and brainstem (3) The StructSeg 2019 dataset used for the Automatic Structure Segmentation for Radiotherapy Planning Challenge 2019 [9, 65]. (4) Cancer Genome Atlas Head–Neck Squamous Cell Carcinoma database [10, 65]. In the literature, the most commonly used dataset for segmentation of mandible is PDDCA.

### 3 Evaluation Metrics

Most of the studies in the literature used in house dataset while few others used PDDCA dataset. Various evaluation metric are used for segmentation of facial bone surface from the CT images. Mainly the evaluation metrics are categorized into two groups: (1) overlap-based metrics and (2) distance-based metrics. The metrics used in the overlap-based methods find the difference in the measurement between the ground truth image and the prediction made automatically. This is done on the basis of true positive (TP), true negative (TN), false negative (FN), and false positive (FP) values from the confusion matrix. The most commonly used overlap-based metrics in the literature is dice coefficient. Distance-based metrics use the difference in contour between the ground truth image and the automatic segmentation. The most commonly used metrics used in this category are average symmetric surface distance (ASD), Hausdorff distance (HD), root mean square error (RMSE), and mean square error (MSE) [58, 65]. Table 1 shows various important quality metrics used in the literature.

### 4 Methodology

This section discusses the various segmentation techniques used to find region of interest during orthographic surgery. Few major commonly used techniques are (1)

**Table 1** Performance metrics used for performance measurement of facial bone segmentation [60, 65]

Metric	Abbreviation overlap-based metrics (%)	Definition
Dice similarity index	Dice	$\text{Dice} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$
<i>Distance-based metrics (%)</i>		
Average symmetric surface distance	ASD	$\text{ASD}(A, B) = \frac{d(A, B) + d(B, A)}{2}$ where $d(A, B) = \frac{1}{N} \sum_{a \in A} \min_{b \in B} \ a - b\ $
Hausdorff distance	HD	$D_{\text{HD}}(A, B) = \max(h(A, B), h(B, A))$ Where $h(A, B) = \max_{a \in A} \min_{b \in B} \ a - b\ $
Mean square error	MSE	$\text{MSE} = \frac{1}{n} \sum_{i=1}^n \ A_i - B_i\ ^2$
Root mean square error	RMSE	$\text{RMSE} = \sqrt{\text{MSE}}$

statistical shape model technique, (2) atlas-based technique, (3) level set-based technique, (4) artificial neural network classifier, and (5) deep learning approaches. A brief analysis of each technique is done in this section.

#### 4.1 Statistical Shape Models

For robust segmentation based on prior shape information, statistical shape models (SSM) are used. Active form models (ASM) and active appearance models (AAM) are the two most widely used methods of statistical models.

Manual segmentation of various medical volumetric images originally produces a set of training samples. From this training dataset, a statistical form model is then built that extracts the middle form and the most relevant variation modes. The training set is often used to encode the feature of an object into a statistical model appearance [11]. The search algorithm employs both the form and the appearance model to extract an item of interest from a previously unseen image.

The appearance model (AM) of the object of interest must also be designed to apply the SSM for object segmentation. There are two well-established AM families in principle: border and region-oriented. The profiles perpendicular to the object surface is tested for the necessary landmarks to create an AM-based boundary. PCA or k-NN-classifier techniques of machine learning are then introduced to identify each landmark globally. The whole internal object area is analyzed in the region-based AM. The model of active appearance is one proven approach All samples of training are transformed first into a typical form, e.g., into an SSM's mean form. For each image transformed, a vector function encodes the characteristics of an inner

object area. A PCA is then used on the functional vectors to construct the appearance model.

Overall in SSM technique for segmentation of mandible, four steps are required. They are image preprocessing techniques, construction of correspondence maps, principal component analysis, and segmentation of mandible [12]. Table 2 shows summary of SSM-based, ASM-based, and AAM-based techniques.

## 4.2 *Atlas-Based Segmentation*

If the relationship between regions and pixel intensity is unspecified due to a lack of border or noise, or if the artifacts to be segmented have the same texture, then atlas segmentation can be used. The atlas-based segmentation is successful when information on differences between these objects is included in the spatial relationship between the objects or compliance with their morphometric properties. The segmentation of the atlas is primarily used if the information on a gray level is not sufficient to segment an image. The benefit of atlas-based segmentation is that the shape of an object is determined or morphological differences are observed among patient groups. Atlas structure [18] or dynamic non-rigid registry is a drawback of an atlas-based segmentation. The segmentation based on the atlas is usually done when gray-level knowledge isn't enough. Table 3 shows summary of atlas-based segmentation.

## 4.3 *PDE-Based Image Segmentation*

PDE-based method does segmentation by solving PDE equation. There active contour segmentation or snakes is the most commonly used PDE-based image segmentation methods [25]. The key advantage of partial differential equation methods is that they are fast in comparison with others and can be used in critical time environments.

Active snake contour models are models for the delineation of an entity using closed parametric contours which can be deformed using internal and external forces [26].

A closed contour must initially be initialized near enough to the desired limit. Then the contour is evolved by energy minimization against the target object contour. The active contour technique is based on minimization of internal energy and external energy. The internal energy is reduced as the curve reaches a shape close to the target object. The external energy is reduced as the regularized gradient close to the contour reaches peak value. For a dividing the image into various segments, this technique takes into consideration the homogeneity of the intensity in various regions and the similarity between the contours.

**Table 2** Summary of SSM-based, ASM-based, and AAM-based technique

**Table 3** Summary of atlas-based segmentation

Author and year	Technique	Quality metric			Dataset	Outcome	Imaging modality
		Dice	ASD	HD			
Li Wang et al. 2014	Atlas-based [19]	0.92 ± 0.02	0.65 ± 0.19	0.96 ± 0.53	In house	Mandible Maxilla	CT
Li Wang et al. 2013	Atlas- and convex-based [20]	0.91 ± 0.02	—	0.92 ± 0.47	In house	Mandible	CT
Ayyalusamy et al. 2019	Atlas-based [21]	85.00	—	—	In house	Mandible	CT
Haq et al. 2019	Atlas-based [22]	85.00	—	—	In house	Mandible	CT
		83.00	—	—	PDDCA		
McCarroll et al. 2018	Atlas-based [23]	84.00 ± 7.00	1.89 ± 1.55	18.63 ± 14.90	In house	Mandible	CT
Huang et al. 2019	Atlas-based [24]	84.50 ± 1.60	—	—	In house	Mandible	CT

Later an advanced modified snake algorithm called level set-based image segmentation is developed. In this technique, an initial curve is plotted and then the evaluation of the curve is done by reducing the functional energy. Using this method, the segmentation problem is converted into convex optimization problem [30]. Table 4 shows summary of PDE-based segmentation.

## 4.4 *Classical Machine Learning-Based Methods*

With the advancement in the field of computer vision, researchers started using classical machine learning (CML) methods widely for various medical image processing researches. Various CML techniques used in the field of facial bone segmentation in the literature are 1) thresholding techniques, 2) region growing, 3) classifiers, 4) clustering.

### A. Thresholding Techniques

Thresholding is the most frequently used method to segment an image. The thresholding technique finds a value, called “threshold,” based on the intensity of global distribution of the pixel [31]. Local thresholding method is called adaptive thresholding [32]. Bi-level thresholding is a method of partitioning the image into white and black. If a gray value threshold divides images into many parts, then multiple threshold divisions can be extended [33]. It is called as segmentation of several thresholds [34]. The segment 1 contains all pixels of the same value as the first threshold, and the segment 2 contains all the pixels between the first threshold and the second threshold [35], etc. After that, the image is divided into  $n + 1$  segments. The histogram is a method for finding the acceptable threshold value. Thresholding is used to get a rough idea about the shape of the region of interest. If the image is noisy with uneven background and poor illumination the accuracy of thresholding will not be good.

### B. Region Growing

The technology behind the region growing approach [36] is that in some preset condition [37] all pixels belonging to the region of interest are identical. The criterion for homogeneity to direct the original region toward its final target region differs in various ways [32]. One of the most used criteria is if the gray value of a pixel is within a particular predefined range then that pixel belongs to the ROI [38].

The region growing algorithm works with one connected region at a time. But in seeded region growing approach, in the set of  $s_1, s_2, \dots, s_n$  seeds, the  $n$  pixels in each step are added and the image is divided into disjoint subregions [39]. In the following step, the centers of the newly formed sub-regions replace these seeds. The pixels are labeled with different symbols in different regions [41]. The pixels are referenced as pixels, and the unmarked pixels are referred to as pixels not assigned. Unallocated pixels are applied according to pixel gray value to the mean gray value measurements in different sub-regions. Seed generation and pixel labeling are the main difficulties of

**Table 4** Summary of PDE-based segmentation

Author and year	Technique	Quality metric			Dataset	Outcome	Imaging modality
		Dice	ASD	HD			
Wang et al. 2014	Local level set model [27]	0.92 ± 0.02	0.65 ± 0.19	0.96 ± 0.53	—	—	In house Mandible CT
Gan et al. 2018	Local level set model [28]	95.64 ± 2.12	0.25 ± 0.02	1.07 ± 0.36	—	—	In house Tooth and Alveolar bone CT
M. Brandariz et al. 2014	Deformable snake model [29]	—	—	—	4.30	4.01	In house Mandible CT
Wu et al. 2018	Fuzzy Model [51]	89.00	—	—	In house	Mandible CT	Wu et al. 2018 Fuzzy Model[51]

seeded region growing technique [40]. Segmentation using regional techniques can cause noise to cause holes or disconnected areas. Pixel labeling and seed generations are the problems of seeded region growing [39].

### C. Classifiers

In clustering, any pixel in an image may be given a tissue label and the labels are chosen for identification purpose. Groups can be tangible, alveolar, and other tissue in the case of mandibular segmentation. A functional area is extracted from training datasets that divide the picture into a number of regions. A function on the original or smoother image is used to create the feature space. Intensity types and textures are the most common features.

In order to train samples, two forms of classification are primarily parametric and nonparametric. The nearest neighbor algorithm (k-NN) is the most common nonparametric classification algorithm. It classifies artifacts in a region of operation with the closest distance training samples. The approach for identifying patterns is one of the simplest [42]. The distance can be measured by pixel intensity.

Maximum likelihood classifier is the most commonly used model parametric classifier [43]. The classifier groups each pixel to a particular class on the basis of probability. It follows Gaussian distribution [44]. “The parameters of the Gaussian models are mean and the covariance matrix. Some cases maximum likelihood classifier will not work like k-NN even though it is computationally efficient. It leads to misclassifications in segmentation.”

“Classifiers are used to segment medical images which have quantifiable features. The interested structures in new datasets” can be identified with the help of labeled training sets. But the disadvantage is that the labeling is done by manual segmentations which is a tedious task which needs the help of subject experts.

### D. Clustering

Clustering without fixed classification marks is an unsupervised technique. K-Means and C-Means are the most frequently used clustering techniques for image segmentation [45, 46]. K-means dividing the pixels in the picture into k cluster where the average intensity determines each pixel as one cluster. The cluster is marked as the pixel with the closest value to the mean strength of the cluster. If the new pixel has been labeled, the medium strength of the cluster is optimized iteratively. A pixel is often included in several clusters [47] in the fuzzy C-means clustering technique. The partition matrix contains to what extent a pixel is part of the cluster [48].

Clustering algorithms involve the initialization of segmentation parameters that affect segmentation efficiency. They are vulnerable to homogeneity since the spatial information in the images is generally overlooked. Table 5 shows summary of classical machine learning model classifier.

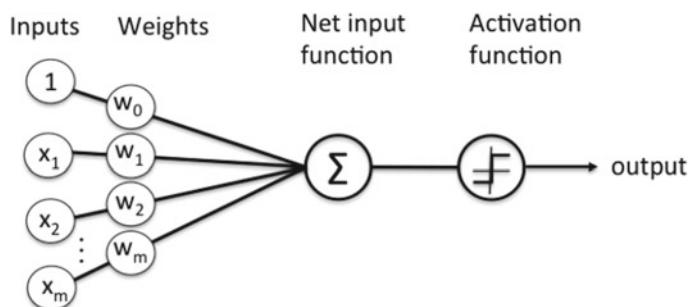
**Table 5** Summary of classical machine learning model classifier

Author and year	Technique	Quality metric			Dataset	Outcome	Imaging modality
		Dice ratio (%)	ASD (mm)	HD (mm)			
Gan et al. 2018	Random transform and local level set model [28]	95.64 ± 2.12	0.25 ± 0.02	–	In house	Tooth and Alveolar Bone	CT
Li Wang, Yaozong Gao et al. 2016	Feature extraction using Random Forest Classifier [49]	0.94 ± 0.02	0.42 ± 0.15	0.74 ± 0.25	In house	Mandible and Maxilla	CT
Linares et al. 2019	Clustering [50]	92.88	–	–	In house	Mandible	CBCT
Wu et al. 2019	Classic Machine Learning Model [52]	89.00	–	–	In house	Mandible	CT
Torosdagli et al. 2017	Classic Machine Learning Model [64]	91.00			PDDCA	Mandible	CT

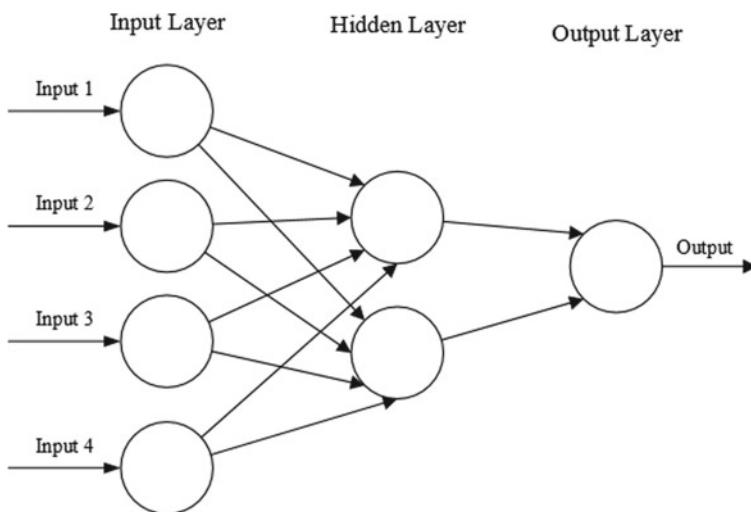
#### 4.5 Artificial Neural Network Classifier

Artificial neural network is a supervised learning algorithm used for classification. There are three layers in artificial neural network, and they are input, hidden, and output layers. In each layer, there is a place for computation to happen and it is called as node. Data and weights are combined in each node it amplifies or quite each node. Thus, it assigns a value of significance for each node which helps in classifying the data without any mistake [67]. The product of input and weight is summed up and given to an activation function to the extent to which the input value pass through the neural network which helps in classification. The node of the neural network is given in Fig. 4 a three-layer neural network is given in Fig. 5

The equation is given :



**Fig. 4** Node in a neural network



**Fig. 5** Three-layer neural network

$$f(x) = (Wr.x + b) \quad (1)$$

where sigma is activation function, Wr is weight matrix, and b is bias. This technique is known as forward propagation. Neural network can improve the process of learning which indirectly increase the accuracy of classification with the help of backpropagation technique. In this technique, the output after processing is compared with the expected output and error is calculated. The calculated error is propagated back from output layer to the input layer. Each node is adjusted with the help of gradient descent. Backward propagation of neural network is expressed in Eq. 2

$$dx = mc * dxprev + lr * (1 - mc) * dperf/dx \quad (2)$$

lr is learning parameter, value of momentum is mc, dx derivative of performance is dx, change in weight and bias id dxprev.

## 4.6 Convolution Neural Network(CNN)

Computational complexity was the disadvantage of ANN due to that a new method called convolution neural network came into existence. CNN is used most commonly to classify huge set of image data. Convolutional neural network uses filtering process simplify the processing of complex image data. CNN considers data which is also called as input neuron, as a spatial component with three dimensions height, width, and depth. Convolutional neural network architecture has three layers. They are convolution layer, pooling layers, and fully connected layer. In this technique, neurons in each layer are not connected to all neurons in the previous layer but they are only connected to nearby neurons which have same weight.

### CNN Architecture

**Convolution Layer:** This layer extracts low level features like orientation, edges, gradient, and color. When more layers are added, it extracts high level features also which in turn provides us with a network which gives a complete understanding of the images available in the dataset.

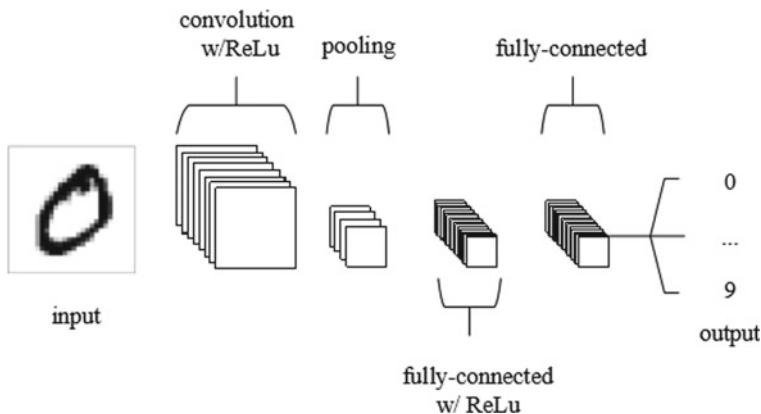
**Pooling Layer:** This layer helps to reduce the dimension of the features extracted in the convolution layer. This helps to reduce the computational complexity to classify the data. Thus, this layer helps to extract rotationally and positional invariant highly dominant features of the data which helps in efficient training of the model.

**Fully Connected Layer:** This layer takes the output of convolution and pooling layer as input and assigns a label to an image in the dataset. For doing this, it converts the output of convolution and pooling layer to a single vector which represents the probability of a feature to belong to a class. Table 7 shows summary of DL techniques.

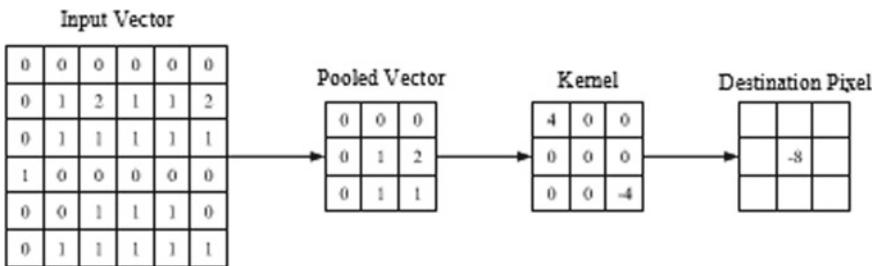
Figure 6 shows CNN architecture, and Fig. 7 shows convolutional layer framework.

**Table 6** Summary of deep learning model models

Author and year	Technique	Quality metric			Dataset	Outcome	Imaging modality
		Dice Ratio(%)	ASD	HD			
Neves, C.A et al. 2021	CNN [53]	0.91	—	0.25	In house	Temporal Bone	CT
Jaskari, J et al. 2021	CNN [54]	—	0.45 mm	—	In house	Mandibular Canal	CT
Xue et al. 2021	DL [55]	94.00 ± 2.00	0.49 ± 0.18 mm	2.36 ± 0.62	PDDCA	Mandible	CT
Liang et al. 2020	DL [56]	94.10 ± 0.70 91.10 ± 1.00(L) 91.40 ± 2.00 (R)	0.28 ± 0.14 0.76 ± 0.13 (L) 0.86 ± 0.14 (R)	—	PDDCA In house	Mandible	CT
Gou et al. 2020	DL [57]	94.00 ± 1.00	0.47 ± 0.11	1.40 ± 0.02	PDDCA	Mandible	CT
He et al. 2020	DL [59]	90.30 (L); 90.80 (R)	—	—	StructSeg2019	Mandible	CT
Qiu et al. 2020	DL [60]	97.53 ± 1.65 95.10 ± 1.21	—	—	In house PDDCA	Mandible	CT
Zhu et al. 2019	DL [61]	92.30	—	—	PDDCA	Mandible	CT
Tong et al. 2019	DL [62]	93.91 ± 1.30 81.64 ± 4.44	0.55 ± 0.14 1.13 ± 0.48	—	PDDCA In house	Mandible	CT
Nikolov et al. 2018	DL [63]	93.10 ± 1.90 92.90 ± 3.50 93.80 ± 1.90	—	—	In house TCIA PDDCA	Mandible	CT



**Fig. 6** CNN architecture



**Fig. 7** Convolution layer

## 5 Discussion

Apart from the brief description of each segmentation technique, we present a useful statistics and discussion the segmentation of facial bone from CT images. In this review paper, the author has tried to analyze 30 papers on segmentation of mandible, maxilla, and other facial bone surface. The various segmentation techniques which are applied in the literature are statistical shape model techniques, atlas-based techniques, PDE-based techniques, CML techniques like thresholding, region growing, classification, clustering techniques, and CNN techniques. The number of papers analyzed under each technique is given in Table 7.

Among all the segmentation techniques atlas-based technique is the oldest well-known technique used for facial bone segmentation. Due to the influence of the low contrast and artifacts in the overall process, the accuracy of the atlas-based technique is low when compared to other techniques. As well, the overall process of atlas-based segmentation technique is time-consuming. The main disadvantage of level set method is that it requires initialization of the contour for stretching the boundary and for finding the region of interest for segmentation of facial bones structures. The

**Table 7** Automatic facial bone segmentation techniques and the reference details

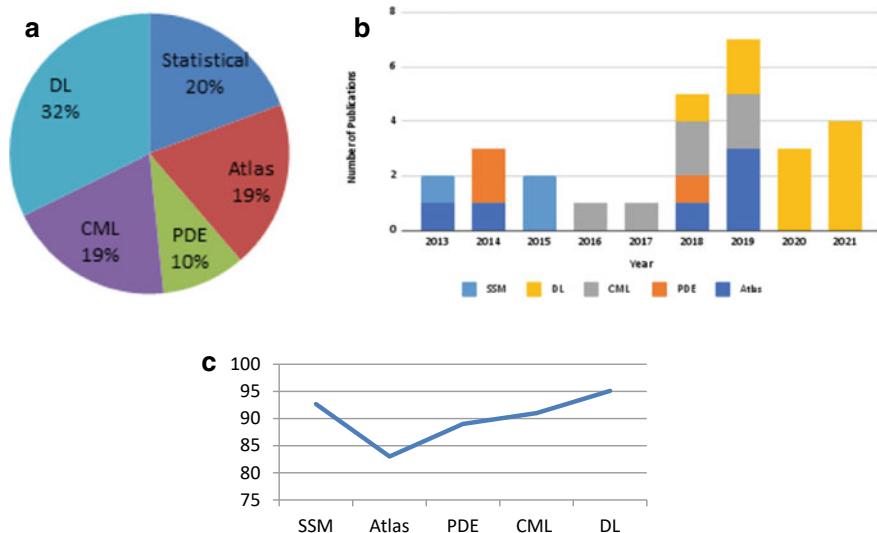
Methodology categories	Publications	Number of publications
SSM-based Techniques	[12–15]	4
ASM-based Techniques	[16]	1
AAM-based Techniques	[17]	1
Atlas-based Techniques	[19–24]	6
PDE-based Techniques	[27–29, 51]	4
CML Techniques	[28, 49, 50, 52]	5
DL Techniques	[53–63]	10

SSM-based techniques used shape as well as texture information for segmentation. As a result of that, the accuracy is better than atlas-based technique. The most popular and efficient facial bone segmentation technique in the literature is DL technique. The disadvantage of manual extraction of features in machine learning technique is overcome with the help of DL technique. The major problem with DL technique is the requirement of huge amount of annotated facial CT images to achieve good amount of accuracy and avoid overfitting. The availability of public dataset is limited in this area of study so to overcome this issue various data augmentation techniques are used. In the current study, we have mainly focused on CT images as it is economical and has low level of radiation and it provides more clarity of bone anatomy when compared to magnetic resonance imaging technique. Figure 8a and b shows the share of methodology categories and distribution of methodology categories, respectively.

The most commonly used public dataset throughout the literature is PDDCA [6] dataset while StructSeg2019 [9], TCIA [7] datasets are used in few studies. The dice ratio is the most common quality metrics used for evaluating the segmentation process. The gradual increase in the dice ratio is shown in Fig. 8c. The maximum dice ratio achieved by DL technique is 95.10 [60]. Major difficulty in this field of research is the varying structure, shape, and density of the bones in the facial area. The proposed study will definitely help the researchers to understand the state of the art in this area of research and come out with very efficient hybrid techniques.

## 6 Conclusion

Accuracy of mandible segmentation is very important for facial reconstruction, 3D modeling, and mesh generation of facial bone surface in orthognathic surgery as well as in 3D VSP in the treatment of cancer patients. Especially due to the diversity in the



**Fig. 8** **a** Share of methodology categories **b** distribution of methodology categories **c** gradual increase in the dice ratio

facial anatomy in different racial groups in India, there is no standard cephalometric norm which makes it very difficult to come up with a standard accurate segmentation algorithm. Before the boom of computer vision techniques, atlas-based technique was used widely for the segmentation and later with the advent of computer vision techniques researchers started exploring various deep learning techniques for the same. In the future work, we will try to combine various deep learning techniques along with image processing techniques to improve the accuracy and thereby address the issues of current mandible segmentation techniques more efficiently.

## References

1. Eder, D., Brealey, R., Bertram, J., Eder, D., Kaminsky, S., Waters, J.: Laboratory Atlas of Anatomy and Physiology. Mc Graw-Hill, New York, 4th edn, (2003)
2. Leggott, B.: The Anatomical Basis of Dentistry. Mosby Elsevier (2011)
3. Lo, L.J., Weng, J.L., Ho, C.T., Lin, H.H.: Three-dimensional region-based study on the relationship between soft and hard tissue changes after orthognathic surgery in patients with prognathism. PLoS ONE **13**(8), e0200589 (2018). <https://doi.org/10.1371/journal.pone.0200589>. PMID:30067766;PMCID:PMC6070212
4. Vaitiekunas, M., Jegelevičius, D., Sakalauskas, A., Grybauskas, S.: Automatic method for bone segmentation in cone beam. Comput. Tomography Data Set. Appl. Sci., **10**, 236 (2020). [CrossRef]
5. Abdolali, F., Zoroofi, R.A., Abdolali, M., Yokota, F., Otake, Y., Sato, Y.: Automatic segmentation of mandibular canal in cone beam CT images using conditional statistical shape model and fast marching. Int. J. Comput. Assist. Radiol. Surg. **12**, 581–593 (2017). [CrossRef]

6. Raudaschl, P.F., Zaffino, P., Sharp, G.C., Spadea, M.F., Chen, A., Dawant, B.M., Albrecht, T., Gass, T., Langguth, C., Lüthi, M., et al.: Evaluation of segmentation methods on head and neck CT: auto-segmentation challenge 2015. *Med. Phys.*, **44**, 2020–2036 (2017). [CrossRef] [PubMed]
7. Clark, K., Vendt, B., Smith, K., Freyman, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al.: The cancer imaging archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging*, **26**, 1045–1057 (2013). [CrossRef] [PubMed]
8. Bosch, W., Straube, W., Matthews, J., Purdy, J.: Head-neck cetuximab. The Cancer Imaging Archive. 2015. Available online: <https://wiki.cancerimagingarchive.net/display/Public/Head-Neck+Cetuximab>. Accessed on 23 Feb 2021
9. Tang, H., Chen, X., Liu, Y., Lu, Z., You, J., Yang, M., Yao, S., Zhao, G., Xu, Y., Chen, T., et al.: Clinically applicable deep learning framework for organs at risk delineation in CT images. *Nat. Mach. Intell.*, **1**, 480–491 (2019). [CrossRef]
10. Zuley, M.L., Jarosz, R., Kirk, S., Lee, Y., Colen, R., Garcia, K., Arede, N.: Radiology data from the cancer genome atlas head-neck squamous cell carcinoma [TCGA-HNSC] collection. *Cancer Imaging Arch.* **10**, 9 (2016)
11. Cremers, D., Tischhauser, F., Weickert, J., Schnorr.: Diffusion snakes: introducing statistical shape knowledge into the mumford-shah functional. *Int. J. Comput. Vision* **50**, 295–313 (2002)
12. Kim, S.-G., et al.: Development of 3D statistical mandible models for cephalometric measurements. *Imaging Sci. Dentistry* **42**(3), 175–182 (2012)
13. Chang, Y.-B., Xia, J.J., Yuan, P., Kuo, T.-H., Xiong, Z., Gateno, J., Zhou, X.: 3D Segmentation of maxilla in cone-beam computed tomography imaging using base invariant wavelet active shape model on customized two-manifold topology. *J. Xray Sci. Technol.* **21**(2), 251–282 (2013)
14. Gollmer, S.T., Buzug, T.M.: “Fully automatic shape constrained mandible segmentation from cone-beam CT data.” In: 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI), Barcelona, 2012, pp. 1272–1275
15. Rueda S., Gil J.A., Pichery R., Alcañiz M. (2006) Automatic segmentation of jaw tissues in CT using active appearance models and semi-automatic landmarking. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) Medical Image Computing and Computer-Assisted Intervention—MICCAI 2006. MICCAI 2006. Lecture Notes in Computer Science, vol. 4190. Springer, Berlin, Heidelberg
16. Albrecht, T., Gass, T., Langguth, C., Lüthi, M.: Multi atlas segmentation with active shape model refinement for multi-organ segmentation in head and neck cancer radiotherapy planning. In: Proceedings of the Head Neck Auto-Segmentation Challenge (MICCAI), Munich, Germany, 5–9 October 2015
17. Mannion-Haworth, R., Bowes, M., Ashman, A., Guillard, G., Brett, A., Vincent, G.: Fully automatic segmentation of head and neck organs using active appearance models. In: Proceedings of the Head Neck Auto-Segmentation Challenge (MICCAI), Munich, Germany, 5–9 October 2015
18. Heimann, T., Meinzer, H.P.: Statistical shape models for 3d medical image segmentation: a review. *Med. Image Anal.*, **13**(4), 543–563, 5, 18, 19, 21 (2009)
19. Wang, L., et al.: Automated bone segmentation from dental CBCT images using patch-based sparse representation and convex optimization. *Med. Phys.* **41**(4), 043503 (2014)
20. Wang, L., et al.: Automated segmentation of CBCT image using spiral CT atlases and convex optimization. In: Medical Image Computing and Computer-Assisted Intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention, 16(03), 251–258 (2013)
21. Ayyalusamy, A., Vellaiyan, S., Subramanian, S., Ilamurugu, A., Satpathy, S., Nauman, M., Katta, G., Madineni, A.: Autosegmentation of head and neck organs at risk in radiotherapy and its dependence on anatomic similarity. *Radiat. Oncol. J.*, **37**, 134 (2019). [CrossRef]
22. Haq, R., Berry, S.L., Deasy, J.O., Hunt, M., Veeraraghavan, H.: Dynamic multiatlas selection-based consensus segmentation of head and neck structures from CT images. *Med. Phys.*, **46**, 5612–5622 (2019) [CrossRef]

23. McCarroll, R.E., Beadle, B.M., Balter, P.A., Burger, H., Cardenas, C.E., Dalvie, S., Followill, D.S., Kisling, K.D., Mejia, M., Naidoo, K., et al.: Retrospective validation and clinical implementation of automated contouring of organs at risk in the head and neck: a step toward automated radiation treatment planning for low-and middle-income countries. *J. Glob. Oncol.*, **4**, 1–11 (2018) [CrossRef]
24. Huang, C., Badiei, M., Seo, H., Ma, M., Liang, X., Capaldi, D., Gensheimer, M., Xing, L.: Atlas Based Segmentations via Semi-Supervised Diffeomorphic Registrations. arXiv 2019, [arXiv:1911.10417](https://arxiv.org/abs/1911.10417)
25. Paragios, N., Mellina-Gottardo, O., Ramesh, V.: Gradient vector flow fast geometric active contours. *IEEE Trans. Pattern Anal. Mach. Intell.*, **26**, 402–407 (2004)
26. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: active contour models. *Int. J. Comput. Vision*, **1**, 321–331 (1988)
27. Wang, L., Chen, K.C., Gao, Y., Shi, F., Liao, S., Li, G., Shen, S.G., Yan, J., Lee, P.K., Chow, B., et al.: Automated bone segmentation from dental CBCT images using patch-based sparse representation and convex optimization. *Med. Phys.*, **41**, 043503 (2014). [CrossRef]
28. Gan, Y., Xia, Z., Xiong, J., Li, G., Zhao, Q.: Tooth and alveolar bone segmentation from dental computed tomography images. *IEEE J. Biomed. Health Inform.*, **22**(1), 196–204 (2018)
29. Brandariz, M., Barreira, N., Penedo, M.G., Suárez-Cunqueiro, M.: “Automatic segmentation of the mandible in cone-beam computer tomography images.” In: 2014 IEEE 27th International Symposium on Computer-Based Medical Systems, New York, NY, pp. 467–468 (2014)
30. Gamboa, A., Cosa, A., Benet, F., Arana, E., Moratal, D.: “A semiautomatic segmentation method, solid tissue classification and 3D reconstruction of mandible from computed tomography imaging for biomechanical analysis.” In: 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI), Barcelona, pp. 1483–1486 (2012)
31. Chow, C., Kaneko, T.: Automatic boundary detection of the left ventricle from cine angiograms. *Comput. Biomed. Res.*, **5**, 388–410 (1972)
32. Kim, D., Chung, S., Park, J.: Automatic navigation path generation based on two-phase adaptive region-growing algorithm for virtual angiscopy. *Med. Eng. Phys.*, **28**, 339–347 (2006)
33. Bezdek, J., Hall, L., Clarke, L.: Review of image segmentation techniques using pattern recognition. *Med. Phys.*, **20**, 1033–1048 (1993)
34. Jiang, X., Mojon, D.: Adaptive local thresholding by verification based multi threshold probing with application to vessel detection in retinal images. *IEEE Trans. Pattern Anal. Mach. Intell.*, **25**, 131–137 (2003)
35. Cao, L., Shi, Z., Cheng, E.: Fast automatic multilevel thresholding method. *Electron. Lett.*, **38**(868–870), 25 (2002)
36. Adams, Bischof, L.: Seeded region growing. *IEEE Trans. Patt. Anal. Mach. Intell.*, **16**(6), 641–647 (1994)
37. Fabijaska, A.: Two-pass region growing algorithm for segmenting airway tree from mdct chest scans. *Comput. Med. Imaging Graph.*, **33**, 537–546. 26, 43 (2009)
38. Modayur, B., Prothero, J., Ojemann, G., Maravilla, K., Brinkley, J.: Visualization-based mapping of language function in the brain. *Neuroimage*, **6**, 245–258 (1997)
39. Fan, J., Zeng, G., Body, M., Hacid, M.: Seeded region growing: an extensive and comparative study. *Patt. Recogn. Lett.*, **26**, 1139–1156 (2005)
40. Fan, J., Yau, D., Elmagarmid, A., Aref, W.: Automatic image segmentation by integrating color-based extraction and seeded region growing. *IEEE Trans. Image Process.*, **10**, 1454–1466 (2001)
41. Mehmet, A., Jackway, P.: An improved seeded region growing algorithm. *Patt. Recogn. Lett.*, **18**, 1065–1071 (1997)
42. Duda, R., Hart, P. & Stork, D. (2000). *Pattern Classification*. Wiley, 2nd edn.(14)
43. Cam, L., Lucien: Maximum likelihood-an introduction. *ISI Review*, 58(2), 153–171. 30 (1990)
44. Rahmati, P., Adler, A., Hamarneh, G.: Mammography segmentation with maximum likelihood active contours. *Med. Image Anal.*, **16**, 1167–1186 (2012)
45. Chen, C., Luo, J. & Parker, K.: Image segmentation via adaptive k-mean clustering and knowledge-based morphological operations with biomedical applications. *IEEE Trans. Image Process.*, **7**(12), 1673–83. 32 (1998)

46. Chuang, K., Tzeng, H., Chen, S., Wu, J., Chen, T.: Fuzzy c means clustering with spatial information for image segmentation. *Comput. Med. Imaging Graph.* **30**(1), 9–15 (2006)
47. Antila, K., Lilja, M., Kalke, M.: Segmentation of facial bone surfaces by patch growing from cone beam CT volumes. *Dent maxillofacial Radiol.* **45**(8), 20150435 (2016)
48. Mohamed, N., Ahmed, M., Farag, A.: Modified fuzzy c-mean in medical image segmentation. *ICASSP* **6**, 3429–3432. 32(30) (1999)
49. Wang, L., et al.: Automated segmentation of dental CBCT image with prior-guided sequential random forests. *Med. Phys.* **43**(1), 336–346 (2016)
50. Linares, O.C., Bianchi, J., Raveli, D., Neto, J.B., Hamann, B.: Mandible and skull segmentation in cone beam computed tomography using super-voxels and graph clustering. *Vis. Comput.* **35**, 1461–1474 (2019)
51. Wu, X., Udupa, J.K., Tong, Y., Odhner, D., Pednekar, G.V., Simone, C.B., McLaughlin, D., Apinorasethkul, C., Lukens, J., Mihailidis, D., et al.: Auto-contouring via automatic anatomy recognition of organs at risk in head and neck cancer on CT images. In: *Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling; International Society for Optics and Photonics*; Bellingham, WA, USA, Vol. 10576, p. 1057617 (2018)
52. Wu, X., Udupa, J.K., Tong, Y., Odhner, D., Pednekar, G.V., Simone, C.B. II, McLaughlin, D., Apinorasethkul, C., Apinorasethkul, O., Lukens, J., et al.: AAR-RT—A system for auto-contouring organs at risk on CT images for radiation therapy planning: principles, design, and large-scale evaluation on head-and-neck and thoracic cancer cases. *Med. Image Anal.*, **54**, 45–62 (2019)
53. Neves, C.A., Tran, E.D., Kessler, I.M., et al.: Fully automated preoperative segmentation of temporal bone structures from clinical CT scans. *Sci Rep* **11**, 116 (2021). <https://doi.org/10.1038/s41598-020-80619-0>
54. Jaskari, J., Sahlsten, J., Järnstedt, J., et al.: Deep learning method for mandibular canal segmentation in dental cone beam computed tomography volumes. *Sci Rep* **10**, 5842 (2020). <https://doi.org/10.1038/s41598-020-62321-3>
55. Xue, J., Wang, Y., Kong, D., Wu, F., Yin, A., Qu, J., Liu, X.: Deep hybrid neural-like P systems for multiorgan segmentation in head and neck CT/MR images. *Expert Syst. Appl.*, **168**, 114446 (2021) [CrossRef]
56. Liang, S., Thung, K.H., Nie, D., Zhang, Y., Shen, D.: Multi-view spatial aggregation framework for joint localization and segmentation of organs at risk in head and neck CT images. *IEEE Trans. Med. Imaging*, **39**, 2794–2805 (2020) [CrossRef]
57. Gou, S., Tong, N., Qi, S., Yang, S., Chin, R., Sheng, K.: Self-channel-and-spatial-attention neural network for automated multi-organ segmentation on head and neck CT images. *Phys. Med. Biol.*, **65**, 245034 (2020) [CrossRef] [PubMed]
58. Tam, C., Yang, X., Tian, S., Jiang, X., Beitler, J., Li, S.: Automated delineation of organs-at-risk in head and neck CT images using multi-output support vector regression. In: *Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging; International Society for Optics and Photonics*, Bellingham, WA, USA, Vol. 10578, p. 1057824 (2018)
59. He, B., Jia, F.: Head and neck CT segmentation based on a combined U-net model. *J. Integr. Technol.* **9**, 17–24 (2020)
60. Qiu, B., Guo, J., Kraeima, J., Glas, H.H., Borra, R.J., Witjes, M.J., Ooijen, P.M.V.: Recurrent convolutional neural networks for mandible segmentation from computed tomography. *arXiv* 2020, [arXiv:2003.06486](https://arxiv.org/abs/2003.06486)
61. Zhu, W., Huang, Y., Tang, H., Qian, Z., Du, N., Fan, W., Xie, X.: AnatomyNet: Deep 3D Squeeze-and-excitation U-Nets for fast and fully automated whole-volume anatomical segmentation. *arXiv* 2018, [arXiv:1808.05238](https://arxiv.org/abs/1808.05238).
62. Tong, N., Gou, S., Yang, S., Cao, M., Sheng, K.: Shape constrained fully convolutional DenseNet with adversarial training for multiorgan segmentation on head and neck CT and low-field MR images. *Med. Phys.*, **46**, 2669–2682 (2019). [CrossRef]
63. Nikolov, S., Blackwell, S., Mendes, R., De Fauw, J., Meyer, C., Hughes, C., Askham, H., Romera-Paredes, B., Karthikesalingam, A., Chu, C., et al.: Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv* 2018, [arXiv:1809.04430](https://arxiv.org/abs/1809.04430)

64. Torosdagli, N., Liberton, D.K., Verma, P., Sincan, M., Lee, J., Pattanaik, S., Bagci, U.: Robust and fully automated segmentation of mandible from CT scans. In: Proceedings of the 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), Melbourne, Australia, pp. 1209–1212 (2017)
65. Qiu, B., van der Wel, H., Kraeima, J., Glas, H.H., Guo, J., Borra, R.J.H., Witjes, M.J.H., van Ooijen, P.M.A.: Automatic segmentation of mandible from conventional methods to deep learning-a review. *J. Pers. Med.* **11**(7), 629 (2021). <https://doi.org/10.3390/jpm11070629>. PMID:34357096;PMCID:PMC8307673
66. Manoharan, S.: Early diagnosis of lung cancer with probability of malignancy calculation and automatic segmentation of lung CT scan images. *J. Innov. Image Process. (JIIP)* **2**(04), 175–186 (2020)
67. Balasubramaniam, V.: Artificial intelligence algorithm with SVM classification using dermoscopic images for melanoma diagnosis. *J. Artific. Intell. Capsule Netw.* **3**(1), 34–42 (2021)

# Development of a Fully Convolutional Network for the Segmentation of Adipose Tissues on Abdominal MRI



B. Sudha Devi and D. S. Misbha

**Abstract** The excess accumulation of visceral adipose tissue (VAT) and subcutaneous adipose tissue (SAT) in the abdomen that causes obesity needs to be measured precisely for clinical evaluation of obesity. The availability of accurate and reliable imaging tools is needed to segment and quantify adipose tissues that cause serious chronic conditions. Computed tomography (CT) and magnetic resonance imaging (MRI) are the most commonly used imaging modalities to distinguish and quantify VAT and SAT. However, CT is subjected to ionizing radiation, and so, MRI is highly preferred. In this work, a fully automated deep learning model has been developed to segmentation VAT and SAT from MRI images of the abdomen. First, a fully convolutional network (FCN)-based U-Net architecture was used to separate SAT. In the second step, the K-means clustering algorithm was used to separate the adipose and non-adipose tissues, thereby extracting the VAT content. The proposed FCN-based method produced a Pearson correlation coefficient of 0.99 and 0.99 for VAT and SAT content. The results show that the new FCN method is highly accurate and reliable.

**Keywords** Obesity · Computed tomography · Magnetic resonance imaging · Segmentation · Deep learning · Fully convolutional network

## 1 Introduction

In humans, body fat is disseminated into two areas—visceral adipose tissue (VAT) and subcutaneous adipose tissue (SAT) [1]. Visceral adipose tissue borders the internal organs in the abdominal cavity, whereas subcutaneous tissue is seen under the skin. An excess distribution of VAT in the abdominal region is known as central obesity or

---

B. S. Devi (✉)

Department of Computer Science, Nesamony Memorial Christian College Affiliated To  
Manonmaniam Sundaranar University, Tirunelveli, India  
e-mail: [sudhanixen@gmail.com](mailto:sudhanixen@gmail.com)

D. S. Misbha

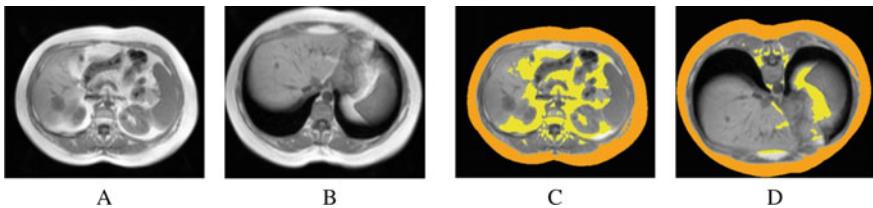
Department of Computer Applications, Nesamony Memorial Christian College Affiliated To  
Manonmaniam Sundaranar University, Tirunelveli, India

abdominal obesity, closely related to many health risks like type 2 diabetes, cardiovascular diseases, cancer, musculoskeletal disorders, and high cholesterol [2, 3]. Many methods exist to determine the distribution of body fat. In the past years, anthropometric measurements like body mass index (BMI), umbilicus circumference, waist circumference (WC), and waist-to-hip ratio (WHR) were considered as easy techniques to estimate the distribution of body fat [4, 5]. But the results were inconsistent due to the change of body position during measurement. Also, these methods do not differentiate SAT and VAT [6]. Therefore, there is a necessity for the advancement of accurate, specific, and reliable tools for the measurement of SAT and VAT.

Many imaging techniques like bioelectrical impedance, dual-energy X-ray absorptiometry, ultrasound, CT, and MRI have been proposed to measure abdominal fat distribution [7]. Among these, various research attempts have shown that MRI is a valuable and reliable tool to investigate and evaluate abdominal fat distribution. Several approaches for the segmentation of VAT and SAT have been proposed in the literature that includes manual, semi-automated, and fully automated techniques. Manual segmentation is the best choice, but it requires trained medical experts and long time span which is subjected to human errors. As VAT is irregular in shape and extensively distributed in the abdominal cavity, it is very difficult to segment VAT manually. To solve this problem, some semi-automated approaches were developed which consumes less time, but the results are based on the operator. Recently, various researchers have proposed several fully automated methods to segregate VAT and SAT and produced good results. The automated segmentation methods includes explicit models [8], random forests [9], multi-atlas segmentation model [10], and deep learning [11]. There exists various studies that demonstrates the accuracy of automated segmentation and quantification of adipose tissues using deep learning. In this work, it is proposed to improve adipose tissue segmentation by using a fully automated method that combines FCN-based U-Net and clustering method. The core of this fully automated segmentation method is the clustering in the peritoneal cavity avoids the misinterpretation of adipose and non-adipose tissues. U-Net is designed in the form of layers that performs pixelwise predictions of the whole image in a short duration of time. This provides rapid segmentation of adipose tissues for medical applications. Experimental results show that this automated method produces high segmentation accuracy with a Pearson correlation coefficient of 0.99 for VAT and SAT which is greater than existing deep learning-based approaches. An illustration of the segmentation of adipose tissues is shown in Fig. 1.

The main objective of this work is the development of an automated adipose tissue segmentation system using U-Net and K-means clustering algorithm to predict chronic diseases earlier. The U-Net architecture separates SAT in the abdominal region from MRI images. The adipose tissues are separated from the background, and thereby, the VAT content is isolated using K-means clustering algorithm.

The state-of-the-art segmentation methods are reviewed in detail in Sect. 2. The proposed U-Net based segmentation methodology is presented in Sect. 3. Experimental results and discussion is described in Sect. 4. Finally, the work is concluded in Sect. 5.



**Fig. 1** Illustration of delineation of adipose tissues. A and B Original T1-weighted MRI slices of the abdomen C and D Separation of adipose and non-adipose tissues—yellow-coloured area indicates VAT and orange-coloured area indicates SAT

## 2 Related Work

Deep learning with convolutional neural networks (CNNs) has become popular in recent years in areas such as image identification, classification, and segmentation. SAT and VAT regions were separated by Wang et al. [12] based on clustering method using water–fat separation MRI. First, adipose tissues were separated from the non-adipose tissues using K-means clustering algorithm. A deformable model initiated by a uniquely generated deformation field that points to the inner SAT boundary was used to differentiate SAT and VAT. The advantage of this method is the increased level of automation and objectivity, as well as less bias among different operators. However, testing was done only on a specific number of people with dyslipidaemia. Hui et al. [13] proposed a method using spoke template for separating SAT and VAT. Midpoint circle algorithm and Bresenham’s line algorithm were used to develop the spoke template which is round in shape. When the specified condition is met, VAT and SAT were isolated automatically. This method does not require any user intervention to produce the segmented output, but a portion of the VAT is left in the SAT results when the connecting area between SAT and VAT is thick. Langner et al. [14] introduced two fully convolutional network architectures U-Net and V-Net which were trained for the segmentation and quantification of SAT and VAT on MR images. The segmentation output produced by U-Net is highly reliable and accurate for the quantification of abdominal adipose tissues. The V-Net is less robust and results in oversegmentation of random patterns. Park et al. [15] performed abdominal muscle and fat segmentation using deep learning concept. A training dataset consisting of 467 subjects and 883 CT scans was used to develop a fully convolutional network (FCN) model. The network model resulted in high performance and accuracy, and external validation was carried out using data from a small number of subjects FCN generated accurate segmentation results, but in a few cases, the results were erroneous in the validated dataset related to SAT and muscle fat segmentation.

A graph theoretic segmentation algorithm was introduced by Sadanathan et al. [16] to segment VAT, deep (DSAT), and superficial (SSAT) from abdominal MRI images. The segmentation process is carried out in two steps. In the first step, graph cut method is used to separate SAT and VAT, and in the second step, a modified level set method is used to distinguish DSAT and SSAT. The segmentation algorithms

produced the results rapidly, and precise segmentation was done. The major drawback is that accurate segmentation of VAT and SAT compartments is difficult in lean subjects. The data was obtained from 44 normal and 38 overweight population for the segmentation. For the quantification of SAT and VAT from MRI, Ning Shen et al. [17] proposed a machine learning technique using IDEAL-IQ sequences. A deep neural network was trained first to obtain the SAT patterns. Then, VAT clustering is done using the AFK-MC<sup>2</sup> algorithm. Although this method is highly reliable, a considerable amount of training time is required for the successful completion of the segmentation process. Kim et al. [17] used a separation mask method to separate SAT and VAT from CT images. The unwanted space between a closed path and muscle area is reduced by the separation mask region. The shortest closed path was estimated by the Convex Hull algorithm. Although this method is accurate and reliable, there is a gap in SAT segmentation because the muscular fat near SAT is included as part of VAT. A 3D convolutional neural network (DCNet) was presented by Kustner et al. [18] for separating SAT and VAT on MRI images. Training and testing of DCNet on whole-body MRI data is done from the epidemiological patient databases of various multi-centres. It provides a robust architecture and can be generalized to different sequence of images. The network was not trained to differentiate bone marrow and other tissues, resulting in misclassifications. The accuracy produced is estimated as 98.4%. Estrada et al. [19] proposed a deep learning pipeline for the segmentation of VAT and SAT using DIXON MRI images. In this three step process, first, the localization of the abdominal area is done, next, the adipose tissue segmentation is done, and finally, aggregation of the previously generated label maps are done. In this method, the number of training parameters is less which is an added advantage. When the input images are of poor quality, segmentation reliability decreases. Samira Masoudi et al. [20] introduced a segmentation method for the separation of adipose tissues on MR images of the abdomen using cross modality adaptation and deep learning algorithms. In order to make the separation of adipose tissues easier, a cycle generative adversarial network (C-GAN) is used. C-GAN converts the MR images into CT format, and then, two U-Net models separate VAT and SAT. The basic limitation of this method is the loss of many features during the transformation of MR images to CT images.

Deep learning models require less pre-processing steps and depend on the availability of large-scale dataset. For the past few years CNNs have influenced the area of adipose tissue segmentation. In this work, a new method is proposed to identify and segment SAT and VAT. For the extraction of SAT pattern, a FCN-based U-Net architecture was used, followed by K-means clustering for classifying the abdominal region into adipose tissue and non-adipose tissue and then obtain the VAT pattern.

### 3 Proposed Methodology

This section covers in detail the structure and technique of the proposed U-Net-based VAT and SAT separation system and the benefits of using it. The suggested method

consists of two critical modules: image pre-processing and the development of a skip connection-based U-Net model for SAT and VAT segmentation.

### **3.1 Abdominal MR Images Pre-Processing**

Although deep learning approaches do not require pre-processing in MR image segmentation, image enhancement is required to minimise computing costs, eliminate false positives caused by MR image bias fields, and get a better result with limited training data. This proposed abdominal VAT segmentation method uniform sizing, Gaussian normalization and data augmentation methods are used for MR image pre-processing.

#### **Uniform sizing**

A change in the size and quality of the MR images can be observed when different configurations of MRI scanners are used for different clinical investigations. Consequently, when the size of the MR images is high, it takes a significant amount of time and computer power to construct a model. In this proposed study, the image is resized to a resolution of  $256 \times 256$  pixels before being trained and tested with the U-Net model.

#### **Gaussian normalization**

Due to variations in the MRI scanner, the brightness and contrast of the images are not uniform in all MR images. Furthermore, MRI scans are frequently obtained using a variety of collection techniques or methods. This has a significant impact on the deep learning model's VAT segmentation efficiency. In this study, Gaussian normalisation was utilised to reduce segmentation error by uniformizing the MR images' luminance and brightness used for the U-Net model's training and validation. Gaussian normalisation is accomplished through the use of the following formula.

$$I(x, y, z) = \frac{I(x, y, z) - \mu}{\sigma} \quad (1)$$

where  $I$  represents the pixel intensity values in MR images, and  $(x, y, z)$  denotes the pixel dimensions.  $\mu$  and  $\sigma$  represent the mean and standard deviation of all pixel dimensions, respectively.

#### **MR image augmentation**

The primary goal of image augmentation in deep learning models is to produce a more accurate segmentation or classification model with fewer training images. In this study, abdomen MR images are rotated artificially in a variety of directions, including horizontally, 180 degrees, 360 degrees, and vertically. Image augmentation multiplies the amount of training images by 400%.

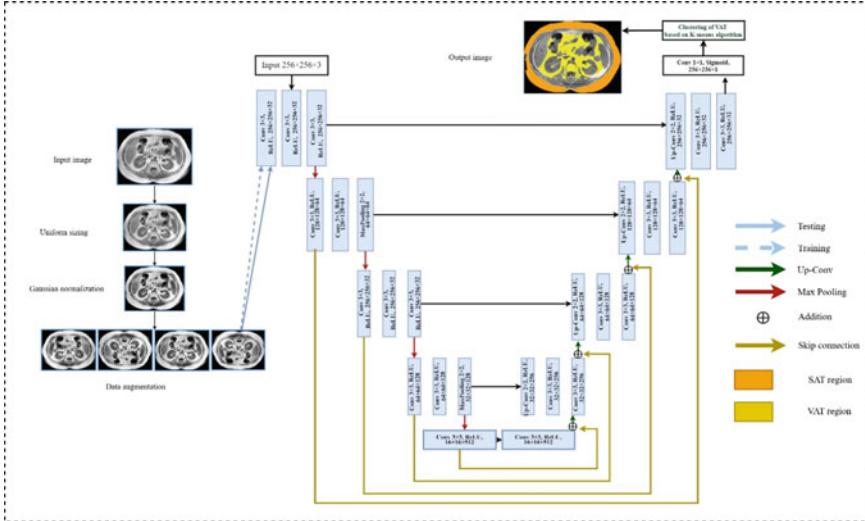
### 3.2 Skip Connection Based U-Net Architecture for SAT and VAT Segmentation

FCN architecture is universally utilised in clinical image processing to perform semantic segmentation. A FCN encodes image pixels into image classes using a convolutional neural network. FCNs are composed of convolutional, pooling, and upsampling layers. Images of varying sizes can be handled by FCNs, which is a notable feature of the technique. The convolution and pooling layers represent the objects in the image, whereas the up sampling layer indicates the area where the object is detected. A skip connection is employed in order to maintain the smooth representation of spatial data that has been lost during the pooling levels of the network. The U-Net architecture is believed to be the most effective FCN for doing end-to-end segmentation. The U-Net, which is formed from FCNs, has evolved into a fundamental network structure for biomedical image segmentation. Figure 1 depicts the complete flow diagram for the automated segmentation of SAT and VAT. Figure 1 depicts the overall flow diagram for automated SAT and VAT segmentation, including pre-processing, training, and testing. The proposed skip connection-based U-Net architecture is divided into three modules: encoding, decoding and bridge. The encoding layer receives pre-processed and augmented images as input. The encoding block is made up of four blocks, each of which has down convolutional layers, a non-linear activation function, and a max pooling layer. This encoding block's primary job is to extract high-level and low-level texture features from the input abdominal MR images. Then, high-level and low-level features with positive values are chosen and fed into the bridge layer. The bridge module connects the encoding and decoding modules. The bridge module has two normalisation units, two ReLU units, two convolutional units and a concatenation unit. Finally, the textural properties of the MRI fat region are extracted by the decoding unit.

A novel skip connection-based U-Net model is proposed to separate the SAT and VAT from abdominal MR images in this study. This improved U-Net architecture establishes skip connectivity between both the up convolution and down convolution parts to enhance convergence rate and VAT and SAT segmentation efficiency. It carries out convolution and cropping operations. The yellow arrow on the figure indicates this. The outputs of the down convolution (1st, 2nd, 3rd and 4th blocks) are connected to the input component of the up convolutional layer (7th, 8th, 9th and 10th block) in this architecture. The skip connection method can be calculated as follows:

$$S_C = f(W_{l-1}X_{l-1}) + X_{l-\text{nl}} \quad (2)$$

In Eq. 2,  $X$  represents the input,  $W$  represents the FCN's network weight, and  $f$  denotes the non-linear activation function (ReLU). The gradient descent (GD) algorithm is commonly used to train the FCN model. When using GD to train the FCN model, the gradient vanishing problem emerges. The proposed skip connection



**Fig. 2** The proposed U-Net-based system for automatic VAT and SAT segmentation, includes pre-processing, training and testing

model more effectively addresses the gradient vanishing problem. As a result, the proposed U-Net model's training efficiency has substantially increased.

Adam optimizer is used for training the input images and the related feature maps [21] for segmentation. Soft-max [22] over the final feature map and Dice loss function are combined to update the weighing values. Soft-max is defined as,

$$S_k(L) = \exp(m_k(L)) / \left( \sum_{k'=1}^k \exp(m_{k'}(L)) \right) \quad (3)$$

where  $m_k(L)$  indicates the activation at the pixel position  $L$  in feature channel  $k$ .  $S_k(L)$  is the maximum-function approximation. The Dice loss function is given by:

$$D(\theta) = \frac{2 \sum_{L \in \omega} p_{\theta}(L)P(L)}{\sum_{L \in \omega} p(L)^2 + \sum_{L \in \omega} p_{\theta}(L)^2} \quad (4)$$

where  $p(L)$  represents the pixel label, and  $\theta$  represents the feature map. The weighing values are calculated as:

$$rt = \nabla_{\theta} D_t(\theta_{t-1}) \quad (5)$$

$$st = \alpha_1 s_{t-1} + (1 - \alpha_1)r_t \quad (6)$$

$$\mathbf{u}_t = \alpha_2 \mathbf{u}_{t-1} + (1 - \alpha_2) \mathbf{r}_t^2 \quad (7)$$

Here,  $t$  represents the time period,  $s_t$  and  $u_t$  represent the first and second moment vectors. The rate of exponential decay of the first and second estimation of moments are  $\alpha_1 = 0.9$  and  $\alpha_2 = 0.99$ , respectively. The entire process flow of the proposed method is described in algorithm 1.

---

**Algorithm 1 VAT and SAT segmentation**


---

1. ***Input Abdominal MR Image dataset***
  2. ***Output Segmented VAT and SAT pixels***
  3. ***Model Training ( $U$  – Net)***
  4. ***Randomly initialize training variable  $U$  – Net with random weights***
  5. ***Accuracy***  $\leftarrow 0$
  6. ***For each***  $MR\text{Image} = 1, 2, \dots, n$  ***MR Image do***
  7.      $MR_{image} \leftarrow$  Uniform resizing(*Input MR Image*)
  8.      $MR_{image} \leftarrow$  Gaussian Normalization ( $MR_{image}$ )
  9.      $I(x, y, z) = \frac{I(x, y, z) - \mu}{\sigma}$
  10.     $MR_{image} \leftarrow$  Image augmentation( $MR_{image}$ )
  11.    *Input layer  $u(t)$  takes image data and sends it to the hidden layer*
  12.    ***Feature Extraction by Encoding Unit***
  13.    Highlevel texture features  $\leftarrow$  UNet( $MR_{image}$ )
  14.    Highlevel texture features  $\leftarrow$  UNet( $MR_{image}$ )
  15.    Generate Skip Connections using  $S_c = f(W_{l-1}X_{l-1}) + X_{l-nl}$
  16.    ***Decoding Unit***
  17.    Select the most relevant features
  18.    ***Training UNet model and Variable Updating***
  19.    Calculate error rate  $e(t)$
  20.    Update weight using back propagation
  21.    Minimize error rate  $e(t)$
  22.    ***End for***
- 

### 3.3 Clustering of VAT Based on K-means Algorithm

The automatic segmentation of VAT and the segregation of adipose and non-adipose tissues are carried out using K-means ( $K = 2$ ) clustering method based on the intensity of images. Some observations without labels can be clustered using the K-means clustering algorithm [23–25]; however, the total number of categories must be clearly represented, such as a total of  $K$  categories. It is necessary to find centre of clusters for each category which is the core of the algorithm. A rough seed point is needed to induce the algorithm. But it seems to be complex to get an adequate output when

random seed algorithm is used. Better clustering results are obtained using AFK-MC<sup>2</sup> algorithm without the distribution of data. It enhances the creation of initial seed point, and its clustering speed is also high. The VAT clustering is done in the peritoneal cavity to avoid misinterpretation of VAT. Several clusters are generated with varying signal intensity. The cluster whose average intensity was high are considered as VAT.

## 4 Experimental Results

This study includes 65 non-contrast MRI datasets that is collected at random from the prospective EISNER trial obtained at NIMS Medical Institute. The study cohort included people without any previous record of coronary heart disease but with diabetes and hypercholesterolemia. The population is represented in Table 1. An average of 55 transverse slices were used per scan, and each axial slice has a dimension of  $512 \times 512$  pixels of  $0.684 \text{ mm} \times 0.684 \text{ mm}$ . For the implementation of this proposed work, a workstation with Dual Intel Xeon Processor E5-2630 v2 (Six-Core HT, 2.6 GHz Turbo, 15 MB), 4 GB NVIDIA Quadro K5000,  $1 \times 8$  GB DDR3, 1 TB 7200RPM SATA, and Windows 10 Pro workstation is used as a platform. Python 3.7 and open-source machine learning libraries such as Keras 2.2.4, TensorFlow 1.13.1, and Scikit-learn 0.20.3 have been used for the development of working model.

In this method, the SAT results produced by the automated method are represented as orange-coloured area, and the VAT results are shown in yellow colour as represented in Fig. 3. In Table 2, the proposed method was compared with the existing literatures, and it is observed that the proposed method provides a high correlation coefficient (SAT—0.99 and VAT—00.99).

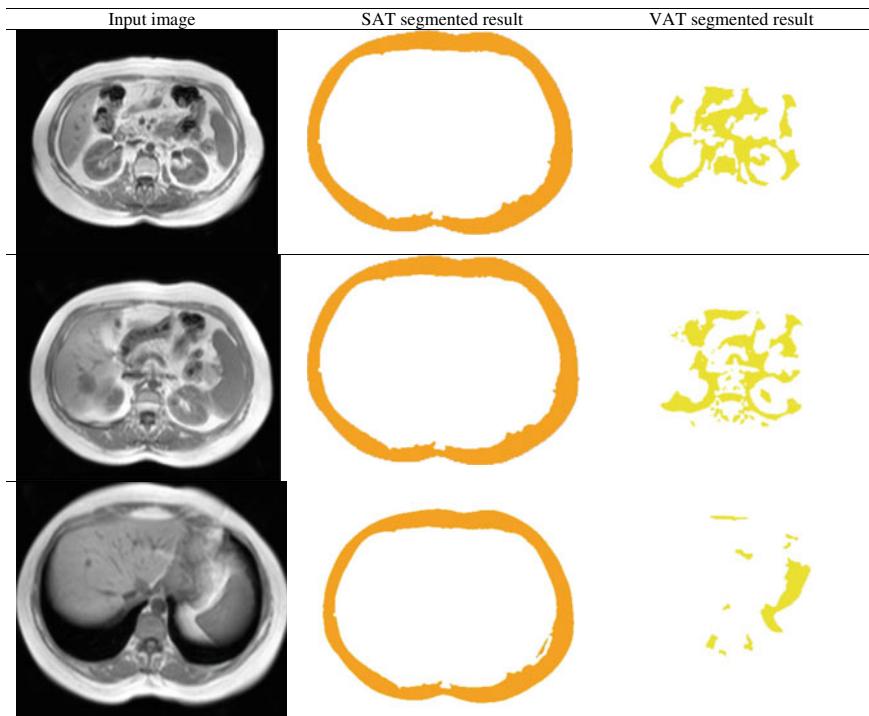
The suggested VAT and SAT segmentation model's overall performance is determined by its total accuracy, which is calculated using Eq. 3.

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (8)$$

This study used the FCN-based U-Net architecture and clustering method to automatically quantify VAT and SAT accurately. In VAT clustering, at certain times, the non-VAT areas in the peritoneal cavity shows a signal intensity similar to VAT. In this case, there is a chance for over estimation of VAT content, and so VAT clustering

**Table 1** Subject characteristics

Subjects	60
Male	40
Female	20
Age (years)	4–65
Diabetes	10
Hypercholesterolemia	35



**Fig. 3** Column1 represents the original image, column2 and column3 indicate the result of segmentation of SAT and VAT, respectively

**Table 2** Comparison performance with recent literatures

Author	Segmentation method	Total segmentation accuracy	
		VAT	SAT
Addeman et al.	Threshold-based	0.98	0.99
Shen et al.	Multi-atlas-based segmentation	0.97	0.96
Jiang et al.	Deep neural network	0.96	0.96
Ning Shen et al.	Deep neural network	0.99	0.99
Hui et al.	Mid-point circle algorithm, Bresenham's algorithm	0.87	0.98
Mendez et al.	Shape-based Otsu algorithm	0.97	0.98
Kuistner et al.	Convolutional neural network(CNN)	0.94	0.94
Estrada et al.	Dense fully convolutional network	0.85	0.97
Samira Masoudi et al.	U-Net model	0.94	0.97
Proposed method	U-Net model	0.99	0.99

is carried out in the peritoneal cavity contour. To avoid misinterpretation of VAT, the fat images were mapped to the water images depicted by U-Net peritoneal contour to obtain the same position.

## 5 Conclusion

Even though various approaches for VAT and SAT segmentation have been proposed, this deep learning and clustering-based segmentation method improves the efficiency and accuracy of segmentation. The experimental results show that there is a good correlation between the proposed and manual method. The automated segmentation of adipose tissue volumes are considered to be beneficial for the creation of new biomarkers of various disorders. This work helps the clinicians to access central obesity and the risks associated with it by segmenting VAT and SAT. Overall, the proposed deep learning-based approach could save the excess time that doctors spend on the segmentation of visceral adipose tissue.

## References

1. Al-Radaideh, A., Tayyem, R., Al-Fayomi, K., Nimer, N., Malkawi, A., Al-Zu’bi, R., Agraib, L., Athamneh, I., Hijjawi, N.: Assessment of abdominal fat using high-field magnetic resonance imaging and anthropometric and biochemical parameters. *Am. J. Med. Sci.* **352**, 593–602 (2016). <https://doi.org/10.1016/j.amjms.2016.09.009>
2. Hussein, S., Bagci, U., Green, A., Watane, A., Reiter, D., Chen, X., Papadakis, G.Z., Wood, B., Cypess, A., Osman, M.: Automatic segmentation and quantification of white and brown adipose tissues from PET/CT scans. *IEEE Trans. Med. Imaging.* **36**, 734–744 (2017). <https://doi.org/10.1109/TMI.2016.2636188>
3. Grainger, A.T., Krishnaraj, A., Quinones, M.H., Tustison, N.J., Epstein, S., Fuller, D., Jha, A., Allman, K.L., Shi, W.: Deep learning-based quantification of abdominal subcutaneous and visceral fat volume on CT images. *Acad. Radiol.* S107663320304268 (2020). <https://doi.org/10.1016/j.acra.2020.07.010>
4. Sri Kumar, T., Siegel, E.M., Gu, Y., Balagurunathan, Y., Garcia, A.L., Chen, Y.A., Zhou, J.-M., Zhao, X., Gillies, R., Clark, W., Gamenthaler, A., Choi, J., Shibata, D.: Semiautomated measure of abdominal adiposity using computed tomography scan analysis. *J. Surg. Res.* **237**, 12–21 (2019). <https://doi.org/10.1016/j.jss.2018.11.027>
5. Kucybala, I., Tabor, Z., Ciuk, S., Chrzan, R., Urbanik, A., Wojciechowski, W.: A fast graph-based algorithm for automated segmentation of subcutaneous and visceral adipose tissue in 3D abdominal computed tomography images. *Biocybernetics Biomed. Eng.* **40**, 729–739 (2020). <https://doi.org/10.1016/j.bbe.2020.02.009>
6. Wald, D., Teucher, B., Dinkel, J., Kaaks, R., Delorme, S., Boeing, H., Seidensaal, K., Meinzer, H., Heimann, T.: Automatic quantification of subcutaneous and visceral adipose tissue from whole-body magnetic resonance images suitable for large cohort studies. *J. Magn. Reson. Imaging.* **36**, 1421–1434 (2012). <https://doi.org/10.1002/jmri.23775>
7. Mattsson, S., Thomas, B.J.: Development of methods for body composition studies. *Phys. Med. Biol.* **51**, R203–R228 (2006). <https://doi.org/10.1088/0031-9155/51/13/R13>
8. Heimann, T., Meinzer, H.-P.: Statistical shape models for 3D medical image segmentation: a review. *Med. Image Anal.* **13**, 543–563 (2009). <https://doi.org/10.1016/j.media.2009.05.004>

9. Criminisi, A.: Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *FNT in Comput. Graph. Vision.* **7**, 81–227 (2011). <https://doi.org/10.1561/0600000035>
10. Decazes, P., Rouquette, A., Chetrit, A., Vera, P., Gardin, I.: Automatic measurement of the total visceral adipose tissue from computed tomography images by using a multi-atlas segmentation method. *J. Comput. Assisted Tomography.* **42**, 139–145 (2018). <https://doi.org/10.1097/RCT.0000000000000652>
11. Shen, D., Wu, G., Suk, H.-I.: Deep learning in medical image analysis. *28* (2017)
12. Wang, D., Shi, L., Chu, W.C.W., Hu, M., Tomlinson, B., Huang, W.-H., Wang, T., Heng, P.A., Yeung, D.K.W., Ahuja, A.T.: Fully automatic and nonparametric quantification of adipose tissue in fat–water separation MR imaging. *Med. Biol. Eng. Comput.* **53**, 1247–1254 (2015). <https://doi.org/10.1007/s11517-015-1347-y>
13. Hui, S.C.N., Zhang, T., Shi, L., Wang, D., Ip, C.-B., Chu, W.C.W.: Automated segmentation of abdominal subcutaneous adipose tissue and visceral adipose tissue in obese adolescent in MRI. *Magn. Reson. Imaging* **45**, 97–104 (2018). <https://doi.org/10.1016/j.mri.2017.09.016>
14. Langner, T., Hedström, A., Mörväld, K., Weghuber, D., Forslund, A., Bergsten, P., Ahlström, H., Kullberg, J.: Fully convolutional networks for automated segmentation of abdominal adipose tissue depots in multicenter water–fat MRI. *Magn. Reson. Med.* **81**, 2736–2745 (2019). <https://doi.org/10.1002/mrm.27550>
15. Park, H.J., Shin, Y., Park, J., Kim, H., Lee, I.S., Seo, D.-W., Huh, J., Lee, T.Y., Park, T., Lee, J., Kim, K.W.: Development and validation of a deep learning system for segmentation of abdominal muscle and fat on computed tomography. *Korean J Radiol.* **21**, 88 (2020). <https://doi.org/10.3348/kjr.2019.0470>
16. Sadanathan, S.A., Prakash, B., Leow, M.K.-S., Khoo, C.M., Chou, H., Venkataraman, K., Khoo, E.Y.H., Lee, Y.S., Gluckman, P.D., Tai, E.S., Velan, S.S.: Automated segmentation of visceral and subcutaneous (deep and superficial) adipose tissues in normal and overweight men: Automated Segmentation of Adipose Tissue. *J. Magn. Reson. Imaging.* **41**, 924–934 (2015). <https://doi.org/10.1002/jmri.24655>
17. Kim, Y.J., Park, J.W., Kim, J.W., Park, C.-S., Gonzalez, J.P.S., Lee, S.H., Kim, K.G., Oh, J.H.: Computerized automated quantification of subcutaneous and visceral adipose tissue from computed tomography scans: development and validation study. *JMIR Med Inform.* **4**, e2 (2016). <https://doi.org/10.2196/medinform.4923>
18. Küstner, T., Hepp, T., Fischer, M., Schwartz, M., Fritzsche, A., Häring, H.-U., Nikolaou, K., Bamberg, F., Yang, B., Schick, F., Gatidis, S., Machann, J.: Fully automated and standardized segmentation of adipose tissue compartments by deep learning in three-dimensional whole-body mri of epidemiological cohort studies. *30* (2020)
19. Estrada, S., Lu, R., Conjeti, S., Orozco-Ruiz, X., Panos-Willuhn, J., Breteler, M.M.B., Reuter, M.: FatSegNet: a fully automated deep learning pipeline for adipose tissue segmentation on abdominal dixon MRI. *Magn Reson Med.* **83**, 1471–1483 (2020). <https://doi.org/10.1002/mrm.28022>
20. Masoudi, S., Anwar, S.M., Harmon, S.A., Choyke, P.L., Turkbey, B., Bagci, U.: Adipose tissue segmentation in unlabeled abdomen mri using cross modality domain adaptation. In: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 1624–1628. IEEE, Montreal, QC, Canada (2020). <https://doi.org/10.1109/EMBC44109.2020.9176009>
21. Tripathi, M.: Analysis of convolutional neural network based image classification techniques. *JIIP.* **3**, 100–117 (2021). <https://doi.org/10.36548/jiip.2021.2.003>
22. Dr. Manoharan, S., Sathish: Improved version of graph-cut algorithm for CT images of lung cancer with clinical property condition. *JAICN.* **2**, 201–206 (2020). <https://doi.org/10.36548/jaicn.2020.4.002>
23. Balasubramaniam, V.: Artificial intelligence algorithm with SVM classification using dermoscopic images for melanoma diagnosis. *JAICN.* **3**, 34–42 (2021). <https://doi.org/10.36548/jaicn.2021.1.003>

24. Babiker Hamdan, P. K., Sathish, Y.: Early prediction of autism spectrum disorder by computational approaches to fMRI analysis with early learning technique. JAICN. **2**, 207–216 (2020). <https://doi.org/10.36548/jaicn.2020.4.003>
25. Sungheetha, Dr. A., Sharma, R., Dr.R.: GTIKF-gabor-transform incorporated K-means and fuzzy C means clustering for edge detection in CT and MRI. JSCP. **2**, 111–119 (2020). <https://doi.org/10.36548/jscp.2020.2.004>

# Improvement of Clarity for Foggy/Dusty Weather Images Using Triple Threshold Method



**Minu Inba Shanthini Watson Benjamin, N. S. Kalyan Chakravrthy, Badetti Syam, R. Navaneethakrishnan, Jee Joe Michael, and J. N. Swaminathan**

**Abstract** The visual quality of collected photographs can be greatly reduced by inclement dusty weather, which makes it hard to notice crucial image details. Image capture in such conditions frequently results in unwanted artifacts such as weak contrast, insufficient hues, or color cast. As a result, several strategies for processing such undesirable events and recovering lucid outcomes using appropriate colors have been developed. Because of the differences in the processing ideas used, these methods range from basic to complex. This paper introduces a new technique for processing poor quality photos acquired in adverse dusty weather that utilizes customized fuzzy intensification operators. Several tests were conducted to assess the proposed techniques processing ability, with the results demonstrating its ability to filter diverse degraded images. It was particularly good at delivering acceptable and revealing fine details in the processed photographs.

**Keywords** Dusty weather · Fuzzy intensification operators · Membership function

---

M. I. S. W. Benjamin

Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, India

N. S. K. Chakravrthy · J. N. Swaminathan (✉)

QIS College of Engineering and Technology, Ongole, Andhra Pradesh 523272, India  
e-mail: [swaminathan@qiscet.edu.in](mailto:swaminathan@qiscet.edu.in)

B. Syam

Mandava Institute of Engineering and Technology, Jaggayyapeta, Andhra Pradesh, India

R. Navaneethakrishnan

Kumaraguru College of Technology, Chinnavaddampati, Coimbatore, India

J. J. Michael

Saveetha School of Engineering, SIMATS, Chennai 602105, India

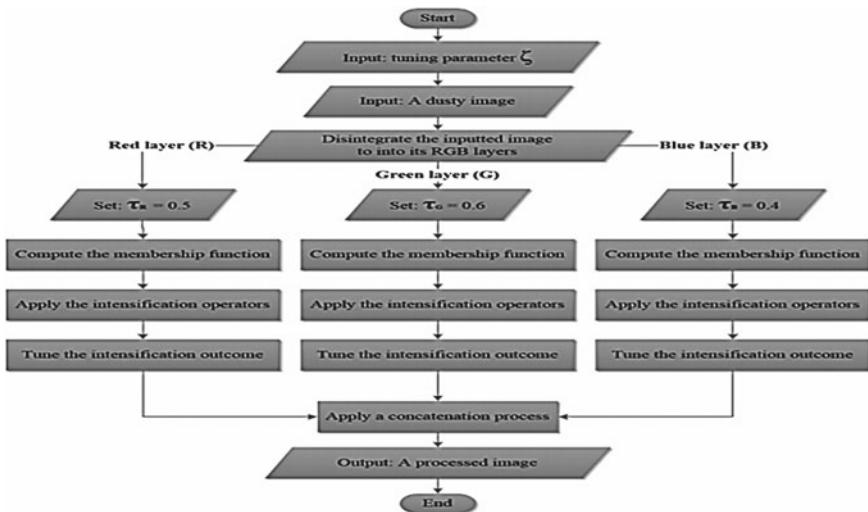
## 1 Introduction

AP Singh et al. proposed the (CAP) color attenuation prior algorithm which utilize depth map to figure the spread and sophisticated with the bilateral filtering and reestablish scene radiance by the advanced transmission map and ambient light [1]. The method of transforming a crisp input value to a fuzzy value with knowledge base information is known as fuzzification [2]. Although numerous types of curves can be found in the literature, the most frequent member functions utilized in the fuzzification process are Gaussian, triangular, and trapezoidal membership functions [3]. Embedded controllers can readily implement these types of membership functions [4]. The aim is to strengthen the image quality of photographs shot in dusty conditions [5]. This is accomplished using a novel adjustment method using tri-threshold fuzzy intensification operators [6]. Extensive trials with various real-world degraded photos were carried out to assess the efficiency of the new framework [7]. There are various techniques like image denoising and deblurring, illumination and color contrast improvement with edge sharpening are used to improve the clarity of the image [8]. There is a brief survey of the literature on the usage of fuzzy intensification operators [9]. Many researchers have researched and used such operators to tackle various difficulties in the path of digital images for processing [10]. All these studies have the same goal in mind to recover significant features from degraded photos [11]. Since capturing photographs in such conditions can have a substantial impact on image quality, severe dusty weather can have a significant impact on image quality weather frequently causes negative consequences, such as color cast, weak contrast, and insufficient color [12, 13].

## 2 Proposed Method

Because of the presence of various real-life challenges, the level of complexity of the suggested research tasks ranges from easy to difficult. All of these efforts are aimed on retrieving important data while processing corrupted photos. The proposed improvement method is described in great depth. As a result, a quick rundown of the processing framework is given. The current technique begins with the produced image color fidelity change using ( $\zeta$ ), a tuning parameter.

The corrupted image then sent to the computer and decomposed into its three basic channels: RGB. Two aspects of the operators of amplification should be evaluated. The first step is to calculate the tau ( $\tau$ ) parameter, which represents the operator's limits for image segmentation. The operator's capacity to handle image pixels is improved by the usage of tau. This investigates proposed enhancement technique in depth. As a result, it provides a quick overview of the processing framework in use. The proposed method begins to receive ( $\zeta$ ) that controls the comfortable experience of the input images. The corrupted image will be input and split into red, green, and blue main channels (RGB). Two aspects are needed to measure the intensification



**Fig. 1** Block diagram of the proposed method to remove foggy/dusty images

operators. The first step is to calculate tau ( $\tau$ ); it indicates the operator's thresholding limits. The operator's ability to process image pixels is enhanced using ( $\tau$ ). A membership function is required in the second step as it changes the default range of pixel values in a device to zero to one. This function has to be implemented in order for the intensification operators to function properly. Each channel's membership function is calculated as in Fig. 1.

Digital image processing has evolved into a powerful tool for processing of signal across a several applications. Image processing is used in most of the applications such as grayscale alteration, earth sciences, diagnostic imaging, remote sensing, and biometric identification. Various techniques in the field of digital image processing have been developed over the last four to five decades.

## 2.1 Algorithm Steps for Methodology

1. Start
2. Read input image
3. Disintegrate the input image in to RGB layers
4. Following this, calculate the membership function for each layer
5. Apply the intensification operators for each layer
6. Tune the intensification outcome for each layer
7. Apply a concatenation process for RGB layers
8. We will get the output image
9. Stop.

## 2.2 Calculation of the Intensification Operators Using Membership Function

Two variables must be considered while evaluating intensification operators. The first step is to look at a parameter called tau ( $\tau$ ); it indicates thresholding boundaries of the operator. The operator's capacity to process image pixels is improved by using ( $\tau$ ).

As it resets the values of pixels in the channel to the default range from zero to one, membership functions are also necessary. For each channel, the membership function is calculated as follows (Fig. 2),

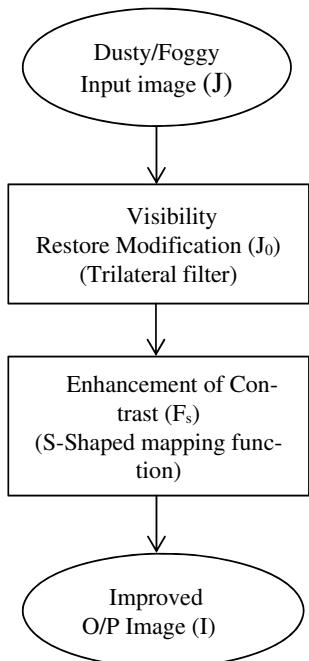
$$F_r = \frac{[r - \min(r)]}{[\max(r) - \min(r)]} \quad \text{for Red Component} \quad (1)$$

$$F_g = \frac{[g - \max(g)]}{[\max(g) - \min(g)]} \quad \text{for Green Component} \quad (2)$$

$$F_b = \frac{[b - \max(b)]}{[\max(b) - \min(b)]} \quad \text{for Blue Component} \quad (3)$$

In numerous image processing submissions, fuzzy intensifying operatives are useful to progress the contrast or color reliability of an input images. The

**Fig. 2** Flowchart of the proposed method to remove foggy/dusty images



intensification operators for each channel are represented as below,

$$K_R = \begin{cases} \{2 * (f_R(x, y))^2, & \text{if } f_R(x, y) \leq \tau_R \\ 1 - 2 * (1 - f_R(x, y))^2, & \text{otherwise} \end{cases} \quad (4)$$

$$K_G = \begin{cases} \{2 * (f_G(x, y))^2, & \text{if } f_G(x, y) \leq \tau_G \\ 1 - 2 * (1 - f_G(x, y))^2, & \text{otherwise} \end{cases} \quad (5)$$

$$K_B = \begin{cases} \{2 * (f_B(x, y))^2, & \text{if } f_B(x, y) \leq \tau_B \\ 1 - 2 * (1 - f_B(x, y))^2, & \text{otherwise} \end{cases} \quad (6)$$

After executing these operators, the result for every channel is adjusted using the recommended tuning process, that can be described as

$$u_R = (k_R)^{\tau_R + \xi} \quad (7)$$

$$u_G = (k_G)^{\tau_G + \xi} \quad (8)$$

$$u_B = (k_B)^{\tau_B + \xi} \quad (9)$$

The outputs uR, uG, and uB are then concatenated to generate the colorful image that depicts the proposed technique's end outcome. The formation of dusty mineral deposit particles is proof of a mystical interplay of processes. A passport of dusty creation must explain the entry and exit of a weigh at the earth's surface. In the scenario of dusty forming the cloud-caps of hills, the practically standard practice of cloud-formation, the cooling of shining relay merit to level, cannot be invoked for dusty conception crack down on.

### 3 Result and Analysis

The model is implemented only for dusty images, and the same is applicable to foggy images also. The S-mapping function used to improve the color contrast of the image, and the trilateral filter is utilized in restoring process (Figs. 3 and 4).

### 4 Conclusion

A novel fuzzy logic-based visibility processing method. This article explains how to improve the quality of images that have degraded due to dusty weather shot during an emergency. The suggested method uses a basic membership function to identify the values of pixels in the range of zero to one for a certain device, as well as fuzziness



**Fig. 3** Captured dusty image and clear image without dust effect



**Fig. 4** Captured foggy image and clear image without fog effect

intensification operators that are utilized depending on the situation and are based on numerous thresholds values, and a new adjustment approach. The proposed method provided satisfactory result with refined color and lucid qualities, according to the findings of the testing. Original and processed images are compared visually, and also interpretation of the histograms is provided for every image, were also used to discover. Finally, this technique is thought to be applicable to other deteriorating images, which are taken in foggy, hazy, or misty weather.

## References

1. Parihar, A. S., Gupta, G.: "Prior based single image dehazing using decision image." In: 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 960–965 (2020)
2. Yan, T., Wang, L., Wang, J.: Method to enhance degraded image in dust environment. *J. Softw.* **9**(10), 2672–2677 (2014)
3. Narasimhan, S., Nayar, S.: Contrast restoration of weather degraded images. *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(6), 713–724 (2003)

4. Huang, S., Ye, J., Chen, B.: An advanced single image visibility restoration algorithm for real-world hazy scenes. *IEEE Trans. Industr. Electron.* **62**(5), 2962–2972 (2015)
5. Chen, B., Huang, S.: An advanced visibility restoration algorithm for single hazy images. *ACM Trans. Multimed. Comput. Commun. Appl.* **11**(4), 1–21 (2015)
6. Huang, S.: An advanced motion detection algorithm with video quality analysis for video surveillance systems. *IEEE Trans. Circuits Syst. Video Technol.* **21**(1), 1–14 (2011)
7. Huang, S., Chen, B., Cheng, Y.: An efficient visibility enhancement algorithm for road scenes captured by intelligent transportation systems. *IEEE Trans. Intell. Transp. Syst.* **15**(5), 2321–2332 (2014)
8. Huang, S., Do, B.: Radial basis function based neural network for motion detection in dynamic scenes. *IEEE Trans. Cybernetics* **44**(1), 114–125 (2014)
9. Chacon, M., Gonzalez, S.: An adaptive neural-fuzzy approach for object detection in dynamic backgrounds for surveillance systems. *IEEE Trans. Industr. Electron.* **59**(8), 3286–3298 (2012)
10. Zhang, X., Hu, W., Chen, S., Maybank, S.: Graph embedding-based learning for robust object tracking. *IEEE Trans. Industr. Electron.* **61**(2), 1072–1084 (2014)
11. Kaur, K., Gupta, N.: Performance evaluation of modified DBLA Using dark channel prior & CLAHE. *Int. J. Intell. Syst. Appl.* **7**(5), 48–56 (2015)
12. Wang, J., Pang, Y., He, Y., Liu, C.: “Enhancement for Dust-sand storm images.” In: *Lecture Notes in Computer Science (Multi Media Modeling)*, (eds.) Springer International Publishing, pp. 842–849 (2016)
13. Mohamad, A.: A new image contrast enhancement in fuzzy property domain plane for a true color images. *Int. J. Signal Process. Syst.* **4**(1), 45–50 (2016)

# The Olympic Gold Medalists on Instagram: A Data Mining Approach to Study User Characteristics



Amirhosein Bodaghi and Jonathan J. H. Zhu

**Abstract** Olympic champions have been real idols for a significant portion of society, and by the advent of social media, their influence has increased rapidly. Despite their impact, they have been less studied. A primary step to grasp their cybercharacter is to examine their Instagram characteristics with possible gender differences and correlations between these characteristics. By applying a data-driven approach, this study utilizes a content analysis method to analyze photos of Olympic gold medalists on Instagram. In this vein, male gold medalists show a monotonously positive relationship between their following/follower ratio and the engagement/follower ratio. Also, the ratio of self-presentation turned out to have a solid monotonous negative relationship with age in both male and female gold medalists, which even takes a linear form in men. In line with the related theories and literature, these findings can help athletes manage and grow their brand on social media.

**Keywords** Instagram · Self-presenting · User characteristics · Olympic · Gold medalists · Data mining

## 1 Introduction

In the realm of ubiquitous computing [1] and cyberspace, the sports world has been taking a stand. Unlike traditional media, social media allows athletes to express themselves in the way they prefer most. By sharing content on social media, athletes interact with their audiences, which leads to active engagement from the individuals [2]. Rui and Stefanone [3] found that users who strategically manage their online self-presentation are those who care most about public evaluations. Many, as well as athletes, may belong to this category. Karg and Lock [4] recognized the importance of fan communities and brand awareness to sport entities. Lebel and Danylchuk [5]

---

A. Bodaghi (✉) · J. J. H. Zhu

Department of Media and Communication, City University of Hong Kong, Hong Kong, Hong Kong

e-mail: [abodaghi@cityu.edu.hk](mailto:abodaghi@cityu.edu.hk)

stated that social media had conferred a platform for athletes to express themselves, so they must be aware of how they present themselves in cyberspace. Also, DeAndrea and Walther [6] analyzed Facebook users' self-presentations and reached the end that each post a person shares could promote or ruin that person's cybercharacter. Also, there are some famous athletes whose page on social media does not receive much attention from public to make any personal brand [7, 8].

By the expansion of visual aspects of social media, interactive involvement would be yielded by sharing media [9]. Social media confers the freedom to people to decide what they want to see and which identity they like to have [10]. Pegoraro and Jinnah [11] and Hambrick and Kang [12] claim that athletes who properly utilize social media as a relationship marketing tool to create their online brand would gain brand loyalty. The widespread use of Instagram for marketing proves the importance of further investigations of this media in different domains, particularly in the sport context.

Hence, this research studies the Instagram characteristics of the athletes to recognize the gender discrepancies and relevancies. In particular, the study focuses on these research questions:

RQ1: How do women and men gold medalists differ from each other in terms of self-presentation on Instagram, and what characteristics influence this trait?

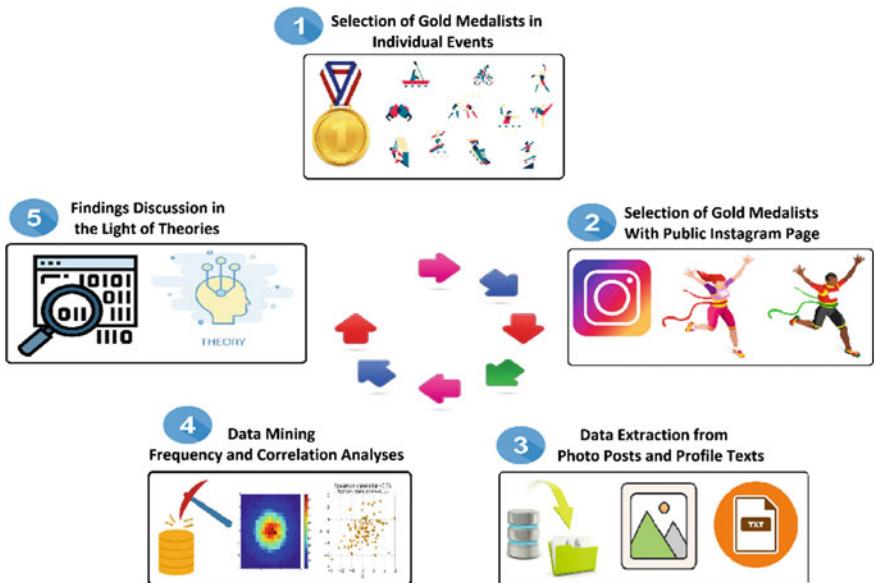
RQ2: What characteristics significantly impact the engagement rate from the gold medalists' followers on Instagram?

The findings will be explained by social theories and compared with the literature. The athletes may utilize the implications to develop their branding on online social networks. Furthermore, as seen from the literature [13, 14], the behaviors of users have deep roots in past activities. This fact indicates the importance of user characteristics whose epiphany can be seen in recent rumor-spreading models [15–18].

To continue, after presenting the literature, the study begins with data extraction. Then, the authors perform a gender breakdown of the Olympic gold medalists' user characteristics such as following/follower, engagement/follower, age, and the ratio of self-presenting posts. The results will be shown by 2D histograms that reveal the differences between men and women regarding the frequency of user characteristics' values. Figure 1 shows the conceptual illustration of the stages in the study.

## 2 Literature Review

Regarding sociotechnical aspects of using social networks, Instagram has been subject to a series of research. Souza et al. [19] studied the possible patterns out of people activities related to selfie sharing at the aggregate level. Pittman and Reich [20] investigated the relationship between the use of image-based media and users' loneliness. Hochman and Schwartz [21] focused on cultural analytics visualization



**Fig. 1** Conceptual illustration of the stages in the study

techniques to study Instagram images, to trace cultural visual rhythms. Silva et al. [22] characterized users' activities related to photo sharing.

Research showed that athletes perform behind-the-scenes performances to discuss their personal lives and get in touch with people [5, 23]. Furthermore, the authors examined gender-specific differences in self-expression and found no differences between the sexes. Research on athletes' self-portraits on social media has so far focused on Twitter and Facebook. Geurin-Eagleman and Clavio [24] examined the Facebook accounts of specialist athletes labeled as those who are not treated in the media and who must rely on their efforts to be seen. Conventional athletes are characterized as those who have reported in the mainstream media. Geurin and Burch [25] conducted a gender analysis on athletes' self-presentation activities. Zillich and Riesmeyer [26] examined the relative importance of personal, descriptive, and omission norms for adolescents' self-portraits on Instagram. They found that teens incur thoughtful norm violations when faced with conflicting norms of self-expression. Stsiampkouskaya et al. [27] focused on how differences between expected and received feedback could affect photo-sharing behavior, noting that when choosing a photo to post, people often take the perspective of their audience and consider whether the selected image generates enough interest to generate feedback.

The literature shows the existence of a gap in research on athletes' characteristics related to age and gender. However, the findings based on a measurement study prove the existence of such relationships would aid further investigations about the athletes' cybercharacters.

### 3 Methods

This study utilizes a content analysis method. The content analysis has shown to be a reproducible method [28]. According to the official Web site of the Olympics in Rio2016, 213 individuals won the gold medals of the individual events. However, only 149 (85 men and 64 women) gold medalists had a public page on Instagram.

The team sports are excluded. The information gathered from each individual gold medalist is as follow: 1—#posts; 2—#followers; 3—#followings; 4—#the max number of likes and comments among the last ten posts; 5—#self-presenting posts from the previous ten posts; 6—#pure self-presenting posts (posts in which the champion stands alone) from the previous ten posts; 7—#gender; and 8—#age. The non-photo posts were excluded. Also, all posts in which the user's face is recognizable were counted as self-presenting posts and those with the athletes as the only person in the photo were counted as pure self-presenting posts. The data extraction was done in four days, all by human referees. The code and data are available to the public.

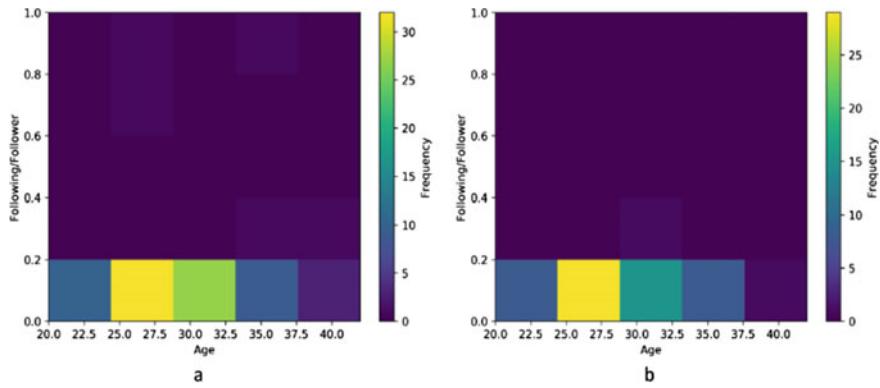
## 4 Results

### 4.1 *Gender Breakdown of the Characteristics*

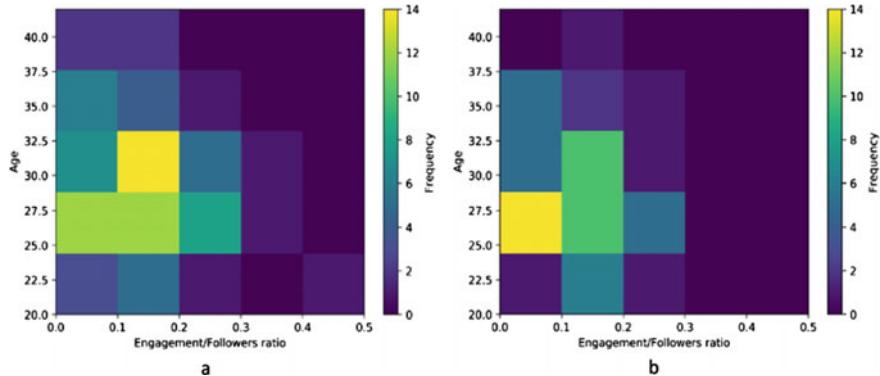
In this section, the goal is to find any difference between the characteristics of men and women gold medalists. To have the measured characteristics more comparable, they are used in a ratio instead of applying them merely on their own. Thus, the characteristics used in this analysis would be as follows: 1—the ratio of following/follower, 2—the ratio of engagement (like + comment)/follower, 3—the ratio of self-presenting posts among the last ten photo posts, 4—the ratio of pure self-presenting posts among the last ten photo posts, 5—the ratio of pure self-presenting posts/self-presenting posts, 6—age, and 7—gender.

To have better comparisons of the data in an apparent compact form, two-dimensional histograms are applied. Figure 2 shows the ratio of following/follower and age in both men and women gold medalists. As shown, the value of following/follower less than 0.2 is dominant for both men and women. This fact is no surprise since champions are the center of attention, and their accounts achieve lots of followers without any obligation to follow back the followers. Moreover, it can be seen that champions of 24–33 years of age have almost swept the gold medals for both men and women events. However, it should be considered that this research was conducted years after the Olympic Rio2016, so those champions took gold medals when they were younger.

When the relation of age with the ratio of engagement/follower is examined, it turns out that men champions at the age of 28.5–33 attain the most engagement out of their followers, while the same happens for women champions at the age of 24–28.5 (Fig. 3). Furthermore, almost less than 0.3 of their followers bother to



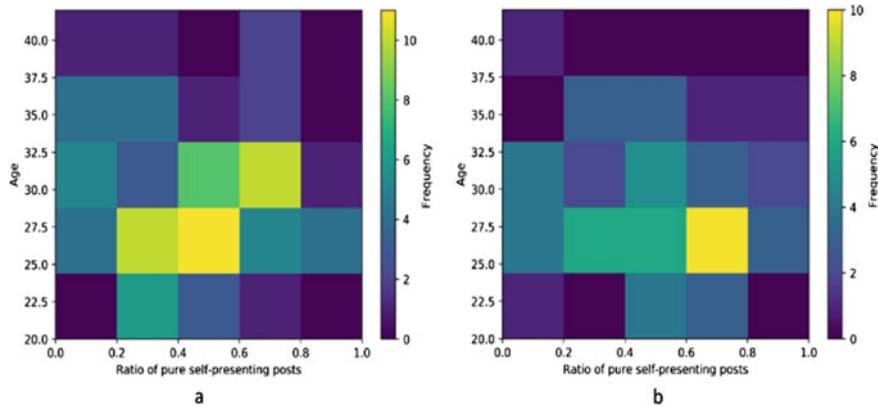
**Fig. 2** 2D histogram of following/follower and age. **a** for men and **b** for women. The squares with warm colors denote areas with more frequency in the dataset. A similar pattern can be seen for men and women since most gold medalists have a following/follower ratio less than 0.2 while their ages fall in the range of 24–33



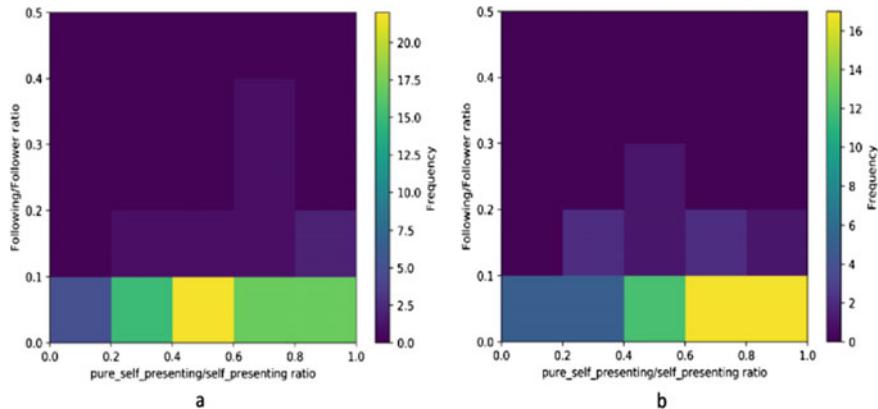
**Fig. 3** 2D histogram of engagement/followers and age. **a** for men and **b** for women. The squares with warm colors denote areas with more frequency in the dataset. Most of the men gold medalists are at the age of 29–33 with an engagement/followers ratio between 0.1 and 0.2, while their female counterparts mostly are at the age of 24–29 with an engagement/followers ratio between 0 and 0.1

like or comment on their posts for both men and women champions. Thus, finding champions who get a higher level of engagement would not be easy.

With the age characteristic, it is possible to compare athletes by a gender-based approach in terms of the ratio of pure self-presenting posts (Fig. 4). By looking at the most crowded range, i.e., 24 to 33, it can be seen that men champions tend to share less pure self-presenting posts than women do. This conclusion could also be drawn saliently from the 2D histogram of the ratio of engagement/follower and ratio of pure self-presenting posts/ self-presenting posts (Fig. 5). Indeed, when it comes



**Fig. 4** 2D histogram of the ratio of pure self-presenting posts and age. **a** for men and **b** for women. The squares with warm colors denote areas with more frequency in the dataset. The ratio of pure self-presenting posts for men reaches its most frequency at the range of 0.4–0.6 while the same for women befalls at the range of 0.6–0.8



**Fig. 5** 2-D histogram of the ratio of pure self-presenting/self-presenting and following/follower. **A** for men and **b** for women. The squares with warm colors denote areas with more frequency in the dataset. The higher ratio of pure self-presenting/self-presenting posts for women indicates they share more pure self-presenting posts out of their self-presenting posts than men do

to self-presenting posts of Olympic champions, chances are women be alone in the photo while men stand with their friends or fans.

## 4.2 Correlations Between the Characteristics

In this section, the possible relationships between different characteristics of gold medalists are examined. To this aim, we computed Pearson and Spearman correlations for each pair of characteristics in three categories, 1—general (both men and women), 2—only men, and 3—only women. Table 1 shows the achieved results, including coefficient correlations and p values for each pair of characteristics. The significant relationships are bold in the table.

The acceptable value of Spearman correlation between the characteristics of following/follower and engagement/follower for general users indicates a significant positive monotonic relationship. Women do not take any share since all their weight falls on the men's side. Moreover, the lack of any Pearson correlation rejects the premise of being linear for this relationship.

Significant relationships emerge from both Pearson and Spearman correlations when it comes to the characteristics of age and ratio of self-presenting posts. The negative coefficients of these correlations accompanied by small p values indicate a remarkable inverse linear and monotonic relationship between age and ratio of self-presenting posts. These strong relationships can be seen in both men and women categories, too, except for women in Spearman correlation in which the p value remains above 0.05 and ushers a non-monotonic linear correlation between their age and the ratio self-presenting posts.

Aside from the strong relationships mentioned above, if a slight bit of violation from the bound of (for instance) be tolerated, other considerable relationships which may not be as strong as aforementioned would be found. The correlations' coefficients of these relationships are underlined in Table 1. First, a negative Spearman correlation can be seen between the characteristics of following/follower and the ratio of pure self-presenting posts, which signals an inverse monotonic relationship

**Table 1** Spearman and Pearson correlations between the characteristics. The significance of the values presented at three levels \* <0.1, \*\* <0.05, and \*\*\* <0.01

		Ratio of self-presenting posts		Ratio of pure self-presenting posts	
		Pearson	Spearman	Pearson	Spearman
Following/follower	General	−0.093	−0.112	−0.028	−0.062
	Women	−0.131	−0.207	−0.162	−0.226*
	Men	−0.117	−0.031	0.008	0.030
Engagement/follower	General	0.054	0.098	0.058	0.055
	Women	0.034	−0.048	−0.077	−0.052
	Men	0.056	0.190*	0.141	0.150
Age	General	−0.392***	−0.276***	−0.156**	−0.113
	Women	−0.352***	−0.107	−0.224*	−0.194
	Men	−0.449***	−0.402***	−0.088	0.050

between them. Second, it can be seen there is a positive Spearman correlation between the characteristics of following/follower and the ratio of pure self-presenting posts, which indicates a positive monotonic relationship between them. Also, there is a negative Spearman correlation between age and ratio of pure self-presenting posts in the general category, which has been inherited from the women gold medalists.

The salient findings of this section can be summed as follows:

- F1: Women gold medalists tend to share more self-presenting posts than men gold medalists, but their rate of pure self-presenting posts is higher.
- F2: The ratio of engagement/follower monotonically and positively is related to the ratio of following/follower for men gold medalists.
- F3: The ratio of self-presenting posts monotonically and negatively is related to the age for gold medalists, which even takes a linear form for men gold medalists.

## 5 Discussion

In this section, the outcomes are explained by related social theories.

### 5.1 *Gender Differences of Self-Presenting and Theory of Evolutionary Psychology (F1)*

The first finding indicates the higher rate of self-presenting photos for women gold medalists than their male counterparts. This finding is aligned closely with the result of similar research about Olympic athletes conducted by Geurin-Eagleman and Burch [25]. They showed that female athletes appear on their Instagram photos more often than male athletes. In contrast, male athletes are more successful in inspiring their fans than female athletes. However, their results were limited to just eight selected Olympians. Similar findings have been achieved from studying on the athletes' self-presentation [29]. Evolutionary psychology, a theory that contains many assumptions about gender differences, explains this result. The evolutionary theory is based on natural selection, which claims that survival traits are more likely to be passed to the next generations [30]. On the other hand, the term self-concept exists, defined as a set of attributes, attitudes, and values that individuals believe define themselves [31].

Women have developed skills to better make connections while men went for domination [32]. In addition, a series of studies have shown women tend to be passive in their sports photos which emphasize the tradition in gender roles [33, 34]. Thus, along with their defined self-concept, women tend to convey emotions, so they exhibit more facial cues, which requires a close shot (mostly pure self-presenting posts) or at least a photo with friends, family, or fans (mostly self-presenting posts). While, men tend to present their dominance and possessions, which can also be shown by not making themselves the subject of the photo, for instance, a runner

posts photos of his shoe collection or a weight lifter post a photo of another weight lifter (usually his rival) who has put all his strengths and efforts to lift a weight that is lighter than his record, even though he admires and encourages that rival with his tribute, but in fact, he put the message of his dominance in the mind of his followers.

## **5.2 *Gender Differences of Self-Presenting and Narcissism (F1)***

The frequency of self-presenting posts can be interpreted as narcissism, a personality trait in which a person has an exaggerated self-image and a desire to be admired [35]. Some studies focus on narcissism in social networks [36], and some confirm the existence of a positive relationship between narcissism and the frequency of selfie posts [37]. Narcissism is found to be related to Instagram use and the time users spend in this social media [37]. Cook et al. [38] found gender issues in the presence of lawmakers on social media, but in a way that was heavily determined by party expectations. However, not much is known about how these gender differences of self-presentation should correctly map into different levels of narcissism in men and women, particularly in sports. The finding of this research, which has been conducted on men and women athletes who are of the same rank (both are Olympic gold medalists), may outline the biased norms of self-presentation on social media concerning gender. Even though now there exists this primitive understanding that women athletes who have more self-presenting posts relative to men athletes are not necessarily more narcissist, but much more research would be required to clarify the variety of forms by which narcissism finds its way into the cybercharacter of athletes, and then quantifying their level of narcissism.

## **5.3 *Engagement and the Ratio of Following to Follower (F2)***

It can be seen in the literature, athletes tend to share their backstage performances. Based on this, one may argue sharing more lifestyle photos boosts the engagement rate [12]. Geurin and Burch [25] found that showcasing diverse photo types aids athletes in promoting their brand. In addition, Rouse and Salter [39] found that cosplayers on Instagram build relationships with potential clients by liking or commenting on their comments and promoting their work.

The findings evince another promising way to boost engagement from the athletes' followers: to increase the ratio of following/follower. In other words, gold medalists who want to get more out of their followers better follow others too. However, the finding showed that this strategy only works for men athletes, which itself opens a new subject to discuss why women gold medalists do not benefit from it. Even though answering this question requires further research, the authors argue that the answer

may have roots in why followers come along to follow the athletes. Indeed, reciprocal relationships that take the form of following and following back on social media befalls on situations in which both sides see their benefit in mutual friendship almost at the same time. Women athletes, particularly at the high rank of gold medalists, aside from their sports fame, benefit from a variety of privileges such as sexuality and attractiveness, particularly for their men followers who engage in their posts without any expectations to be followed back, or at least, behaving the same for women followers of men gold medalists may not be prevalent as well.

### **5.4 Aging and Self-Presentation (F3)**

Roberts et al. [40] concluded that narcissism declines with age because the narcissistic trait of not interacting with others conflicts with normative pathways. Another NPI study of narcissism found a steady decline in narcissism between the ages of 15 and 54 and increased slightly after age 55 [41]. Litchfield and Kavanagh [42] have shown that male athletes tend to be presented in an active posture and are less likely in a passive posture than female athletes. Krane et al. [43] examined the preferred type of sports images young women preferred to see to find authenticity, and self-reflection was features of positively viewed images.

The findings show that aging gold medalists lose interest in sharing their self-presenting posts; however, this decrease for men takes a linear form while women only remain monotonic. The authors argue that the reason resides in the difference of gender-basis tendencies for self-presentation on social media. Aging men gold medalists who used to share their self-presentations as active posts with the spirit of an unbeatable man now feel less strength to be active as in the past. However, women gold medalists who were more satisfied with presenting themselves in passive posts would not run out of interest in self-presentation by aging as much as their male counterparts.

## **6 Conclusion**

This study, by focusing on user characteristics of Olympic gold medalists, has led to new insights about the cyberbehavior of these champions on Instagram. The findings may help the athletes to better promote their media brand and practitioner to devise efficient strategies to boost engagement rate in Instagram.

### *Limitations*

In some rare cases, the authors faced with posts that were not clear to be a pure self-presenting post; for example, in some photos, the gold medalists were in the photo without any friend or fan close to them; however, there were other people in the photo placed in different locations, which seemed there had been no intention to have them

in the photo; thus, those photos were considered as pure self-presenting posts. Also, a few posts were the paintings that depicted the gold medalists' faces as the only subject of the photos, and the authors considered them as pure self-presenting posts.

It could also be argued that the Instagram accounts analyzed in the study were not the authentic gold medalist's account but someone else's. Nearly all the accounts that were surveyed had been tagged by the official Instagram blue registration mark. For the few gold medalists who did not officially register their Instagram accounts, their posts had been explored to confirm the account belonged to them [44]. In addition, some applied graph analysis can be instrumental when we need to characterize users and identify their identities [45], particularly by new techniques for sentiment analysis [46, 47]. Moreover, a real-time implementation of this study can facilitate future research by applying new AI-based techniques both in text [48] and in photo [49] contexts. This would help the researchers run longitudinal studies [50] on the athletes' characteristics where feature extractions could be done with high accuracy without human interference [51], particularly the challenges of selecting photo posts and identifying the self-presentation traits.

**Funding** The study was partially funded by City University of Hong Kong Centre for Communication Research (No. 9360120) and Hong Kong Institute of Data Science (No. 9360163).

## References

1. Bodaghi, A.: A novel pervasive computing method to enhance efficiency of walking activity. *Heal. Technol.* **6**, 269–276 (2016)
2. Frederick, E.L., Lim, C.H., Clavio, G., Walsh, P.: Why we follow: an examination of parasocial interaction and fan motivations for following athlete archetypes on Twitter. *Int. J. Sport Commun.* **5**, 481–502 (2012)
3. Rui, J.R., Stefanone, M.A.: Strategic image management online. *Inf. Commun. Soc.* **16**, 1286–1305 (2013)
4. Karg, A., Lock, D.: Using new media to engage consumers at the Football World Cup. In S. Frawley, D. Adair (Eds.), *Managing the Football World Cup*. Palgrave MacMillan, Melbourne (2014)
5. Lebel, K., Danylchuk, K.: How tweet it is: a gendered analysis of professional tennis players' self-presentation on Twitter. *Int. J. Sport Commun.* **5**, 461–480 (2012)
6. DeAndrea, D.C., Walther, J.B.: Attributions for inconsistencies between online and offline self-presentations. *Commun. Res.* **38**(6), 805–825 (2011)
7. Eagleman, A.N.: Acceptance, motivations, and usage of social media as a marketing communications tool amongst employees of sport national governing bodies. *Sport Manage. Rev.* **16**(4), 488–497 (2013)
8. Parmentier, M., Fischer, E.: How athletes build their brands. *Int. J. Sport Manage. Market.* **11**(1/2), 106–124 (2012)
9. Marshall, P.D.: The promotion and presentation of the self: celebrity as a marker of presentational media. *Celebrity Stud.* **1**(1), 35–48 (2010)
10. Bullingham, L., Vasconcelos, A.C.: 'The presentation of self in the online world': Goffman and the study of online identities. *J. Inf. Sci.* **39**(1), 101–112 (2013)

11. Pegoraro, A., Jinnah, N.: Tweet ‘em and reap ‘em: The impact of professional athletes’ use of Twitter on current and potential sponsorship opportunities. *J. Brand Strategy.* **1**(1), 85–97 (2012)
12. Hambrick, M. E., Kang, S. J.: Pin it: exploring how professional sports organizations use Pinterest as a communications and relationship-marketing tool. *Commun. Sport.*, (2014). <https://doi.org/10.1177/2167479513518044>
13. Casaló, L., Flavián, C., Ibáñez-Sánchez, S.: Antecedents of consumer intention to follow and recommend an Instagram account. *Online Inf. Rev.* **41**(7), 1046–1063 (2017)
14. Lee, M., An, H.:m “A study of antecedents influencing eWOM for online lecture, website: personal interactivity as moderator”. *Online Inf. Rev.*, (2018)
15. Liu, W., Wu, X., Yang, W., Zhu, X., Zhong, S.: Modeling cyber rumor spreading over mobile social networks: a compartment approach. *Appl. Math. Comput.* **343**, 214–229 (2019)
16. Bodaghi, A., Goliae, S.: A novel model for rumor spreading on social networks with considering the influence of dissenting opinions. *Adv. Complex Syst.* **21**(6), 1850011 (2018)
17. Bodaghi, A., Goliae, S., Salehi, M.: The number of followings as an influential factor in rumor spreading. *Appl. Math. Comput.* **357**, 167–184 (2019)
18. Bodaghi, A., Oliveira, J.: The characteristics of rumor spreaders on Twitter: a quantitative analysis on real data. *Comput. Commun.* **160**, 674–687 (2020). <https://doi.org/10.1016/j.comcom.2020.07.017>
19. Souza, F., Casas, D., Flores, V., et al.: “Dawn of the selfie era: the whos, wheres, and hows of selfies on Instagram”. In: Proceedings of ACM on Conference on Online Social Networks, Palo Alto, CA, pp. 221–231. FF (2015)
20. Pittman, M., Reich, B.: Social media and loneliness: why an Instagram picture may be worth more than a thousand Twitter words. *Comput. Human Behav.* **62**, 155–167 (2015)
21. Hochman, N., Schwartz, R.: “Visualizing instagram: tracing cultural visual rhythms.” AAAI Technical Report WS-12-03 Social Media Visualization (2013)
22. Silva, T., Melo, P., Almeida, J., Salles, J., Loureiro, A.: “A picture of Instagram is worth more than a thousand words: workload characterization and application”. In: Proceedings of IEEE International Conference on Distributed Computing in Sensor Systems, pp. 123–132 (2013)
23. Burch, L.M., Clavio, G., Geurin-Eagleman, A.N., Major, L.H., Pedersen, P., Frederick, E.L., et al.: Battle of the sexes: gender analysis of professional athlete tweets. *Global Sport Bus. J.* **2**(2), 1–21 (2014)
24. Geurin-Eagleman, A.N., Clavio, G.: Utilizing social media as a marketing communication tool: an examination of mainstream and niche sport athletes’ Facebook pages. *Int. J. Sport Manage.*, **16**(2) (2015)
25. Geurin, A., Burch, L.: Communicating via photographs: a gendered analysis of Olympic athletes’ visual self-presentation on Instagram. *Sport Manage. Rev.* **19**(2), 133–145 (2016)
26. Zillich, A.F., Riesmeyer, C.: Be yourself: the relative importance of personal and social norms for adolescents’ self-presentation on instagram. *Soc. Med. + Soc.*, **7**(3) (2021). <https://doi.org/10.1177/20563051211033810>
27. Stsiampkouskaya, K., Joinson, A., Piwek, L., Stevens, L.: Imagined audiences, emotions, and feedback expectations in social media photo sharing. *Soc. Med. + Soc.*, **7**(3) (2021). <https://doi.org/10.1177/20563051211035692>
28. Riffe, D., Lacy, S., Fico, F.G.: Analyzing media messages: using quantitative content analysis in research, 2nd edn. Lawrence Erlbaum Associates, Inc. Mahwah, NJ (2005)
29. Smith, L.R., Sanderson, J.: I’m going to instagram It! an analysis of athlete self-presentation on instagram. *J. Broadcast. Electron. Med.* **59**(2), 342–358 (2015)
30. Gazzaniga, M.S., Ivry, R. B., Magnun, G.R., Hustler, J.: Evolutionary perspectives. In: Gazzaniga, M.S., Ivry, R. B., Magnun, G. R. (eds.) *Cognitive Neuroscience: The Biology of the Mind*. W. W. Norton and Company, New York (2009)
31. Berk, L.E.: Self and social understanding. In: Berk, L. E. (eds.), *Child Development*. Pearson Education, Boston (2009)
32. Dovidio, J.F., Brown, C.E., Heltman, K., Ellyson, S.L., Keating, C.F.: Power displays between women and men in discussions of gender-linked tasks: a multichannel study. *J. Pers. Soc. Psychol.* **55**(4), 580–587 (1988)

33. Fink, J.S., Kensicki, L.J.: An imperceptible difference: Visual and textual constructions of femininity in sports illustrated and sports illustrated for women. *Mass Commun. Soc.* **5**(3), 317–339 (2002)
34. Hardin, M., Lynn, S., Walsdorf, K.: Challenge and conformity on contested terrain: Images of women in four women's sport/fitness magazines. *Sex Roles* **53**(1/2), 105–117 (2005)
35. Buffardi, L.E., Campbell, W.K.: Narcissism and social networking web sites. *Pers. Soc. Psychol. Bull.* **34**, 1303–1314 (2008)
36. Kapidzic, S.: Narcissism as a predictor of motivations behind Facebook profile picture selection. *Cyber Psychol. Behav. Soc. Netw.* **16**, 14–19 (2013)
37. Sheldon, P.: In Self-monitoring and narcissism as predictors of sharing Facebook. *J. Soc. Media Soc.* **5**(3), 70–91 (2016)
38. Cook, J.M.: Gender, party, and presentation of family in the social media profiles of 10 state legislatures. *Soc. Med. + Soc.*, (2016). <https://doi.org/10.1177/2056305116646394>
39. Rouse, L., Salter, A.: Cosplay on demand? instagram, onlyfans, and the gendered fantrepreneur. *Soc. Med.+ Soc.*, 7(3) (2021). <https://doi.org/10.1177/20563051211042397>
40. Roberts, B.W., Edmonds, G., Grijalva, E.: It is developmental me, not generation me: developmental changes are more important than generational changes in narcissism—commentary on Trzesniewski and Donnellan (2010). *Perspect. Psychol. Sci.* **5**, 97–102 (2010)
41. Foster, J.D., Misra, T.A., Reidy, D.E.: Narcissists are approach-oriented toward their money and their friends. *J. Res. Pers.* **43**, 764–769 (2009)
42. Litchfield, C., Kavanagh, E.: Twitter, team GB and the Australian olympic team: representations of gender in social media spaces. *Sport Soc.*, (2018)
43. Krane, V., Ross, S., Miller, M., Ganoe, K., Lucas-Carr, C., Barak, K.S.: It's cheesy when they smile: what girl athlete prefer in images of female college athletes. *Res. Q. Exerc. Sport* **82**(4), 755–768 (2011)
44. Tifferet, S., Vilnai-Yavetz, I.: Gender differences in facebook self\_presentation: an international randomized study. *Comput. Human Behav.* **35**, 388–399 (2014)
45. Bodaghi, A., Oliveira, J.: The theater of fake news spreading, who plays which role? a study on real graphs of spreading on Twitter. *Expert Syst. Appl.*, (2021). <https://doi.org/10.1016/j.eswa.2021.116110>
46. Pandian, A.P.: Performance evaluation and comparison using deep learning techniques in sentiment analysis. *J. Soft Comput. Paradigm.* **3**(2), 123–134 (2021)
47. Tripathi, M.: Sentiment analysis of Nepali COVID19 tweets using NB, SVM AND LSTM. *J. Artific. Intell. Capsule Netw.* **3**(3), 151–168 (2021)
48. Manoharan, J.S.: Capsule network algorithm for performance optimization of text classification. *J. Soft Comput. Paradigm* **3**(1), 1–9 (2021)
49. Manoharan, J.S.: Study on hermitian graph wavelets in feature detection. *J. Soft Comput. Paradigm.* **1**(1), 24–32 (2019)
50. Bodaghi, A., Oliveira, J.: A longitudinal analysis on Instagram characteristics of Olympic champions. *Soc. Netw. Anal. Mining.* **12**(1), 3 (2022). <https://doi.org/10.1007/s13278-021-00838-9>
51. Bodaghi, A., Oliveira, J., Zhu, J.J.H.: The fake news graph analyzer: an open-source software for characterizing spreaders in large diffusion graphs. *Softw. Impacts.* **10**, 100182 (2021). <https://doi.org/10.1016/j.simpa.2021.100182>

# Performance Analysis of KNN Algorithm to Improve the Process of Hemodialysis on Post-Covid-19 Patients



N. Vijaya, G. Revathy, D. Sivanandakumar, C. Sasikala, and B. Sreedevi

**Abstract** A recent study shows that covid-19 infected patients are having more probability of developing acute kidney injury that may leads to loss of kidney functionality. Hemodialysis is a process of removing the waste and excess fluids from the blood. Nowadays, because of covid-19, people prefer for home dialysis rather than taking dialysis in the hospitals. Generally, in the patients starting dialysis, almost, 23 percent of patients died in first month due to improper monitoring during the process of dialysis. Here, we have proposed an approach for real-time monitoring and health prediction. Our aim is to predict the probability of success, by analyzing the data using the popular classification techniques of machine learning which gives the maximum rate of accuracy to predict the outcome of dialysis. The system designed will collect the patient's parameters such as temperature, blood pressure, and pulse rate during home dialysis. The stored data are then processed to check for any air bubbles or blood leakage occurrences. In such occurrences, the patient's family is immediately alerted through an SMS to take the patient to hospital. This system helps to reduce the mortality rate after the dialysis treatment. Results prove that the KNN algorithm shows improvement in prediction accuracy of about 14%, 8%, and 5% when compared with logistic regression, SVM, and Naïve Bayes algorithms.

---

N. Vijaya

Department of Computer Science and Engineering, K. Ramakrishnan College of Technology, Tiruchirappalli, Tamilnadu, India

G. Revathy (✉)

School of Computing, SASTRA Deemed to be University, Thanjavur, Tamilnadu, India

D. Sivanandakumar

Instrumentation and Control Engineering, Sri Manakula Vinayagar Engineering College, Puducherry, India

C. Sasikala

Computer Science and Engineering, KPR Institute of Engineering and Technology, Coimbatore, India

B. Sreedevi

Department of Computer Science and Engineering, Srinivasa Ramanujan Centre, SASTRA Deemed to be University, Kumbakonam, India

**Keywords** Covid-19 · Hemodialysis · Machine learning · Acute kidney injury · Health prediction · Real-time monitoring

## 1 Introduction

Covid-19 a major threat to humankind was reported during December 2019. Since then, it has taken out many human lives and serves as a huge challenge to the entire world. Covid-19 categorized as severe acute respiratory distress syndrome that affects all age groups, [1] and the covid infected patients develops severe complications like: acute kidney injury (AKI), thrombotic events leading to stroke, acute myocardial infarction, and pulmonary embolism [2, 3]. AKI is one of the severe disease which damages heart, brain, and lungs. When this AKI reaches advanced stage, dangerous levels of fluid, electrolytes, and wastes can build up in the body causing kidney failure leading to mortality. A treatment for this is kidney dialysis that can be done under three different modes like: Hemodialysis—this process removes waste, toxins, and cleans the blood externally using a dialyzer. Peritoneal dialysis (PD)—this is a surgical procedure where PD catheter is placed in the abdomen to remove waste and toxins from the blood. Continuous renal replacement therapy—this process uses special filters to remove waste and toxins from the blood.

During this pandemic situation, it is highly preferable to treat AKI patients at home rather than in hospital. Hence, home hemodialysis for post-covid patients reduces the in-hospital treatment since hospital beds and intensive care units are occupied by corona infected patients. Generally, lower age groups with less co-morbidity burden will favor the patients for home hemodialysis. In this paper, in order to reduce the mortality and to improve the success rate, the following contributions were made:

1. Mortality in AKI patients under home hemodialysis occurs due to improper monitorization of dialyzed patients. During dialysis process, the patients should be monitored for their blood pressure, temperature, and pulse rate. The tub of dialysis should be monitored for blood leakage and air bubbles in blood stream. The data should be collected and monitored at real time by the respective consultant/doctor. If it reaches certain abnormality, the dialysis process should be immediately stopped, and family members should be alerted with buzzer sound or a call, and immediate treatment should be given. In this approach, we can continuously check the health status of the patient which helps to decrease the mortality rate of the dialysis patients.
2. Here, we also predict the occurrence of diverse reactions among the patients who underwent dialysis using k-nearest neighbor algorithm. We also tested the performance of KNN with other leading machine learning algorithms to find its accuracy in prediction.

This paper is structured as follows: Sect. 2 presents the usage of machine learning techniques in the treatment of AKI. Section 3 describes the proposed approach. Section 4 describes the experimental results and discussion. Finally, Sect. 4 concludes with future plans of the proposed work.

## 2 Machine Learning in AKI-State-of-Art

Generally, post-covid patients treated for hemodialysis in clinic has several drawbacks such as poor sleep, long travel, fatigue, and long period of treatment. Many machine learning approaches are proposed in the treatment of home hemodialysis so far, but they are rudimentary and needs improvement. In [4], authors used two kinds of models to predict the hemoglobin levels in chronic renal failure patients. Here, two kinds of models: local and global models were used. For local models, clustering techniques and adaptive resonance theory and a predictor are used. For global models, artificial neural networks, support vector machines, and regression trees were used. In [5], authors used SVM and ANN models to monitor the hematocrit parameters of blood such as hematocrit and oxygen saturation on spectroscopic-based environment. In [6], authors used KernelSVM algorithm and k-means clustering algorithm to predict the mortality rate of dialysis patients. In [7], the authors proposed a system for identifying and predicting cancer cells using singular value decomposition method coupled with Fast Dictionary Learning algorithm using steepest descent method. In [8], the authors proposed a system that predicts the occurrence of any blood leakage in the dialysis process using fog computing. The authors developed a tool to alert the patient when there is blood leakage or blood loss during the dialysis process using an array of photocell sensors and hetero-associative memory model. In [9], the authors proposed a self-organizing algorithm to detect the blood leakage levels using virtual alarm unit, and the observations are communicated via wireless fidelity over cloud. In [10], the authors identified chronic kidney disease using various machine learning models and classification techniques. Here, the model uses various classifier to accurately predict the chronic kidney disease. In [11], the authors used extreme gradient boosting machine learning approach to validate AKI fluid responsiveness model. They have also compared the performance of their approach with logistic regression model. In [12], the authors proposed a deep learning approach for continuous prediction of acute kidney injury before 48 h.

Even though many approaches are proposed for monitoring hemodialysis patients, they are rudimentary and needs improvement. This novel approach of analyzing post-covid patients during dialysis process shows improvement in prediction accuracy of about 14%, 8%, and 5% when compared with logistic regression, SVM, and Naïve Bayes algorithms.

## 3 Prediction of Risk Using KNN

The hemodialyzer acts as an extracorporeal circuit by replacing the kidney's regular function. During this process, the patient blood from the body is redirected to this external machine where the dialyzer processes the blood using dialysis fluid and a semi-permeable dialysis membrane. To promote the healthy function of kidneys, many critical parameters should be monitored and controlled during the dialysis

process. Here, we have intended a system consists of Arduino microcontroller connected with various biometric sensors to monitor the patient's body temperature, blood pressure, and pulse rate continuously. Blood leakage is a thoughtful life-threatening origin in this progression. Hence, we have air bubble sensor and blood leakage sensors. Air bubble sensors are used to detect air bubbles and prevent air intercalation. If endure over looked may block the slender blood vessels and causes death. Blood leakage sensors are to perceive blood circuit disconnection and membrane defects. The critical structures are collected from various sensors and displayed in the LCD display and recorded at a regular interval of time. The collected data are then processed and checked for any abnormal condition. The sensors continuously monitor the patient's parameter values comparing with the normal predefined values (set by therapeutic session). If the chronicled ethics seem to exceed predefined values, the system immediately alerts the patient's family members and doctor through a buzzer for patient's immediate treatment. The data are stored for future reference of the doctor for any prescription that may be recommended.

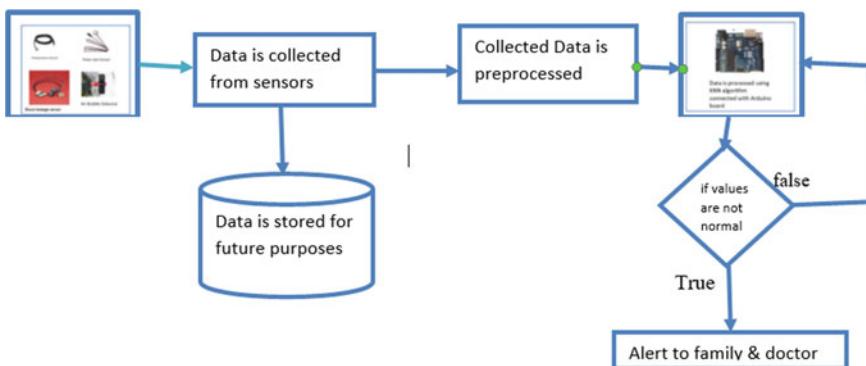
The hemodialysis over the purification of the blood components and the patient risk factors is analyzed. The detection of the casual components identification has made with the physical factors like temperature, pressure, blood leakage, and air bubbles. If any abnormality occurs, the patient's family and the doctor are immediately alerted with a buzzer sound or SMS to deliver further treatment. Figure 1 shows the work flow of proposed system that consists of following stages:

**Data collection:** During this stage, data from various sensors are collected, and the output is interfaced with the analog to digital circuit pins of microcontroller.

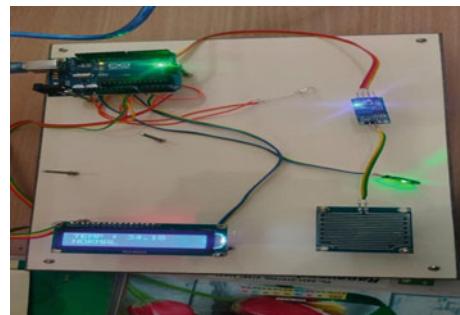
**Data storage:** During this stage, data will be stored for statistical analysis and future prediction.

**Data processing:** Here, the collected data are processed using KNN algorithm for detecting and predicting adverse effects if any.

**Alert system:** During this stage, if the collected data from sensors seem to exceed predefined values, the system will immediately generate SMS to alert the patient's family members and doctor.



**Fig. 1** Workflow

**Fig. 2** Circuit connection

### **3.1 Circuit Connection of the System**

Figure 2 shows the circuit connection of the system that includes the temperature sensor, pulse rate sensor, air bubble detector, blood leakage sensor, Arduino microcontroller, buzzer, and LCD display. All are connected with a low power supply of battery 5 volts.

### **3.2 Implementation**

Tables 1 and 2 show the values of each sensors. The output from various sensors such as temperature, pulse rate, air bubble detector, and leakage detector are given to numerous machine learning algorithms such as constant, SVM, Naïve Bayes, logistic regression, and KNN. Each and every workflow is discussed below.

**Table 1** Sensors with predefined and abnormal values

Sensors used	Predefined values	abnormal value
Temperature sensor	96° F to 99° F	Above 99° F
Pulse rate sensor	70 to 90 bpm	Above 90 bpm

**Table 2** Sensors description

Sensors used	Power supply	Description
Blood leakage sensor	5 V	It detects any blood leakage in the patient's dialysis process
Air bubble sensor	3.5 TO 5 V	It detects air bubble in dialysis process

### 3.2.1 Implementation of KNN and Result Discussion

KNN predicts according to nearest training instances. The KNN algorithm explores for K-closest training samples in feature planetary and customs it middling for prediction [13].

Pseudo-Code for KNN.

1. Compute the Euclidean distance between the points where it starts from one to n number of values.
2. Arrange the values in ascending order based on distance.
3. Let us assume a value ‘x’ which is the first distance value among the number of values.
4. Uniquely split the values based on x points in Euclidean distance
5. Let  $q_i$  represents the quantity of points belonging to a specific class among  $x$  points, i.e., minimum x value must be greater than zero.
6. Compare each and every point if the value of distance is less than  $x$ , it will be combined in the specific class.

Preprocessing in KNN.

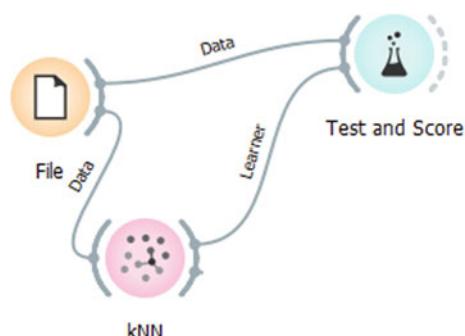
KNN executes in the following order.

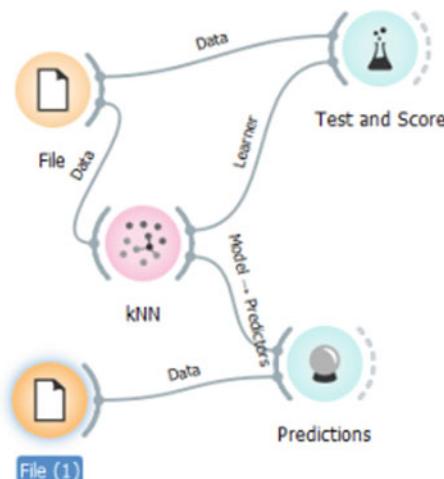
1. Removes instances with unknown target values
2. Continues categorical variable
3. Removes empty column
4. Imputes missing values with mean values
5. Normalizes the data by centering to mean and scaling to standard deviation of 1.

The input dataset is passed to the constant, and temperature, pulse rate, air bubble detection, and blood leakage detection are considered as feature variable, and output is assumed to be target variable. Constant gives an accuracy of 100 percent for the dataset.

Figure 3 shows the hemodialysis data trained with KNN model in Orange tool. The model is applied to the dataset, and the values are tested and scored. Figure 4

**Fig. 3** Hemodialysis data are trained with KNN model and tested





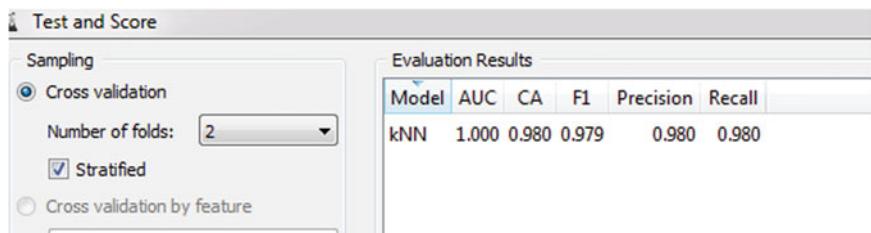
**Fig. 4** Hemodialysis testing data output prediction using KNN

shows the hemodialysis data model trained using KNN, and testing data are applied with KNN, and the prediction values are verified.

Figure 5 KNN shows the prediction with hundred percent accurate results. Figure 6 shows the complete testing prediction of KNN. Figure 7 shows the accuracy of KNN. The same KNN is repeated with different folds. The technique has a single constraint called  $k$  that indicates to the quantity of groups that a preset data sample is to be scrapped into. As such, the practice is often called  $k$ -fold cross-validation. When an obvious charge for  $k$  is preferred, it may be cast off in place of  $k$  in the ailment to the model, such as  $k = 10$  charming tenfold cross-validation. Folds are an inclusive spread progression as it is unassuming to realize and as it commonly consequences in a less partial or less expectant approximation of the model skill than other methods, such as a simple train/test split [14].

The general procedure is as follows:

1. Disclose the data in different orders.
2. Couple the data in  $x$  groups.



**Fig. 5** KNN model evaluation result

	kNN	Output	Temperature in F	Pulse Rate in BPM	ir Bubble detection	bd Leakage Detected
1	0.00 : 1.00 → NORMAL	NORMAL	94.0	72	NO	NO
2	0.00 : 1.00 → NORMAL	NORMAL	90.0	72	NO	NO
3	0.00 : 1.00 → NORMAL	NORMAL	98.4	73	NO	NO
4	0.00 : 1.00 → NORMAL	NORMAL	96.0	82	NO	NO
5	1.00 : 0.00 → ABNORMAL	ABNORMAL	101.0	98	YES	YES
6	0.00 : 1.00 → NORMAL	NORMAL	97.0	96	NO	NO
7	0.00 : 1.00 → NORMAL	NORMAL	96.0	97	NO	NO
8	1.00 : 0.00 → ABNORMAL	ABNORMAL	104.0	99	NO	NO
9	0.00 : 1.00 → NORMAL	NORMAL	91.0	71	NO	NO
10	0.00 : 1.00 → NORMAL	NORMAL	90.0	72	NO	NO

**Fig. 6** Testing data predicton using KNN**Fig. 7** Testing data prediction result with KNN

Model	AUC	CA	F1	Precision	Recall
kNN	1.000	1.000	1.000	1.000	1.000

3. Assume for each specific group
  - (a) Combine the group as hold value or test data
  - (b) Consider the group with high values as training data
  - (c) Specific a value or a group of values in training data and using it apply it on testing data.
  - (d) Every time remember the optimal value and follow the same procedure.
4. Combine the skill models and evaluate using model value.

### 3.2.2 Implementation of Constant and Comparison with KNN

Constant predicts the maximum frequent class or mean value from the training set. This learning model predicts the majority for classification. For classification, when foreseeing the class value with calculations, constant will return relative frequency of the classes in the training set.

Input: Dataset which we got output from IoT sensors.

Constant: Preprocessing methods.

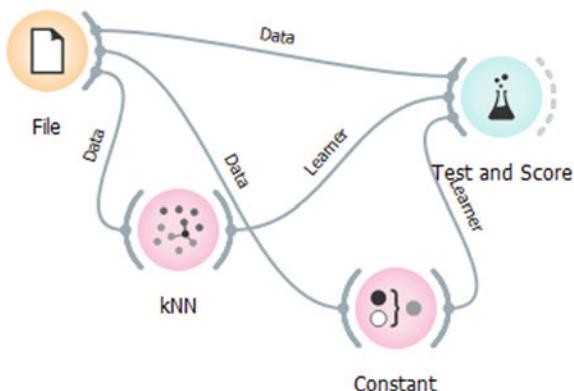
Constant Learner: Majority/mean model.

Model: Trained.

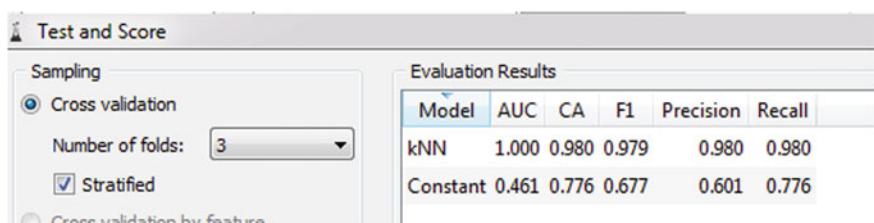
The input dataset is passed to the constant, and temperature, pulse rate, air bubble detection and blood leakage detection are considered as feature variable, and Output is assumed to be target variable. Constant gives an accuracy of 46 percent for the dataset.

Constant model is compared with KNN, and the results are recorded.

Figure 8 shows the model trained by KNN and constant, and testing scores are recorded. Figure 9 shows comparison of KNN and constant. Figure 10 shows the output for each fold with KNN and constant.

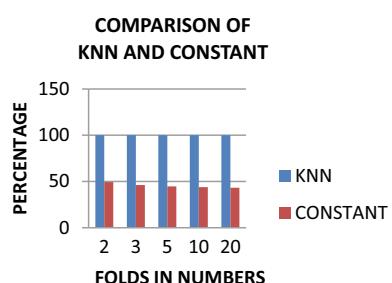


**Fig. 8** KNN comparison with constant



**Fig. 9** KNN output comparison with constant

**Fig. 10** KNN comparison with constant



### 3.2.3 Implementation of SVM and Comparison of KNN with SVM

Support vector machines (SVMs) map involvements to sophisticated dimensional feature spaces. SVM is a ML technique that ruptures the characteristic space with a hyperplane, thus exploiting the margin between the occurrences of different classes. For a dataset entailing of features set and labels set, an SVM classifier builds a model to predict classes for new examples. It dispenses new specimen/facts points to one of the classes. If there are solitary two classes, then it can be called as a binary SVM classifier.

There are two kinds of SVM classifiers:

- Linear SVM classifier
- Non-linear SVM classifier

Preprocessing:

SVM executes in the following order.

- Clear the facts that does not have a feature value.
- Computes the model with the target variable.
- Clear all the unused or unvalued columns.
- Specify the mean values in the place of nulled values.

Input: Dataset which we got output from IOT sensors.

SVM: Preprocessing methods.

SVM Learner: Linear regression models.

Model: Trained.

The input dataset is passed to the constant, and temperature, pulse rate, air bubble detection, and blood leakage detection are considered as feature variable, and output is assumed to be target variable.[15] SVM gives an accuracy of 98 percent for the dataset. SVM is compared with KNN, and the values are recorded.

Figure 11 shows the model trained by KNN and SVM; Fig. 12 shows the output comparison of KNN and SVM. Figure 13 shows the comparison of KNN and SVM with each folds.

### 3.2.4 Implementation of Naïve Bayes and Comparison with KNN

A fast and simplest probabilistic classifier based on Naïve theorem with the assumption of feature independence.

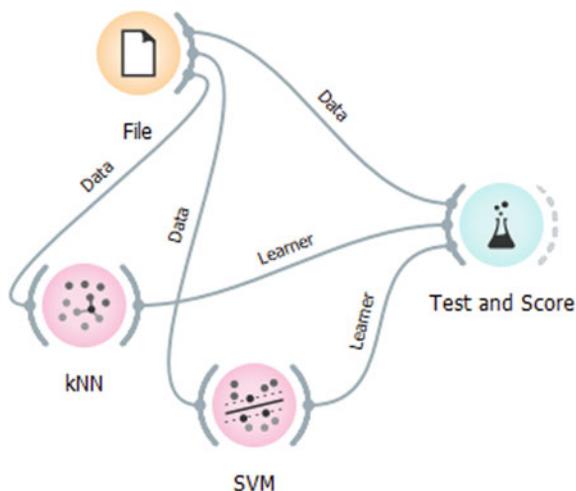
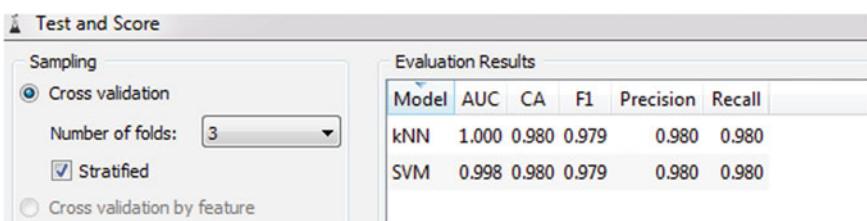
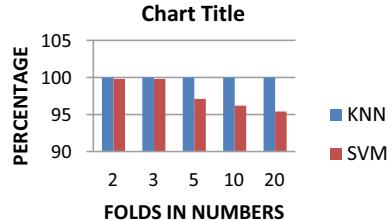
Input: Dataset which we got output from IOT sensors.

Naïve Bayes: Preprocessing methods.

Naïve Bayes Learner: Naïve Bayes learning models.

Model: Trained.

It works on conditional probability. Conditional probability is the probability that something will happen, given that something else has already occurred. Using the

**Fig. 11** KNN and SVM compared**Fig. 12** KNN output comparison with SVM**Fig. 13** KNN comparison with SVM

conditional probability, we can calculate the probability of an event using its prior knowledge.

Below is the formula for calculating the conditional probability.

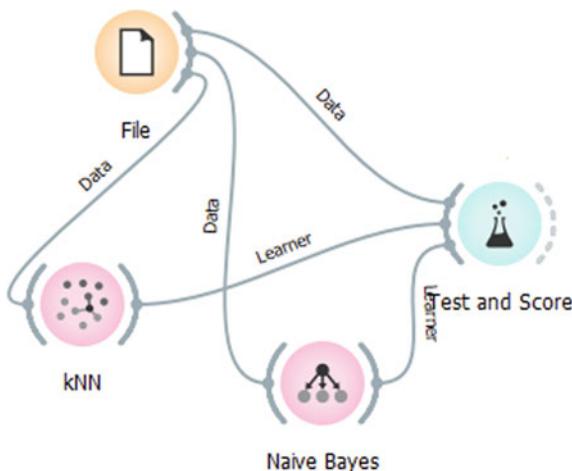
$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

where

- $P(H)$  is the probability of hypothesis H being true. This is known as the prior probability.
- $P(E)$  is the probability of the evidence (regardless of the hypothesis).
- $P(E|H)$  is the probability of the evidence given that hypothesis is true.
- $P(H|E)$  is the probability of the hypothesis given that the evidence is there.

The input dataset is passed to the constant, and temperature, pulse rate, air bubble detection, and blood leakage detection are considered as feature variable, and output is assumed to be target variable. Naïve Bayes gives an accuracy of 99 percent for the dataset. Naïve Bayes is compared with KNN, and the values are recorded.

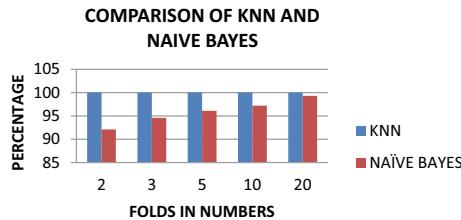
Figure 14 shows the model trained by KNN and Naïve Bayes; Fig. 15 shows the output comparison of KNN and Naïve Bayes. Figure 16 shows the comparison of KNN and Naïve Bayes with each folds.



**Fig. 14** Naïve Bayes comparison with KNN

Test and Score						
Sampling		Evaluation Results				
Model	AUC	CA	F1	Precision	Recall	
kNN	1.000	0.985	0.985	0.986	0.985	
Naive Bayes	0.993	0.978	0.978	0.979	0.978	

**Fig. 15** KNN comparison with Naïve Bayes



**Fig. 16** Comparison graph of KNN and Naïve Bayes

### 3.2.5 Implementation of Logistic Regression and Comparison with KNN

The logistic regression classifier algorithm with LASSO LI or Ridge L2 regularization is studied. Lasso regression is like linear regression, but it uses a technique ‘shrinkage’ where the coefficients of determination are shrunk toward zero.

Linear regression gives you regression coefficients as observed in the dataset. The Lasso regression allows you to shrink or regularize these coefficients to avoid overfitting and make them work better on different datasets.

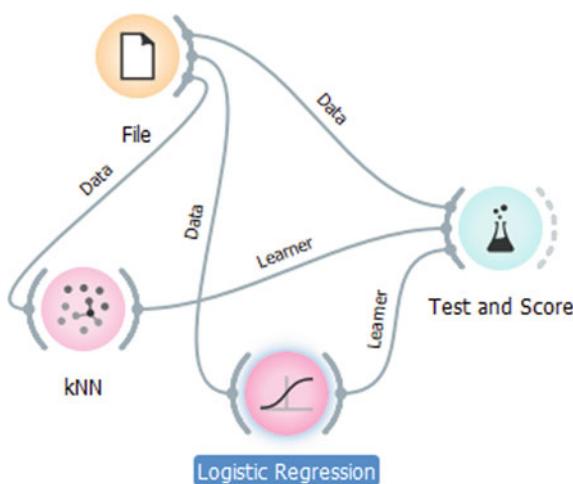
Input: Dataset which we got output from IOT sensors.

Linear Regression: Preprocessing methods.

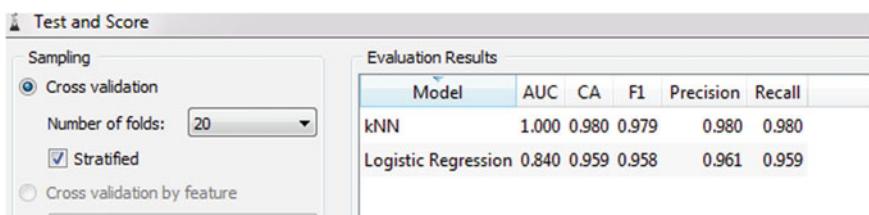
Linear Regression Learner: Linear regression learning models.

Model: Trained.

Figure 17 shows the model trained by KNN and logistic regression; Fig. 18 shows the output comparison of KNN and logistic regression. Figure 19 shows the comparison of KNN and logistic regression with each fold.

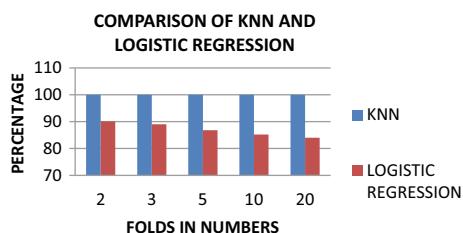


**Fig. 17** Comparison of KNN with logistic regression



**Fig. 18** KNN comparison with logistic regression

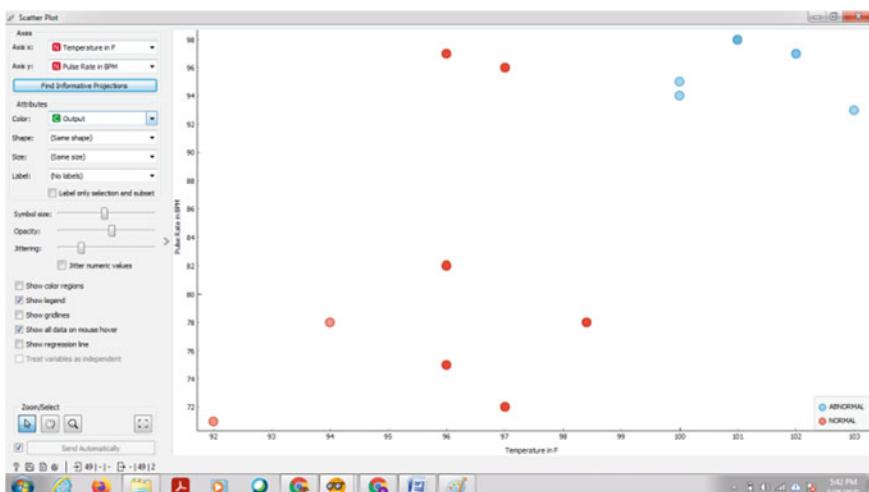
**Fig. 19** Comparison of KNN and logistic regression



### 3.2.6 Overall Comparison

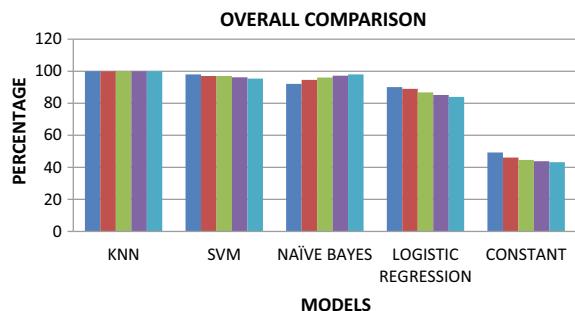
Based on the above results, KNN shows higher accuracy with each folds of data.

Figure 20 scatter plot has two values; the first one is normal values, and second one is abnormal values. Figure 21 describes the comparison of KNN with all other algorithms such as constant, Naïve Bayes, and logistic regression. At each fold, KNN shows higher accuracy results when compared with all other algorithms



**Fig. 20** Scatter plot output for KNN

**Fig. 21** Overall comparison of KNN with SVM, Naïve Bayes, logistic regression, and constant



## 4 Conclusion

During this pandemic period, treating AKI patients, who recovered from covid-19, in the hospital is risky due shortage of ICUs and beds. Hence, home hemodialysis is a better option for patients and doctors. In our approach, we have tested several machine learning algorithms for predicting the abnormal/normal conditions of post-covid patients undergoing home hemodialysis. Various critical parameters are studied, and the results prove that KNN classifier outperforms other machine learning classifiers in predicting the abnormal conditions of patients during dialysis. When compared with all the machine learning algorithms, KNN shows 100 percent accuracy, whereas constant shows an average of 46%, logistic regression 86%, Naïve Bayes 92%, and SVM 95%.

## References

1. Wu, Z., McGoogan, J.M.: Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. *JAMA* **323**(13), 1239–1242 (2020)
2. Cheng, Y., et al.: Kidney disease is associated with in-hospital death of patients with COVID-19. *Kidney Int.* **97**(5), 829–838 (2020)
3. Klok, F.A., et al.: Confirmation of the high cumulative incidence of thrombotic complications in critically ill ICU patients with COVID-19: an updated analysis. *Thromb. Res.* **191**, 148–150 (2020)
4. Martínez-Martínez, J.M., et al.: Prediction of the hemoglobin level in hemodialysis patients using machine learning techniques. *Comput. Methods Programs Biomed.* **117**(2), 208–217 (2014)
5. Decaro, C., et al.: Machine learning approach for prediction of hematocrit parameters in hemodialysis patients. *IEEE J. Trans. Eng. Health Med.* **7**, 1–8 (2019)
6. Radović, N., et al.: “Machine learning approach in mortality rate prediction for hemodialysis patients.” *Comput. Methods Biomech. Biomed. Eng.*, 1–12 (2021)
7. Vigneswari, T., Vijaya, N., Kalaiselvi, N.: Early prediction of cervical cancer using machine learning techniques. *Turkish J. Physiotherapy Rehabil.* **32**(3), 262–269 (2021)
8. Chuang, H.C., et al.: The development of a blood leakage monitoring system for the applications in hemodialysis therapy. *IEEE Sens. J.* **15**(3), 1515–1522 (2014)

9. Lin, C.H., et al.: Assistive technology using integrated flexible sensor and virtual alarm unit for blood leakage detection during dialysis therapy. *Healthcare Technol. Lett.* **3**(4), 290–296 (2016)
10. Jena, L., et al.: “Risk prediction of kidney disease using machine learning strategies.” *Intell. Cloud Comput.*, Springer, Singapore, pp. 485–494 (2021)
11. Zhang, Z., Ho, K.M., Hong, Y.: Machine learning for the prediction of volume responsiveness in patients with oliguric acute kidney injury in critical care. *Crit. Care* **23**(1), 1–10 (2019)
12. Revathy, G et al.: Machine learning algorithms for prediction of diseases. *Int. J. Mech. Eng.* **7**(1), 2672–2676 (2022)
13. Thakur, N., Han, C.Y.: A study of fall detection in assisted living: identifying and improving the optimal machine learning method. *J. Sens. Actuator Netw.* **10**(3), 39 (2021)
14. <https://dias.library.tuc.gr>
15. <https://orange3.readthedocs.io>

# Secure Mobile Internet Banking System Using QR Code and Biometric Authentication



S. Ajish and K. S. Anil Kumar

**Abstract** With the digital technology explosion, Internet banking users increased exponentially due to worldwide accessibility and convenience. The leading challenge of online banking is to ensure security for online transactions and the accounts of customers. Phishing attempts to get user's login credentials like username and password by disguising themselves as a trusted entity in the banking sector. SIM swap is a new cyber fraud, where the attacker collects the personal data of the bank customer and gets a new SIM card. The attacker can easily steal the user's login credentials like user name and password using phishing attack and the OTP using the SIM swap fraud attack. This study analyzes the security of the online banking system and proposes a new anomaly-based fraud detection method to overcome phishing and SIM swap fraud attacks. The login attributes like IP address, device, cookie, operating system, and browser are used to generate and update the user's profile. The primary user profile contains the most recently used login attributes, and the second profile contains the most frequently used login attributes. If the current login attributes match either the primary or secondary user profile, the user can access their account. Otherwise, additional security mechanisms like OTP with QR code or biometric authentication or both are used to identify the suspicious behavior of the user. The proposed method reduces the login burden of the user and provides better security for the online mobile banking system.

**Keywords** Mobile banking · QR code · Biometric authentication · Phishing · SIM swap · Anomaly detection

---

S. Ajish (✉)

Department of Futures Studies, University of Kerala, Thiruvananthapuram, Kerala, India

e-mail: [ajishs2014@gmail.com](mailto:ajishs2014@gmail.com)

URL: <http://www.springer.com/gp/computer-science/lncs>

K. S. Anil Kumar

University of Kerala, Thiruvananthapuram, Kerala, India

## 1 Introduction

The digitalization process changes the lifestyle of human beings in a technologically advanced manner. Internet banking is one of the main contributions of digitalization technology, and the banking sector is more dependent on technology. Nowadays, Internet banking or e-banking is common among banking customers, and it provides flexibility and convenience. E-banking [1] allows bank customers to perform financial transactions using the Internet. The bank's services come in a wide range, including managing the account, checking accounts, bill payments, etc. Additionally, e-banking helps customers to access their bank accounts through bank Web sites from anywhere using the Internet.

The emergence of technologies in the banking sector leads to cyber-attacks. Some common cybersecurity attacks faced by the banking area are phishing, pharming, man-in-the-middle, man-in-the-browser, attacks utilizing Trojan horses, cross-channel attacks, botnets, etc. [2]. Among them, the most popular and dangerous attack is phishing, and phishing means phreaking plus fishing. Studies reveal that around 85% of attacks mainly concentrated on financial institutions.

Phishing [3] is an attempt to get useful information like the username and password of a banking user by disguising it as a trusted entity in communication. The phishing attack tries to steal user's login credentials and credit card numbers using deceptive e-mails and Web sites. When the user enters their sensitive information like username, password, etc., into the deceptive Web sites, the attackers acquire this information. By using this stolen information, the attacker can perform fraudulent banking transactions.

Online banking systems require efficient security models to identify authorized users and authorize their transactions to fight against cyber frauds. According to the Anti-phishing Working Group [4], the rate of phishing attacks increases day by day. The banking system uses two-factor authentication with user name and password and OTP to avoid a phishing attack. SIM swap [5] is a new cyber fraud, where the attacker collects the personal information of the bank customer and obtains a new SIM card. Once the new SIM is activated, the scam can access the mobile number and initiate fraud activities on the bank account. Once the attacker has stolen the user credentials like username and password, the current system is compromised.

This paper proposes a secure anomaly-based mobile Internet banking system that correctly detects unauthorized login attempts in banking applications. The main contribution of this system is to provide efficient security to the user account by checking various attributes at login time. The login attributes like IP address, device, cookie, operating system, and browser are used to generate and update the user's primary profile. The anomaly-based mobile Internet banking system reduces the login burden of genuine users.

In mobile Internet banking, the customer may log in to the banking system using an IP address at the home, office, etc. The IP address attribute of the customer changes when the user logs in using a mobile phone at different locations. The customer has to go through biometric authentication when he switches from office to home and vice versa, and it will degrade the performance of anomaly-based fraud detection. In this

paper, the anomaly-based detection is modified by including a secondary user profile that contains the customer's frequently used login attributes. The primary profile contains the most recently used login attributes, and the second profile contains the most frequently used login attributes.

If the current login attributes match either the primary or secondary user profile, the users are allowed to access their account. Otherwise, additional security mechanisms like OTP with QR code [6] or biometric authentication [7] or both are used to identify the suspicious behavior of the user. If the current login attribute matches the secondary profile, the primary and secondary profiles do not update.

The remaining portion of the paper is organized as follows: Sect. 2 goes through the literature review, and Sect. 3 describes the methodology. Section 4 describes the implementation details; Sect. 5 analyzes the performance of the anomaly detection algorithm. Section 6 analyzes the security, Sect. 7 discusses the results, and Sect. 8 concludes the paper.

## 2 Literature Review

The present research is directed toward cyber-attacks [8] and the identification of those attacks in e-banking systems. The current system provides various security methods to avoid phishing attacks. These security methods generally include device registration, one-time password, short message system, virtual keyboard, captcha, etc. But these security models are vulnerable to some attacks. Now financial institutions use various technologies to combat phishing attacks and are commonly referred to as phishing attack countermeasures, which usually includes web page and e-mail personalization, two-factor authentication [9], protection of software and user awareness programs, etc.

### 2.1 Phishing

Phishing makes a fraudulent effort to achieve sensitive information, such as user names, passwords, credit or debit card details, or other confidential information, by imitating as a reliable entity. The different types of phishing are

1. E-mail phishing [10]: E-mail phishing is the most common phishing type; the e-mail contains a message to the recipient that your account is compromised, and you have to respond by clicking the following link. When the user clicks the link, the attacker asks for sensitive information like username, password, credit card number, etc. Some phishing e-mails are difficult to recognize because phishing e-mails are designed carefully using the language and grammar of the actual e-mail. The e-mail domain of the phishing e-mail is almost similar to the existing e-mail domain, so it is difficult for the victim to identify the phishing e-mail.

2. Smishing [10]: The smishing attack uses the short message service (SMS) or text messaging to steal sensitive information. The message contains a phone number to call or a link to click that results in a smishing attack. The SMS is texted carefully; it is like the SMS from the bank and contains the information that your account is compromised and responds immediately. The attacker tells the victim to verify his account number, SSN, etc.; then, he gets control of the victim's bank account.
3. Vishing [10]: The vishing attack is similar to e-mail phishing and smishing, such that the attacker tries to theft personal or sensitive information. The variation in vishing attacks is that vishing is carried out through phone calls or voice calls.

## 2.2 *SIM Swap Fraud*

Nowadays, most online banking implements a second-factor authentication for login, such as OTP, for added security. SIM swap [11] is a new cyber fraud, where the attacker collects the personal information of the bank customer and obtains a new SIM card. The scam collects the customer's personal information like name, address, date of birth, phone number, Aadhaar number, etc., from social media or by using phishing links, social engineering via SMS, phone calls, etc.

After collecting the personal information, the attacker contacts the victim's mobile service provider and asks for a new SIM card, claiming that the SIM card is damaged or lost [12]. The attacker uses the gathered information to manage and obtain a new SIM card of the registered mobile number of the victim (bank customer). Once the new SIM is activated, the fraudster can access the mobile number and initiate fraud activities on the bank account.

The steps in the SIM swap fraud attack [11] are

- Step 1: The attacker collects the personal information of the bank customer through social media, phishing, or any other means.
- Step 2: The fraudster approaches the mobile service provider along with the ID proof of the customer and requests a new SIM.
- Step 3: The mobile operator deactivates the old SIM and issues a new SIM to the fraudster.
- Step 4: The fraudster then gets the OTP required for online banking in the new SIM set on his mobile phone.

## 2.3 *Methods to Safeguard Online Banking*

The main action by the banking institution to safeguard online baking is to give awareness to the customers about the online banking threats. The bank provides some warning to the online customers that do not compromise their bank account

to attackers. The foremost advice by the banking institutions to the customers is do not disclose confidential information to anyone, do not respond to phishing and scamming e-mails, set up secure passwords, etc.

Almost all banks implement the two-factor authentication method to enhance security where the user has to enter the one-time password (OTP) [13] other than user name and password. Some banks implement two-factor authentication by requesting the code on the backside of the debit card instead of OTP.

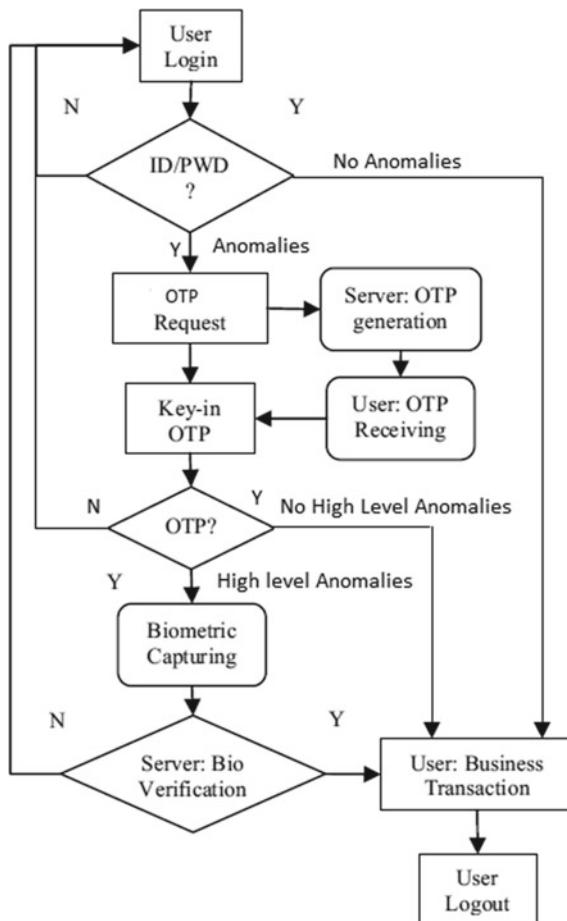
The device registering method [14] restricts the access to the online banking system to the registered and previously known devices. The biometric authentication methods are used to identify the user before the registration of the device. Device identification is commonly used with device registration, but it is used stand-alone in online banking. In the device identification method, the previous transaction details are logged and used to identify the device.

Transaction monitoring is currently used in all online banking systems in which different techniques are used to monitor the transaction. Each customer's transaction history is saved and analyzed using artificial intelligence to identify fraud transaction patterns.

### 3 Methodology

Nowadays, most online banking uses two-factor authentication such as passwords and mobile OTP [13]. The attacker tricks bank customers and collects the user name and password using phishing attacks. Using the SIM swap fraud [11], the attacker can quickly obtain the OTP required for the online banking authentication. The authentication of bank customers using passwords and mobile OTP is not secure, and second-factor authentication like OTP hassles the customer.

This study designs a reliable, efficient, and trustworthy solution to identify genuine login attempts and block login attempts from fraudsters. The proposed method uses an anomaly-based multi-factor method, as shown in Fig. 1. The parameters IP address, device, browser, operating system, and location are considered to identify the anomaly. The anomaly-based multi-factor authentication method allows the user to log in to the account with the user name and password if no abnormalities are detected. If no anomalies are detected, the user can log in with a user name and password that keep away the customer from the burden of the second authentication step, such as OTP with QR code, and the third authentication step, such as biometric authentication. The second-level authentication (OTP with QR code) is initiated when the operating system change anomaly is detected. When high-level anomalies like IP address change, device change, and location change are identified, second-level authentication (OTP with QR code) and third-level authentication (biometric authentication) are initiated for higher security. The different types of anomalies and corresponding security actions are reported in Table 1.



**Fig. 1** Login using OTP and biometric authentication

**Table 1** Security actions for different anomalies

IP address	Device	Operating system	Browser	Action
✓	✓	✓	✓	No action
✓	✓	✓	✗	No action
✓	✓	✗	✓	OTP authentication
✓	✓	✗	✗	OTP authentication
✓	✗	✓	✓	OTP and Biometric authentication
✗	✓	✓	✓	OTP and biometric authentication
✗	✗	✓	✓	OTP and biometric authentication

The different steps in the anomaly-based user identification are:

Step 1: In the primary stage, when the user attempts a login, the IP address, location, cookie, device, operating system, time of day, and browser information are extracted and are used to detect abnormalities in the user behavior. Based on this information, the customer's profile is created, and it is stored in the bank database to identify any anomalies in the login attempt.

Step 2: From the next login attempt onward, the various attributes like IP address, device, location cookie, operating system, browser, etc., are compared with the information present in the database. If no anomalies are detected, and all the factors are compatible with the database information, banking transactions are allowed.

Step 3: After a specific number of login attempts, one frequently used login parameter like IP address, device, location cookie, operating system, browser, etc., of each user is identified, and it is saved as a secondary personal profile in the database.

Step 4: After the secondary user profile is generated, the various login attributes like IP address, device, location cookie, operating system, browser, etc., are compared with the primary and secondary profile of the user to identify anomalies.

Step 5: Other higher security mechanisms like OTP with QR codes and biometric authentications are added based on the anomalies. Based on the user's successful completion of the various security methods, access to the bank details is allowed or denied.

Step 6: After the customer completes all the security steps, the primary profile of the customer is updated, and it is further used for fraud identification. The secondary profile is updated using the most frequently used algorithm [15].

## 4 Implementation

In this study, anomaly-based detection, device identification, IP address identification, OTP with QR code, and fingerprint biometric authentication are significant concepts.

### 4.1 Modified Anomaly Detection Algorithm

Anomaly-based detection monitors the normal behavior of the customer for a specific time, and it is logged. During the registration process, the customer gives some details, and using it, a temporary profile of the customer is created. The temporary profile is updated by analyzing the behavior of the customer for a specific time.

The IP address, device, location, cookie, operating system, browser, login time, etc., of each user are identified and used to generate the profile. The profile of each customer is stored in a local database if any abnormal action that is different from

the user profile detail is identified, which is classified as an unauthorized attempt. Suppose a user attempts a login from a country that is different from the location saved in his profile; that might be an unauthorized attempt.

The customer may use the phone at different locations (home, office, etc.) in mobile banking. The customer's profile is updated each time they log in from home and office. The customer must undergo OTP and biometric added security mechanism each time they login from home and office. A secondary profile for each user is generated, which contains the frequent login parameters like IP address, device location, cookie, operating system, browser, login time, etc., to overcome the difficulty.

After completing the profile generation of a user, it is stored in the database for authorized user identification. At the time of login, the login details such as the IP address, location, cookie, device, operating system, time of day, and browser information are compared with the details of the specific user profile (primary and secondary) stored in the database.

If no anomalies are detected at the validation stage, the user is recognized as an authorized user. If any low-level anomaly like browser change or operating system change is identified, the user has to undergo OTP with a QR code security mechanism. If any high-level anomaly such as IP address change, device change, and location change is identified, the user must undergo OTP and biometric security mechanism.

## **4.2 Device Identification**

When the banking customer registers to avail the online banking facility, the device used is identified using a cookie [16]. The cookie is a tiny piece of information saved by the browser, and it is used to identify the device in the future. When the user attempts to log in through a different device, the device is not recognized, and higher security mechanisms like OTP and biometric authentication are initiated.

## **4.3 IP Address Identification**

The IP address [17] is a unique number assigned to each device in the network to identify the devices. Internet uses the IP address to send and receive IP packets to destination and source. The IP packet received by the server contains the source and destination IP address, so the server can easily identify the client (source) machine. When the user tries to log in, the IP address is identified and validated, and the user login action is allowed if it is valid otherwise denied. If users access their account using any malicious IP address, the login attempt is blocked by the server. The blocklisted IP addresses are saved in the database to compare them with the request's IP address. Once users access their account, the IP address is identified as

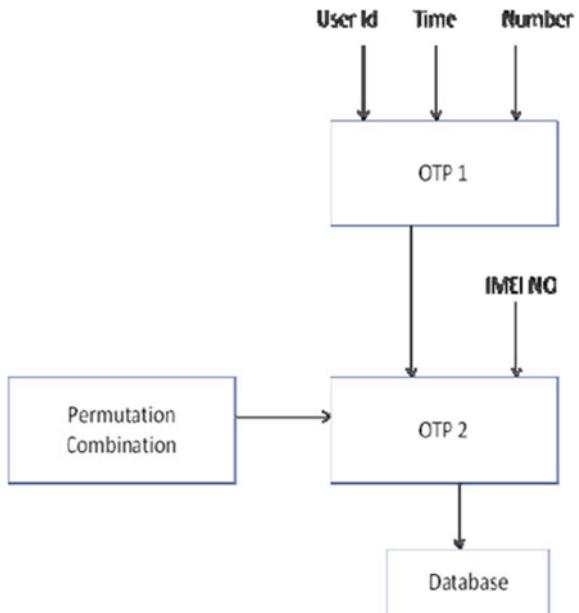
valid, and that information is saved in the database. When users access their account using an IP address different from the stored data, the OTP and biometric-based authentication method authenticate the user.

#### 4.4 Mobile OTP with QR Code

The working of OTP with the QR code authentication method is indicated in Fig. 2 [6]. After the registration and login process, if the anomaly-based detection identifies that the IP address and the device are the same and the operating system is different, the additional security action “OTP with QR code” is initiated. As indicated in Fig. 2 [6], using the user id, system time, and random number, the OTP1 is generated and concealed in the QR code. Then, the QR code image is generated using the parameters OTP1, image format, and image size, and the generated QR code is displayed on the screen.

The displayed QR code is scanned using a QR code scanner application specially designed for this banking application. The scanner application reads OTP1 from the QR code and generates OTP2 using permutation combination logic with the parameters OTP1 and IMEI number [18] of the customer’s mobile phone. The OTP2 is displayed on the screen, and the customer enters OTP2 for authentication. If the OTP2 entered by the customer matches the one stored in the database, the authentication is successful.

**Fig. 2** One-time password generation



#### **4.5 Fingerprint Biometric Authentication**

Fingerprint biometric authentication [19] is the most matured biometric technique that enables customers to access the online service using fingerprint images. The fingerprint biometric authentication [20] consists of two phases: the initial enrollment phase and the verification phase. During the enrollment phase, the fingerprint image of the customer is captured, processed the fingerprint image to extract the features, and stored the characteristics of each user in the database.

During the verification phase, the user's fingerprint image is captured using the fingerprint sensor or using a smartphone camera. The captured fingerprint image is preprocessed to eliminate the redundant information. The user's features are extracted and compared with the features of the specific user stored in the database. If the extracted features and the features of the user stored in the database match, the authentication is a success.

### **5 Performance Analysis of Anomaly Detection with Precision–Recall Curves**

Anomaly detection [21] identifies the abnormalities in events or behaviors. The recognition of anomalies is critical in the case of cyber-attacks. The standard metrics to evaluate the anomaly detection are precision and recall [22]. The precision and recall are defined using Eqs. 1 and 2.

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (1)$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \quad (2)$$

where TP is true positive, FP is false positive, and FN is false negative. Precision is the division of all real anomalies by detected anomalies, and recall is the division of real anomalies detected successfully. Recall and precision are complementary, and it is helpful to evaluate the performance.

Precision is the measure of how well the anomaly-based algorithm picks out the anomalies. The anomaly detection algorithm returns purported anomalies ( $S(t)$ ) for threshold  $t$ . Some of the purported anomalies are real, and some are not real. The proportion of real anomalies is the precision.

The recall is the measure of how well the anomaly-based algorithm identifies all anomalies. The recall is the percentage of the complete set of anomalies from  $S(t)$ .

The PR curve [23] or precision–recall curve is the trade-off connecting the recall and precision for anomaly detection algorithm. For a threshold range  $t$ , the precision and recall are determined, and the curve shows the trade-off.

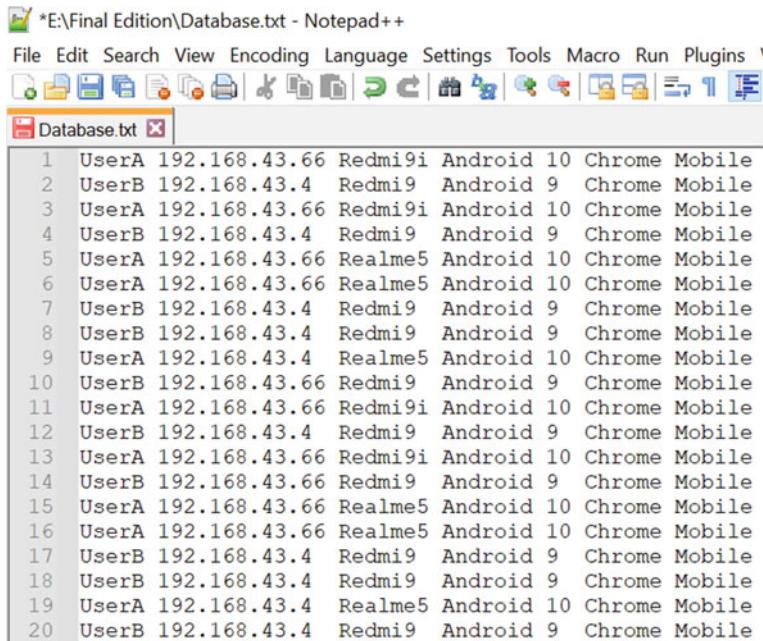
## 5.1 Precision and Recall of Anomaly Detection Algorithm

The profile of user A stored in the database is  $\langle$  User name, User A  $\rangle$ ,  $\langle$  IP Address, 192.168.43.66  $\rangle$ ,  $\langle$  Device, Redmi 9i  $\rangle$ ,  $\langle$  Browser, Chrome Mobile  $\rangle$  and for user B is  $\langle$  User name, User B  $\rangle$ ,  $\langle$  IP Address  $\rangle$ ,  $\langle$  192.168.43.4  $\rangle$ ,  $\langle$  Device, Redmi 9  $\rangle$ ,  $\langle$  Browser, Chrome Mobile  $\rangle$ . The random sample log results of User A and User B are shown in Fig. 3. There are 10 samples for each user, i.e., for User A and User B. The precision and recall for the different thresholds of the log result shown in Fig. 3 are reported in Table 2. The threshold value is the number of samples taken from the log result for analysis. The precision of the normal anomaly detection algorithm is calculated using Eq. 1

$$\text{Precision for User A} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Precision for User A} = \frac{5}{5+5} = 50\%$$

$$\text{Precision for User B} = \frac{6}{6+4} = 60\%$$



```
*E:\Final Edition\Database.txt - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins 
Database.txt
1 UserA 192.168.43.66 Redmi9i Android 10 Chrome Mobile
2 UserB 192.168.43.4 Redmi9 Android 9 Chrome Mobile
3 UserA 192.168.43.66 Redmi9i Android 10 Chrome Mobile
4 UserB 192.168.43.4 Redmi9 Android 9 Chrome Mobile
5 UserA 192.168.43.66 Realme5 Android 10 Chrome Mobile
6 UserA 192.168.43.66 Realme5 Android 10 Chrome Mobile
7 UserB 192.168.43.4 Redmi9 Android 9 Chrome Mobile
8 UserB 192.168.43.4 Redmi9 Android 9 Chrome Mobile
9 UserA 192.168.43.4 Realme5 Android 10 Chrome Mobile
10 UserB 192.168.43.66 Redmi9 Android 9 Chrome Mobile
11 UserA 192.168.43.66 Redmi9i Android 10 Chrome Mobile
12 UserB 192.168.43.4 Redmi9 Android 9 Chrome Mobile
13 UserA 192.168.43.66 Redmi9i Android 10 Chrome Mobile
14 UserB 192.168.43.66 Redmi9 Android 9 Chrome Mobile
15 UserA 192.168.43.66 Realme5 Android 10 Chrome Mobile
16 UserA 192.168.43.66 Realme5 Android 10 Chrome Mobile
17 UserB 192.168.43.4 Redmi9 Android 9 Chrome Mobile
18 UserB 192.168.43.4 Redmi9 Android 9 Chrome Mobile
19 UserA 192.168.43.4 Realme5 Android 10 Chrome Mobile
20 UserB 192.168.43.4 Redmi9 Android 9 Chrome Mobile
```

**Fig. 3** Sample log results of users

**Table 2** Anomaly and recall of user A and user B at different thresholds (anomaly detection algorithms)

Threshold	User A			User B		
	Anomaly	Precision (%)	Recall (%)	Anomaly	Precision (%)	Recall (%)
1	✗	100	20	✗	100	16.66
2	✗	100	40	✗	100	33.33
3	✓	66.66	40	✗	100	50
4	✗	75	60	✗	100	66.66
5	✓	60	60	✓	80	66.66
6	✓	50	60	✓	66.66	66.66
7	✗	57.14	80	✓	57.14	66.66
8	✓	50	80	✓	50%	66.66
9	✗	55	100	✗	55.55	83.33
10	✓	50	100	✗	60	100

## 5.2 Precision and Recall of Modified Anomaly Detection Algorithm

The primary profile of User A stored in the database is ⟨User name, User A⟩, ⟨IP Address, 192.168.43.66⟩, ⟨Device, Redmi 9i⟩, ⟨Browser, Chrome Mobile⟩, and the secondary profile of User A stored in the database is ⟨User name, User A⟩, ⟨IP Address, 192.168.43.4⟩, ⟨Device, Reamlme 5⟩, ⟨Browser, Chrome Mobile⟩. The primary profile of User B stored in the database is ⟨User name, User B⟩, ⟨IP Address, 192.168.43.4⟩, ⟨Device, Redmi 9⟩, ⟨Browser, Chrome Mobile⟩, and the secondary

**Table 3** Anomaly and recall of User A and User B at different thresholds (modified anomaly detection algorithm)

Threshold	User A			User B		
	Anomaly	Precision	Recall	Anomaly	Precision (%)	Recall (%)
1	✗	100%	14.28%	✗	100	10
2	✗	100%	28.57%	✗	100	20
3	✓	66.66%	28.57%	✗	100	30
4	✗	75%	42.85%	✗	100	40
5	✗	80.0%	57.14%	✗	100	50
6	✓	66.66%	57.14%	✗	100	60
7	✗	71.42%	71.42%	✗	100	70
8	✓	62.5%	71.42%	✗	100	80
9	✗	66.66%	85.71%	✗	100	90
10	✗	70.0%	100%	✗	100	100

profile of User B stored in the database is  $\langle$ User name, User B $\rangle$ ,  $\langle$ IP Address, 192.168.43.66 $\rangle$ ,  $\langle$ Device, Redmi 9 $\rangle$ ,  $\langle$ Browser, Chrome Mobile $\rangle$ . The precision and recall of the modified anomaly detection algorithm for the different thresholds of the log result shown in Fig. 3 are reported in Table 3.

The precision of the modified anomaly detection algorithm is calculated using Eq. 1

$$\text{Precision for User A} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{Precision for User A} = 7/(7 + 3) = 70\%$$

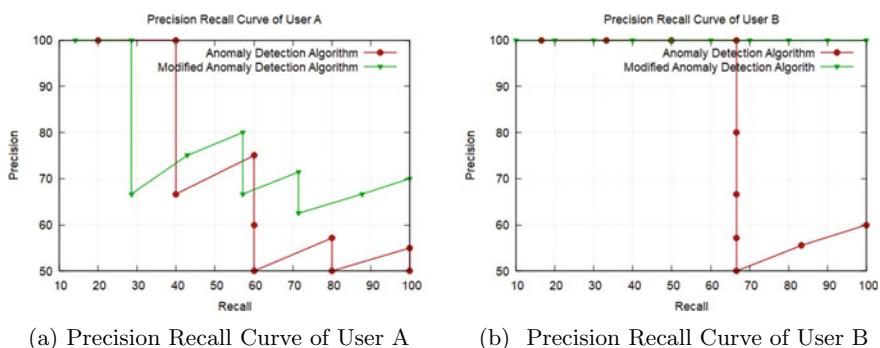
$$\text{Precision for User B} = 10/(10) = 100\%$$

### 5.3 Results

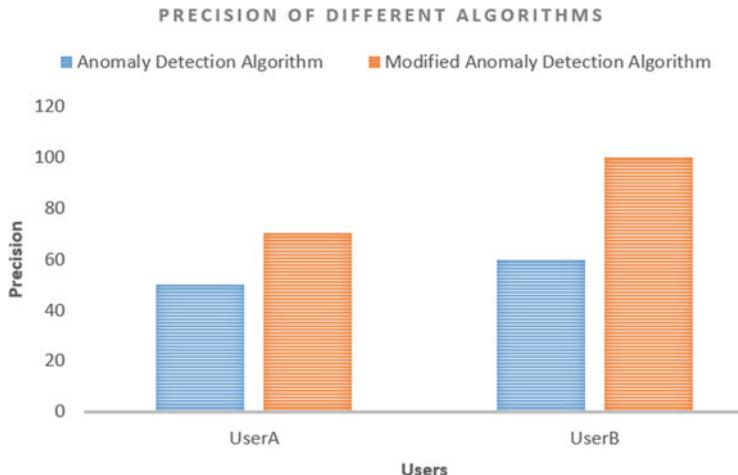
The precision of anomaly-based detection and modified anomaly-based detection is reported in Table 4. As reported in Table 4, the precision of the modified anomaly detection algorithm for User A is 70%, whereas the anomaly detection algorithm is only 50%. The precision of the modified anomaly detection algorithm for User B is 100%, whereas the anomaly detection algorithm is only 60%. The precision-recall curve (PR curve) of User A is shown in Fig. 4a, and for user B is shown in Fig. 4b. According to the PR curve, the algorithm graph, which plots at the top, indicates good accuracy. The PR curve of the modified anomaly detection algorithm plots at the top indicates better accuracy. The introduction of a secondary profile enhances anomaly-based detection precision and reduces the burden of higher-level

**Table 4** Comparison of anomaly detection algorithm and modified anomaly detection algorithm

Algorithm	Precision User A (%)	Precision User B (%)
Anomaly detection algorithm	50	60
Modified anomaly detection algorithm	70	100



**Fig. 4** Precision–recall curve of users



**Fig. 5** Precision of anomaly detection algorithm and modified anomaly detection algorithm

authentication for genuine users. The bar graph of the precision of anomaly-based detection and modified anomaly-based detection is shown in Fig. 5.

## 6 Security Analysis

Nowadays, most online banking uses two-factor authentication [9] such as passwords and mobile OTP. The attacker tricks bank customers and collects the user name and password using phishing attacks [10]. Using the SIM swap fraud [11], the attacker can quickly obtain the OTP required for the online banking authentication. The authentication of bank customers using passwords and mobile OTP is not secure, and second-factor authentication like OTP hassles the customer.

The proposed Internet banking system authenticates the customer by a multi-factor authentication method. If there are no anomalies, the user can log in to the Internet banking system using the username and password. When the browser change or operating system change anomaly is detected, the customer goes through the OTP with a QR code security mechanism. The user must go through the OTP with QR code and biometric authentication security method if IP address change or device change anomaly is detected.

The mobile OTP with QR code authentication method generates OTP2 using OTP1 and the IMEI number of the mobile phone. The user is authenticated using OTP2, so even if the attacker gets OTP1 using a SIM swap fraud attack, he cannot authenticate the bank server. The mobile OTP with QR code authentication method defeats the SIM swap fraud attacks.

**Table 5** Comparison of security and performance of different methods

Authentication method	Security factor	Performance complexity
Single factor authentication (user name and password) [24]	$S(n)$	$P(n)$
Two-factor authentication (user name and password and mobile OTP) [25]	$S(2*n)$	$P(2*n)$
Multi-factor authentication (user name and password, mobile OTP, and biometric) [26]	$S(3*n)$	$P(3*n)$
Proposed method (anomaly-based user name and password or mobile OTP and biometric)	$S(3*n)$	$P(n)$

## 7 Discussion

In this paper, we propose a modified anomaly-based multi-factor authentication for the online mobile banking system. This method introduces a different level of authentication based on the anomaly detected. For a genuine user, if no anomalies are detected, the user can log in with the user name and password authentication method, and it reduces the burden for authentication. The anomaly detection algorithm is modified to adapt to the transportability of mobile phones.

The precision of anomaly detection and modified anomaly detection algorithm is analyzed, and it is found that the precision of the modified anomaly detection algorithm is far better than the anomaly detection algorithm. The precision of the modified anomaly detection algorithm for User B is 100%.

The different online banking authentication methods and their security and performance are compared in Table 5. The two-factor and multi-factor authentication increase the security while increasing the customer's burden (performance complexity). The proposed anomaly detection-based authentication method introduces the high-level authentication only when the anomaly is detected; thereby, it increases the security factor to  $S(3 * n)$  and reduces the performance complexity to  $P(n)$ .

Phishing attempts to get user's login credentials like username and password by disguising themselves as a trusted entity in the banking sector. SIM swap is a new cyber fraud, where the attacker collects the personal data of the bank customer and gets a new SIM card. The attacker can easily steal the user's login credentials like user name and password using phishing attack and the OTP using the SIM swap fraud attack. In the anomaly-based fraud detection algorithm, the login attributes like IP address, device, cookie, operating system, and browser are used to generate and update the user's profile.

## 8 Conclusion

The Internet banking users increase day by day, and the security of online banking transaction attains particular interest. The phishing and SIM swap fraud attacks defeat online banking authentication security with user name, passwords, and OTP. In this paper, an anomaly detection-based multi-factor authentication method is proposed, in which the level of security mechanism introduced to the user depends on the anomaly detected. If no abnormalities are detected, the user can log in with the user name and password. For browser change or operating system change anomaly, OTP with QR code is added. For IP address change or device change, both OTP and biometric authentication security are initiated. The anomaly detection algorithm is modified by including a secondary profile that contains the user's most frequently used login attributes. The precision of the modified anomaly detection algorithm is better concerning precision and PR curve analysis. The proposed method reduces the user's login burden and provides better security for the online banking system.

## References

1. Liao, Z., Cheung, M.T.: Internet-based e-banking and consumer attitudes: an empirical study. *Inf. Manag.* **39**(4), 283–295 (2002)
2. Mehra, P.: Controlling attacks and intrusions on internet banking using intrusion detection system in banks. *Int. J. Adv. Res. Comput. Commun. Eng.* **4**(11), 346–348 (2015)
3. Dhamija, R., Tygar, J.D., Hearst, M.: Why phishing works. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 581–590 (2006)
4. The anti-phishing working group. <http://www.antiphishing.org/>
5. Jordaan, L., von Solms, B.: A biometrics-based solution to combat sim swap fraud. In: International Workshop on Open Problems in Network Security, pp. 70–87. Springer (2010)
6. Brindha, G., G.N.: Secure banking using QR code. *Int. J. Adv. Res. Comput. Eng. Technol.* **3**(12), 4302–4306 (2014)
7. Jain, A.K., Nandakumar, K.: Biometric authentication: system security and user privacy. *IEEE Comput.* **45**(11), 87–92 (2012)
8. Rid, T., Buchanan, B.: Attributing cyber attacks. *J. Strateg. Stud.* **38**(1–2), 4–37 (2015)
9. Mail, A., Box, D.: Two factor authentication (2017)
10. Yeboah-Boateng, E.O., Amanor, P.M.: Phishing, smishing & vishing: an assessment of threats against mobile devices. *J. Emerg. Trends Comput. Inf. Sci.* **5**(4), 297–307 (2014)
11. Awale, S.M., Gupta, P.G.: Awareness of sim swap attack. *Int. J. Trend Sci. Res. Dev.* **4**, 995–997 (2019)
12. Sivaganesan, D.: A data driven trust mechanism based on blockchain in IoT sensor networks for detection and mitigation of attacks. *J. Trends Comput. Sci. Smart Technol. (TCSST)* **3**(01), 59–69 (2021)
13. Tsai, C.L., Chen, C.J., Zhuang, D.J.: Secure OTP and biometric verification scheme for mobile banking. In: 2012 Third FTRA International Conference on Mobile, Ubiquitous, and Intelligent Computing, pp. 138–141. IEEE (2012)
14. Yildirim, N., Varol, A.: A research on security vulnerabilities in online and mobile banking systems. In: 2019 7th International Symposium on Digital Forensics and Security (ISDFS), pp. 1–5. IEEE (2019)
15. O'neil, E.J., O'neil, P.E., Weikum, G.: The lru-k page replacement algorithm for database disk buffering. *ACM Sigmod Record* **22**(2), 297–306 (1993)

16. Park, J.S., Sandhu, R.: Secure cookies on the web. *IEEE Internet Comput.* **4**(4), 36–44 (2000)
17. Ford, M., Boucadair, M., Durand, A., Levis, P., Roberts, P.: Issues with IP address sharing. *IETF Request Comment 6269* (2011)
18. Kumar, K., Kaur, P., Amritsar, G.: Vulnerability detection of international mobile equipment identity number of smartphone and automated reporting of changed IMEI number. *Int. J. Comput. Sci. Mob. Comput.* **4**(5), 527–533 (2015)
19. Sharma, L., Mathuria, M.: Mobile banking transaction using fingerprint authentication. In: 2018 2nd International Conference on Inventive Systems and Control (ICISC), pp. 1300–1305. IEEE (2018)
20. Manoharan, J.S.: A novel user layer cloud security model based on chaotic Arnold transformation using fingerprint biometric traits. *J. Innov. Image Process. (JIIP)* **3**(01), 36–51 (2021)
21. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. *ACM Comput. Surv. (CSUR)* **41**(3), 1–58 (2009)
22. Tatbul, N., Lee, T.J., Zdonik, S., Alam, M., Gottschlich, J.: Precision and recall for time series. *arXiv preprint arXiv:1803.03639* (2018)
23. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd international conference on Machine learning, pp. 233–240 (2006)
24. Bani-Hani, A., Majdalweieh, M., AlShamsi, A.: Online authentication methods used in banks and attacks against these methods. *Procedia Comput. Sci.* **151**, 1052–1059 (2019)
25. Sadri, M.J., Asaar, M.R.: An anonymous two-factor authentication protocol for iot-based applications. *Comput. Netw.* **199**, 108460 (2021)
26. Jacomme, C., Kremer, S.: An extensive formal analysis of multi-factor authentication protocols. *ACM Trans. Privacy Secur. (TOPS)* **24**(2), 1–34 (2021)

# A Sparrow Search Algorithm for Detecting the Cross-layer Packet Drop Attack in Mobile Ad Hoc Network (MANET) Environment



S. Venkatasubramanian, A. Suhasini, and N. Lakshmi Kanthan

**Abstract** Due to the wide range of applications and ease of deployment, mobile ad hoc networks (MANETs) have become increasingly popular. There is no need for a fixed infrastructure in this network. Mobile nodes of this network connect via a wireless channel making it very vulnerable to various attacks. *Purposes:* The packet drop attack is a famous attack type. The easiest way to cause service denial is in such dynamic networks as a packet drop or gray hole attack. This attack reflects the malicious node to be the shortest route, and entire packets are received while the designated packets are dropped to provide the user with the wrong service. It is a special type of attack that defends the system and the user from misuse. When a network is attacked, it suffers from network performance. *Methods:* A sparrow search algorithm for a cross-layer packet drop-off attack (CLPDM-SSA) is proposed in this paper. This proposal is used to identify a malicious node within a real-world data acquisition system that was subjected to a packet drop (PD) attack with a cluster-based meta-heuristic detection device. *Results:* NS-2 is used for implementation, and the validation is carried out in terms of throughput, packet delivery ratio (PDR), mean PD, and false positive rate. *Conclusion:* After integrating SSA intelligence, the results reveal considerable gains. The algorithm detects false positives of suspected malicious nodes by analyzing the CPU and the memory.

**Keywords** Cross-layer packet drop attack · Mobile nodes · Malicious node · Packet delivery ratio · Sparrow search algorithm

---

S. Venkatasubramanian (✉) · N. Lakshmi Kanthan

Department of Computer Science and Engineering, Saranathan College of Engineering, Trichy, Tamil Nadu 620012, India

e-mail: [veeyes@saranathan.ac.in](mailto:veeyes@saranathan.ac.in)

N. Lakshmi Kanthan

e-mail: [lakshmikanthan-cse@saranathan.ac.in](mailto:lakshmikanthan-cse@saranathan.ac.in)

A. Suhasini

Department of Computer Science and Engineering, Annamalai University, Annamalainagar, Tamil Nadu 608002, India

## 1 Introduction

A MANET is a network that consists of several free and independent nodes [1], often made up of mobile strategies or further nodes which can be organized in different ways and functions without the severe top-down governance of a network. MANET configurations are various and the possibility for this type of network is under investigation. The intrinsic vulnerability of MANET creates additional security problems that primarily affect the network and protocol stack data connection layers. Each packet needs to be transmitted quickly via an intermediate node and must travel to the destination from the source. The routing detection or maintenance procedure may not be targeted by malicious routing attacks if the routing protocol requirements are not met [2, 3]. This upsurges the probability of attacks such as wiping, spoofing, DoS, and imitation. Due to the unique features of ad hoc mobile networks, MANET cannot use security approaches to safeguard stationary networks [4]. New risks, such as internal malicious node assaults, Byzantine, cross-layer attacks [5], packet drop attacks, and wormhole attacks, are challenging to protect against. The objective of this work is to focus primarily on cross-layer packet drop attacks by clustering nodes with each cluster having a cluster head.

### 1.1 *Packet Drop (Gray Hole) Attack*

A drop packet attack is a DoS attack when a node drops data packets at a specific point in time to a certain network destination every  $t$  second [6]. It differs marginally from a black hole in that the black hole is a widely denied service attack that, because of its very special key restrictions, rejects packets. The attack causes packets to be dropped [7]. This is an active attack. The attacking node initially agrees but does not do so and then maliciously starts to act [8]. Initially, a routine attacker node will drive RREPs to extra nodes to appeal routine Request (RREQ) messages accept or accept packets sent and then drop a few or any packets to presentation a DoS [9]. If neighborhood nodes try to send data packets over an attack or victim's nodes, they can lose their association and want to find out or reconstruct a route by using the RREQ message. The attacking node sends route response messages (RREP) [10] to establish a route. This process continues until the attacking node reaches its aim, such as the consumption of power and bandwidth.

### 1.2 *Cross-layer Attacks*

Cross-layer attacks are generated as a result of a lack of communiqué between the MAC layer and the routing layer. This sort of attack is often launched from the MAC layer. The network performance is being severely harmed as a result of this attack.

There are several forms of assaults that result in cross-layer attacks. An attacker can induce network congestion by creating particular traffic patterns or generating an excessive quantity of traffic [11]. These attacks originate at the MAC layer and manifest as DoS attacks on the routing level. Sinking is a harmful behavior in which the attacker just drops the next packet without being prepared. Sinking behavior is maliciously expected to either greedily preserve resources, such as power, processing time, and so on, or to disrupt the routing system by losing essential packets. To reduce this packet drop attack by decreasing latency and reducing false positives, it is important to detect a hostile node. As a result, the suggested approach in this article is based on cluster-based technique and false positive; SSA is used for recognizing harmful nodes as assumed false positive for determining whether an alleged node is malicious on CPU use time.

The rest of the paper is prearranged as shadows: Sect. 2 displays the associated study of existing techniques in MANET. The explanation of the proposed methodology with finding the malicious nodes is presented in Sect. 3. The validation of the proposed methodology with existing techniques is provided in Sect. 4. Lastly, the conclusion of the research work is described in Sect. 5.

## 2 Related Works

Vinayagam et al. [12] presented a cross-layer method to identify various types of attacks and improve MANET security. For feature subset selection, the artificial bee colony (ABC) algorithm is integrated with this technique. For packet delivery, a cross-layer tool is created among the MAC layer and the routing layer. Datasets for the MAC and routing layers are collected from individual nodes in this section. The most significant features are chosen with the aid of feature selection.

Manikandan et al. [13] advocated using certificate termination to identify and remediate a rogue node in the routing situation. The proposed solution is divided into three sections. This includes cluster formation, trust, and certificate revocation. During the cluster creation process, a subset of the smallest hopping sum nodes is grouped into a single cluster node, and the cluster node is tracked using a cluster head. The estimate method is used to amount a node's flat of confidence in a trust [14]. During the certificate revocation phase, a node's trust is associated with a threshold value to determine if it is trustworthy or untrustworthy. If a node's fundamental trust rating exceeds a certain level, the node is deliberated untrustworthy, which nodes are eventually eliminated from network routing. Das et al. [15] introduced an energy use calculation method. Earlier and after a data packet communication in this way, the energy level of each knot is maintained. The data packet is sent to an energy level target node. For the reason that a lower-energy node cannot transfer a data packet to a neighboring node, the data packet transmission procedure selects a node with a higher-energy level [16].

Garikipati et al. [17] proposed MANET to notice and mitigate black-hole attacks by mobile trust point approaches. The routing technology is used to track CHs assigned to a collection of nodes using the mobile trust point technology. If the route node (RREP) does not respond, the node is regarded as an attacker node. The cluster head transmits data to cluster nodes and expects maximum timestamps [18]. Following proof of identity of the attacker node, the CH will transmit a grade of the black hole to the mobile faith. The nodes in the black hole are removed from the driving area. For the enhanced metering infrastructure intrusion detection model based on convolutional neural networks (CNN) and long short-term memory (LSTM) network cross-layer feature fusion, Yao et al. [19] have presented a model. The KDD Cup 99 and NSL-KDD datasets are used in the experiments. As a result, the model is built from cross-layer components of CNN and LSTM. The CNN component detects regional characteristics to gain global features, while the LSTM component uses memory function to obtain periodic features. It is possible to combine the two types of intrusion detection features to get a more comprehensive set of features with multi-domain properties.

An AMI intrusion detection ideal based on the ELM was proposed by Yuancheng et al. [20]. When using this model, you get faster detection times without sacrificing accuracy because it makes use of online sequence training. Face the “black-hole onslaught” of AMI. The gain ratio evaluation approach was used in the experiment to lower the sample dataset’s size. Classes are created using the OS-ELM method, which uses the datasets to create new classes. After then, a slew of tests is run to determine the system’s ideal algorithm parameters.

An enhanced K-nearest neighbor (KNN) algorithm was proposed by Khan et al. [21] as a hybrid intrusion detection model. The Bloom filter is then applied to the AMI system to look for anomalous data. To begin, we used several data standardization and scalability strategies in pre-processing. To enhance anomaly detection, dimensionality reduction algorithms are used second. The dataset was balanced using an altered nearest neighbor rule technique, and a signature database was generated using the Bloom filter by noting the system for a specified period during which irregularities did not occur. Last but not least, we developed a hybrid anomaly detection system that coupled our package contents-level detection with another instance-based learner to detect new assaults. Venkatasubramanian et al. [3, 6] presented the use of a ticket-ID-based clustering technique, in which each node is assigned a TID based on the parameters such as bandwidth, delay, and throughput. But, their work does not detect any attack in the clusters.

Concerning constrained wireless communication links in random-access wireless systems, Mehta [22] has studied the cross-layer paradigm, which relies on consolidated network coding design and efficient resource allocation to those links. Non-linear programming preparation with numerous constraints was used in this framework for stable entropy and resource allocation measures in wireless networks that were under stress. There are four layers in the network protocol stack in Mehta

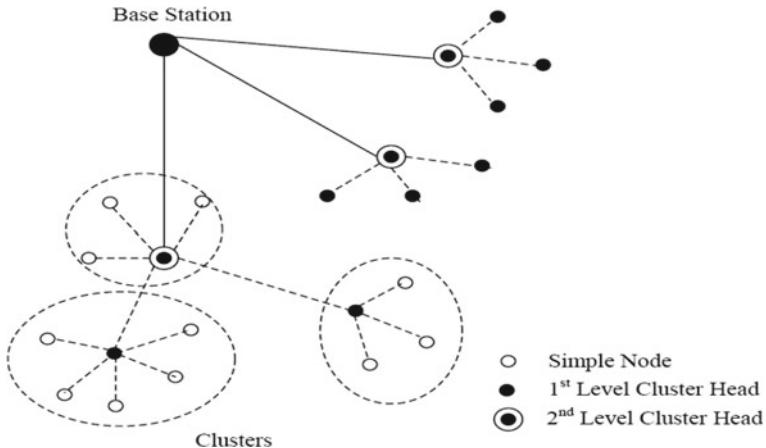
[23], and the adaptive Huffman coding technique is used to determine the probability of nodes persisting in the network. Convex optimization technique used to derive optimal data compression and network service quality constraints for this multiple-flow network scenario with random mobility for the generic routing system.

Shakya [24] developed the adaptive routing protocol that included the genetic bee colony and ant colony algorithm for improving the network lifetime and minimizing the energy consumption. NS-2 was used to implement the protocol and validated by using throughput, energy utilization, and network lifetime. However, the communication was not secured by this adaptive routing protocol. Manoharan [25] proposed the new algorithm called water cycle with evaporation, condensation, and precipitation (WC-ECP) for improving the probabilistic neural network (PNN)'s accuracy. In the search space of exploration and exploitation, the weights of PNN were fine-tuned by WC-ECP and tested with five UCI datasets such as Parkinson's, heart, breast cancer, diabetes, and liver disorder. However, the model is not developed for network implementation and the security of nodes.

Jacob [26] proposed the artificial bee colony algorithm to enhance the throughput of the wireless multi-channel networks. The synchronous and decentralized routing decisions were made by the simple behavior of bee agents. The overall routing protocol was increased by using the breadth-first search variant. The algorithm was not supported for multiple paths communication, because a single threshold value was used. Mugunthan [27] presented the low-energy infected susceptible epidemic model for the various processes such as charging, reinfection, and removal in wireless rechargeable sensor networks (WRSNs). The simulation of disease-free points and epidemic equilibrium of global and local stabilities were carried out and analyzed. The model was well-suited only for the homogeneous static network. The safety hazards and random model problems were not analyzed.

### 3 Detection of Cross-layer Packet Drop Attack using SSA

The attack techniques that objective the interfaces among the MAC and the routing layer have not been entirely lectured, resulting in a new class of assaults known as cross-layer attacks. These assaults create a DoS attack in the objective layer, resulting in a progressive decline in the presentation in terms of increased latency, poor throughput, reduced PDR, and so on. In this paper, we looked at how a cross-layer plan that incorporates both the MAC and the network layer may give a recognition approach and enhance network presentation. Because of its high energy, it is a centralized monitoring method in which the cluster works as a data collector. As a result, the CH is chosen depending on the energy factor. The steps of our technique are as follows: Stage I: cluster head (CH) selection; Stage II: malicious nodes detection; and Stage III: confirmation.



**Fig. 1** Sample of hierarchical clustering

### Stage I: Cluster Head (CH) selection

The topology of MANET changes quite often due to node mobility. An adaptive  $k$ -level hierarchical clustering approach is utilized in this step of cluster creation [22]. Figure 1 depicts an example of hierarchical clustering. A  $k$ -level command chain of clusters is most beneficial when compared to localized low-energy clusters to reduce network power consumption (LLC). This combination is particularly feasible in comparison with the reduction in system vitality (LLC).

In addition, it ensures the system and environmental change in the routing option are independent. Of course, by training new clusters, the system responds quickly to topological changes. It contains no routes, and nodes use the choice of the neighborhood to select the best route. The X-LLC cluster is thus used because it enables the size of the cluster to be reduced by examining the radius via the use of diverse power levels. In comparison with traditional hierarchical methods for cluster formation, this offers a significant advantage for energy transmission reduction [28].

```

Algorithm: SelectionPhase 1:
Loop: ForeachCandidateNode (Nci)
GeneratesRandomvalueK = K{0,1}
If Pth > K
Nci can become a candidatenode
At T0 = 0 transmitsMsg(Ni, Rei)
Loop: ForeachCandidateNodes
If Rej > Rei
Then
T0 expires&T1 starts
Else
T0 continues and Nci is out of the selectionprocess
Nci is considered to be clusterhead & SendsMsg to Rad
Endloop
Else
Nct to be idle till the selection process to finish.
Endloop

AssociationPhase 2:
Loop : For every Ns
SendsMsg (ECo) & Sets T2 = 0
NsClusterHeadsResponds (rcei)
Note the timing of each response
If Trcei < Trcej
Then NChi is selected as the cluster to join
Endif
Endloop

```

Cluster creation consists of two stages: election and organization.

#### (i) Election stage

We evaluated  $k \geq 1$  level of CH and  $k$  unique selection and connotation steps in this stage, and LLC is derived by selecting  $k = 1$ .

#### Procedure

Step 1: Each node sets the sum of assortment messages acquired by a contestant node to one.

Step 2:  $m$  is set to 0 and a consistently distributed random number  $u$  among 0 and 1 is generated.

Step 3: In comparison with the threshold,  $P_{ji}$  determines the likelihood that nodes would participate in the election phase.

Step 4: If  $P_{ji}$  is previously indicated, the node develops a candidate CH and partakes in the election period; otherwise, it remains silent until the election procedure is completed.

Step 5: The node is enabled in a start m old commencing from the value  $a$ .

Step 6: Each candidate node sends a promoting note with communication power  $P_w$  that spans a radius  $R_w$  of space.

Step 7: Each applicant hub collects marketing messages from several rival hubs in the area and tallies the received messages by increasing *them*.

Step 8: At the end of the period, the applicant will change the promotional timer per  $sb$ , where  $sb$  is based on  $m$  and node energy.

Step 9: Finally, after  $sb$  deceases, the candidate timer node becomes the leading  $i$ th cluster and sends a  $Pw$  transmission-powered advertising message.

Step 10: Alternatively, if the timer continues to count down and the node gets an advertising message, it waits for the election procedure to complete.

Step 11: If the CH at the  $i$ th level is not elected, they take an interest in the  $(i + 1)$ th level CH selection; otherwise, they simply stay CHs at the  $i$ th level [23].

## (ii) Organization stage

Following the conclusion of the election method, the organization stage commences and includes  $k$ -specific association sub-periods that are performed in the best down way commencing with the base station and progressing to uncomplicated hubs. At that point, the  $(k - 1)$ th level CHs for the nearest  $k$ th level cluster head, who responds by providing the TDMA table; the operation is repeated down to the regular ad hoc node level.

Furthermore, the accompaniment stays constant.

1. Each CH is in charge of a few hubs.
2. Simple hubs identify the closest CH with a single leap separation.
3. The transmission kind of simple hubs might be reduced in comparison with the one required by LLC. As a result, transmission requires less power and inter-cluster blockage decreases.

The optimum sum of levels for a certain application is determined by the features of the transfer, the offered hierarchy above, the kind of hubs, the entire degree, the available speed, and residual energy.

### **3.1 Phase II: Discovery of the Malicious Node using Sparrow Search Algorithm (SSA)**

The sparrows are sociable birds of several kinds. They are found in maximum regions of the ecosphere and like to reside in areas where there is human life. Furthermore, they are omnivorous birds that mostly graze on grain seeds or weeds [29]. The fact that sparrows are frequent resident birds is widely known. Unlike various other tiny birds, the sparrow is very clever and has a long memory.

#### *Mathematical model and algorithm*

Based on the preceding sparrow description, we can create a mathematical model to build the sparrow search method. For the sake of simplicity, we idealized the following sparrow behavior and developed matching rules.

1. Generally, manufacturers have high-energy reservoirs and provide all scroungers with foraging areas or directions. It identifies regions in which rich food supplies can be obtained. The energy supply level relies on the evaluation of the individual's fitness values.

2. The characters begin chirping as frightening messages once a sparrow spots the predator. If the alarm is higher than the safety threshold, all the scroungers must be led to the safe area by the producers.
3. As long as every sparrow seeks better food resources, it can become a producer, but the proportion of makers and idlers in the population at large is unchanged.
4. Higher-energy sparrows as producers would be acting. More than one hungry scrounger is likely to fly to other places to get more power.
5. The scroungers follow the manufacturer that can search for food for the best food. Meanwhile, some scroungers can monitor the producers constantly and contest for food to increase their rate of predation.
6. Sparrows travel quickly toward the safe place at the periphery of the group, where they are aware of the danger, while sparrows stroll allegedly in the center of the group to be closely related.

We have to employ virtual sparrows to find food in the simulation experiment. In the following matrix, the location of the sparrows:

$$X = \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,d} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,d} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n,1} & X_{n,2} & \cdots & X_{n,d} \end{bmatrix} \quad (1)$$

where  $n$  is the sparrow number and  $d$  the dimensions of the variables which should be optimized are shown. The following vector can therefore express the fitness value of all sparrows:

$$F_X = \begin{bmatrix} f([X_{1,1} \ X_{1,2} \ \cdots \ X_{1,d}]) \\ f([X_{2,1} \ X_{2,2} \ \cdots \ X_{2,d}]) \\ \vdots \ \vdots \ \vdots \ \vdots \\ f([X_{n,1} \ X_{n,2} \ \cdots \ X_{n,d}]) \end{bmatrix} \quad (2)$$

where  $n$  shows the sparrow number, and the  $F_X$  value of each row shows the individual's fitness value. In the SSA, the priority in the searching process is to get food from growers with greater fitness values. Furthermore, the producers are accountable for the search of food and the mobility of the whole population. The creators can therefore seek food in a wide variety of places than scroungers. Under rules (1 and 2), the producer location is updated as follows during every iteration:

$$X_{i,j}^{t+1} = \begin{cases} X_{i,j}^t \cdot \exp\left(\frac{-i}{\alpha \cdot \text{iter}_{\max}}\right) & \text{if } R_2 < \text{ST} \\ X_{i,j}^t + Q & \text{if } R_2 \geq \text{ST} \end{cases} \quad (3)$$

In which the current iteration  $t$  is indicated,  $j = 1, 2, \dots, d$ .  $X_{i,j}^t$  is the value of the  $i$ th sparrow's  $j$ th dimension at iteration  $t$ . The total amount of iterations is a constant,

namely the penultimate. A random number is  $\alpha \in (0, 1]$ . The value of alert  $R_2$  ( $R_2 \in [0, 1]$ ) and the safety threshold ST(ST  $\in [0.5, 1.0]$ ) is the R 2 alarm and the ST alarm.  $Q$  is a random number that is distributed as normal.  $L$  shows a matrix of  $1 \times d$  to  $d$  each of which is 1.

If  $R_2 < ST$ , i.e., no predators, the producer enters the broad mode of search.

When  $R_2 \geq ST$ , particular sparrows have found the predator, and entirely sparrows have to fly fast to extra safe zones.

As for screeners, the rules (4) must be applied and (5). As mentioned earlier, some scroungers are more often monitoring producers. Once the manufacturer has found good food, they consent to their current place immediately to contest for food. If you win, you can get the producer's food right away, otherwise, you will continue to follow the rules (5). This is described in the position update formulation for the scrounger:

$$X_{i,j}^{t+1} = \begin{cases} Q \cdot \exp\left(\frac{X_{\text{worst}}^t - X_{i,j}^t}{I^2}\right) & \text{if } i < \frac{n}{2} \\ X_p^{t+1} + \left|X_{i,j}^t - X_p^{t+1}\right| A^+ L & \text{otherwise} \end{cases} \quad (4)$$

where  $X_P$  is the manufacturer's optimal position.  $X_{\text{worst}}$  is the world's worst current place.  $A$  signifies the matrix of 1 to  $d$  for which 1 or  $-1$  for each element within is assigned randomly, and  $A^+ = A^T (AA^T)^{-1}$ . When  $i > n/2$ , to each element within the matrix  $(-1)$ . When  $I > n/2$ , it seems that the  $i$ th scrounger is more likely to be hungry with the poorer fitness value.

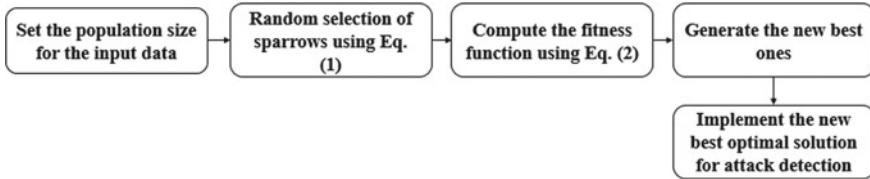
Randomly in the population are the initial places of these sparrows. The mathematical model can be represented as follows by rules (6):

$$X_{i,j}^{t+1} = \begin{cases} X_{\text{best}}^t + \beta \cdot \left|X_{i,j}^t - X_{\text{best}}^t\right| & \text{if } f_i > f_g \\ X_{i,j}^t + K \cdot \left(\frac{\left|X_{i,j}^t - X_{\text{worst}}^t\right|}{(f_i - f_w) + \varepsilon}\right) & \text{if } f_i = f_g \end{cases} \quad (5)$$

where the current ideal global location is  $X_{\text{best}}$ .  $\beta$  is a normal random number distribution with a mean of 0 and variance of 1 as a step-size control limit. A random number is  $K \in [-1, 1]$ . Here,  $f_i$  is the current sparrow's fitness value. The current global fitness values are  $f_g$  and  $f_w$  correspondingly.  $\mu$  is the least constant to prevent null-division-error.

For ease, when the sparrow is on the edge of the collection,  $f_i > f_g$  is shown.  $X_{\text{best}}$  stands for and is safe around the center of the population.  $f_i = f_g$  shows that sparrows are conscious of the risk and need to migrate closer to others, in the center of the population.  $K$  refers to the direction, and the sparrow is moving and the step-size control. Figure 2 shows the flowchart of SSA.

The following steps are used to detect rogue nodes and to determine the average packet drop rate.

**Fig. 2** Flow of SSA

Step 1: CH determines packet drop ( $D_i$ ) based on SSA accreditation.

Step 2: CH determines the average packet drop ( $D_{mn}$ ) founded on entirely cluster nodes data.

$$D_{mn} = \sum D_i / N_c \quad (6)$$

$N_c$  = Total number of nodes in the cluster

Step 3: CH associates the packet drop with a mean drop from specific nodes. If a packet drop is more than the mean drop, the nodes are referred to be malicious suspects.

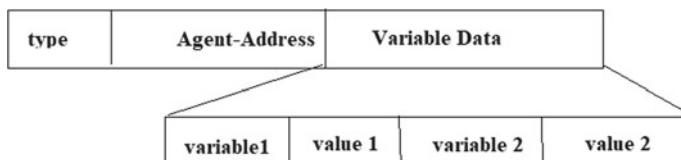
The following method shows malicious node identification:

```

begin procedure malicious node
If ( $D_i > D_{mn}$ )
    return true; //i as malicious node
else
    False // as normal node
Endif
End procedure
  
```

### Stage III: confirmation/False positive

When the node is believed as a malicious node, the suspected node may not be harmful since it is provided some specific packet seeking information about it. CH collects memory and CPU use by a specific packet agent of all suspected nodes, as depicted in Fig. 3.

**Fig. 3** CPU and memory utilization of the suspected node

CPU usage and memory usage are calculated for this node as a percentage for  $C_i$  and  $M_i$ .

## 4 Results and Discussion

The simulation work was led on the NS-2 simulator, which is commonly used in wireless network simulation. Table 1 illustrates the simulation strictures deliberated.

The proposed CLPDM-SSA performance is compared with adapting cross-layer approach for detecting and segregating malicious nodes (ACLDSSM) and CLPDM without SSA in terms of throughput, PDR, PD, and FPR.

### 4.1 Performance Analysis of CLPDM-SSA for Throughput

This is the sum of info packets sent to a destination node per unit of time from a source node. The validation of the suggested method for the throughput is shown in Tables 2 and 4; Fig. 4.

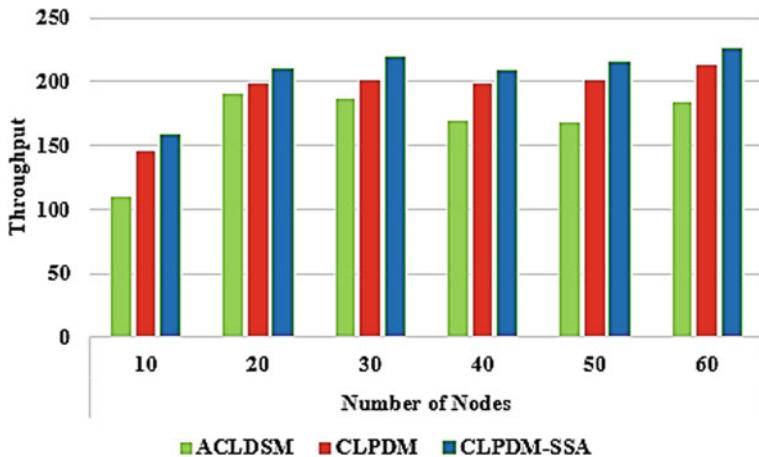
When the sum of nodes increases, the throughput of the proposed method is also increased. For instance, the CLPDM-SSA achieved 219 of throughput and CLPDM achieved 201 of throughput when the node is 30. But, the proposed method achieved 226 of throughput and CLPDM without SSA achieved only 213 and ACLDSSM achieved 184 of throughput when the node is 60. The result indicates clearly that

**Table 1** Simulation factors

Name	Value
Reproduction time	10 s
Network area	200 × 200
Traffic source	CBR
Number of nodes	10–60
Transmission range	260 m
Node speed	Constant (10 m/s)

**Table 2** Validation of proposed methodology for throughput

Methodology	Number of nodes					
	10	20	30	40	50	60
ACLDSSM	110	190	186	170	168	184
CLPDM	145	198	201	198	201	213
CLPDM-SSA	158	210	219	209	215	226



**Fig. 4** Graphical representation of proposed CLPDM-SSA in terms of throughput

SSA performance is higher than without SSA. This is because bad nodes drop packets. We test imitation with a variable sum of nodes since the sum of nodes in the CLPDM-SSA protocol is increased from 10 to 60 and the number of node numbers constantly decreases after node 20.

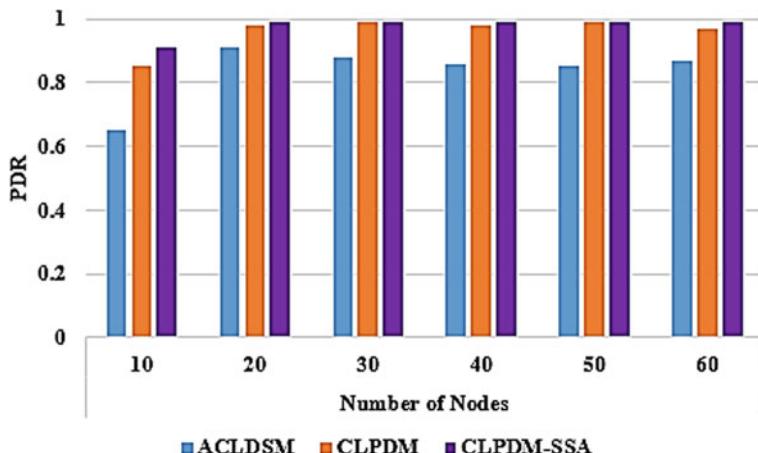
#### 4.2 Performance Analysis of CLPDM-SSA for PDR

This is the ratio between the destination and the source data packets. Table 3 and Fig. 5 display the investigational analysis of the proposed SSA methodology with existing protocols based on PDR.

Figure 5 displays the PDR value attained by our protocol is outstanding. When the node is 20, the ACLDSM achieved 0.91 PDR, CLPDM achieved 0.98 PDR, and proposed CLPDM-SSA achieved 0.99 PDR. When the nodes increase after 20, the ACLDSM achieved nearly 0.86–0.88 PDR. Our protocol is greater PDR than the ACLDSM and CLPDM protocols in all the different numbers of nodes. The protocols

**Table 3** Validation of proposed methodology for PDR

Methodology	Number of nodes					
	10	20	30	40	50	60
ACLDMS	0.65	0.91	0.88	0.86	0.85	0.87
CLPDM	0.85	0.98	0.99	0.98	0.99	0.97
CLPDM-SSA	0.91	0.99	0.99	0.99	0.99	0.99



**Fig. 5** Graphical representation of proposed CLPDM-SSA in terms of PDR

**Table 4** Validation of proposed methodology for false positive rate

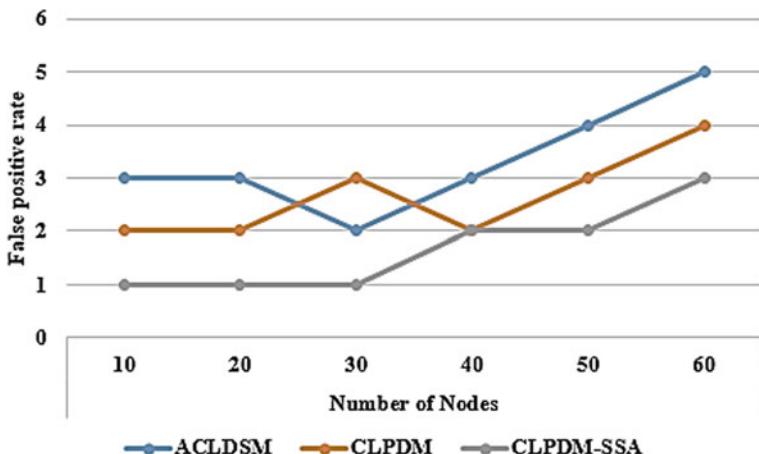
Methodology	Number of nodes					
	10	20	30	40	50	60
ACLDSTM	3	3	2	3	4	5
CLPDM	2	2	3	2	3	4
CLPDM-SSA	1	1	1	2	2	3

ACLDSTM and CLPDM have no swarm intelligence meta-heuristics, therefore the SSA introduction abruptly improves performance in our protocol.

#### 4.3 Performance Analysis of CLPDM-SSA for False Positive Rate

The FPR refers to well-detected mobile nodes as malicious nodes. The proportion is the erroneous node, divided by the entire node number, by the CPU inspection, and by the percentage utilization time. Table 4 and Fig. 6 show the analysis of the proposed protocol employing a false positive rate.

This is obvious from Fig. 6 that the fake positive rate is reduced in comparison with ACLDSM and CLPDM in our protocol, with our sparrow search algorithm. The improvement in our protocol is important. For the performance analysis of our CLPDM-SSA protocol with the existing ACLDSM and CLPDM protocol, we have calculated the performance, PDR, mean packet drop, and false positive rate. In total, the examination of findings demonstrates that CLPDM-SSA performs in terms of all



**Fig. 6** Graph of proposed CLPDM-SSA in terms of false positive rate

computed parameters above the existing methodology. The ACLDSM performance is very low because it does not implement optimization parameters. When SSA is implemented with the proposed CLPDM, its performance is suddenly improved in terms of all parameters.

## 5 Conclusion

MANET is a temporary working network, which is both significant and problematic in terms of protecting this network against assaults and other dangers with the enormous rise of MANETs. A reliable and fast device for detecting cross-layer packet drop attacks in ad hoc mobile networks is proposed in this research study. In this research, we explored an SSA-based cluster that uses a false positive CPU rate and memory usage to detect malicious nodes. This execution of the swarming method is superior and improves network recital measurement and reduces false positives in comparison with other non-swarm ways to discover malicious nodes. The FPR was CPU and memory usage time to check the actual malicious node further. Without the SSA protocol, our protocol works better than ACLDSM and CLPDM. We will be incorporating cryptographic algorithms to secure nodes and compare them with other techniques in future. To test our proposed model, a more dynamic environment may also be included.

## References

- Amiri, E., Keshavarz, H., Heidari, H., Mohamadi, E., Moradzadeh, H.: Intrusion detection systems in MANET: a review. *Procedia. Soc. Behav. Sci.* **129**, 453–459 (2014)
- Laqtib, S., Yassini, K.E., Houmer, M., Oudghiri, M.D.E., Hasnaoui, M.L.: Impact of mobility models on optimized link state routing protocol in MANET. In: 2016 International Conference on Wireless Networks and Mobile Communications (WINCOM) (2016)
- Venkatasubramanian, S., Suhasini, A., Vennila, C.: An energy efficient clustering algorithm in mobile Adhoc network using ticket Id based clustering manager. *Int. J. Comput. Sci. Netw. Secur.* **21**(7), 341–349 (2021)
- Laqtib, S., Yassini, K.E., Hasnaoui, M.L.: Link-state QoS routing protocol under various mobility models. *Indonesian J. Electr. Eng. Comput. Sci. (IJEECS)* **16**(2), 906–916 (2019)
- Shrestha, R., Han, K.H., Choi, D.Y., Han, S.J.: A novel cross layer intrusion detection system in MANET. In: 2010 24th IEEE International Conference on Advanced Information Networking and Applications, pp. 647–654. IEEE (2010)
- Venkatasubramanian, S., Suhasini, A., Vennila, C.: An efficient route optimization using ticket-ID based routing management system (T-ID BRM). *Wirel. PersCommun* (2021). <https://doi.org/10.1007/s11277-021-08731-6>
- Tsou, P.C., Chang, J.M., Lin, Y.H., Chao, H.C., Chen, J.L.: Developing a BDSR scheme to avoid black hole attack based on proactive and reactive architecture in MANETs. In: 13th International Conference on Advanced Communication Technology, Seoul, pp. 755–760 (2011)
- Arunmozhhi, S.A., Venkataramani, Y.: A flow monitoring scheme to defend reduction-of-quality (RoQ) attacks in mobile ad-hoc networks. *Inf. Secur. J. Global Perspect.* **19**(5), 263–272 (2010)
- Hyojin, K., Ramachandra, B.C., JooSeok, S.: Novel defense mechanism against data flooding attacks in wireless Ad Hoc networks. *IEEE Trans. Consum. Electron.* **56**(2), 579–582
- Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of IEEE International Conference Neural Network, vol. 4, pp. 1942–1948 (1995)
- Indrani, G., Selva Kumar, P.: Handling cross-layer attacks using neighbors monitoring scheme and swarm intelligence in MANET. *Int. J. Comput. Appl. Technol. Res.* **2**(1), 41–48 (2013)
- Vinayagam, J., Balaswamy, C.H., Soundararajan, K.: An efficient optimization based cross layer approach to enhance the security of MANET. *Int. J. Pure Appl. Math.* **118**(24), 1–18 (2018)
- Manikandan, N., et al.: Secured key management with trusted certificate revocation in MANET. *Information Systems Design and Intelligent Applications*, pp. 159–168. Springer, Singapore (2019)
- Gurung, S., Chauhan, S.: A survey of black-hole attack mitigation techniques in MANET: merits, drawbacks, and suitability. *Wirel. Netw.*, pp. 1–31 (2019)
- Das, I., Shaw, R.N., Das, S.: Analysis of energy consumption in dynamic mobile Ad Hoc networks. *Data Communication and Networks*, pp. 235–243. Springer, Singapore (2020)
- Alippi, C., Camplani, R., Roveri, M.: An adaptive LLC-based and hierarchical power-aware routing algorithm. *IEEE Trans. Instrum. Meas.* **58**(9), 3347 (2009)
- Garikipati, V., Naga Malleswara Rao, N.: Secured cluster-based distributed fault diagnosis routing for MANET. *Soft Computing and Signal Processing*, pp. 35–51. Springer, Singapore (2019)
- Agarwal, R., Motwani, M.: Survey of clustering algorithms for MANET. *Int. J. Comput. Sci. Eng.* **1**(2), 98–104 (2009)
- Yao, R., Wang, N., Liu, Z., Chen, P., Sheng, X.: Intrusion detection system in the advanced metering infrastructure: a cross-layer feature-fusion CNN-LSTM-based approach. *Sensors* **21**(2), 626 (2021)
- Yuancheng, L., Rixuan, Q., Sitong, J.: Intrusion detection system using online sequence extreme learning machine (OS-ELM) in advanced metering infrastructure of smart grid. *PLoS ONE* **13**, e0192216 (2018)
- Khan, I.A., Pi, D., Khan, Z.U.: HML-IDS: a hybrid-multilevel anomaly prediction approach for intrusion detection in SCADA systems. *IEEE Access* **7**, 89507–89521 (2019)

22. Mehta, R.: Throughput and resource optimization for adaptive coding-based random access networks with correlated sources. *Int. J. Commun. Syst.* **34**(1), e4673 (2020). <https://doi.org/10.1002/dac.4673>
23. Mehta, R.: Optimal Huffman coding performance of Ad-hoc networks based on cross-layer design. *J. Inf. Sci. Eng.* **36**(6), 1375–1386 (2020). [https://doi.org/10.6688/JISE.202011\\_36\(6\).0015](https://doi.org/10.6688/JISE.202011_36(6).0015)
24. Shakya, S., Pulchowk, L.N.: Intelligent and adaptive multi-objective optimization in WANET using bio inspired algorithms. *J. Soft Comput. Paradigm (JSCP)* **2**(01), 13–23 (2020)
25. Manoharan, S.: Population based meta heuristics algorithm for performance improvement of feed forward neural network. *J. Soft Comput. Paradigm (JSCP)* **2**(01), 36–46 (2020)
26. Jacob, I.J., Darney, P.E.: Artificial bee colony optimization algorithm for enhancing routing in wireless networks. *J. Artif. Intell.* **3**(01), 62–71 (2021)
27. Mugunthan, S.R.: Wireless rechargeable sensor network fault modeling and stability analysis. *J. Soft Comput. Paradigm (JSCP)* **3**(01), 47–54 (2021)
28. Bhande, P., Bakhar, M.D.: Cross layer packet drop attack detection in MANET using swarm intelligence. *Int. J. Inf. Technol.* **13**(2), 523–532 (2021)
29. Song, W., Liu, S., Wang, X., Wu, W.: An improved sparrow search algorithm. In: 2020 IEEE International Conference on Parallel and Distributed Processing with Applications, Big Data and Cloud Computing, Sustainable Computing and Communications, Social Computing and Networking (ISPA/BDCloud/SocialCom/SustainCom), pp. 537–543. IEEE (2020)

# Blockchain-Based Internet of Vehicles for Intelligent Transportation System Using Fog Computing



U. Sakthi , J. Dafni Rose, Dahlia Sam , and M. K. Kirubakaran

**Abstract** Blockchain technology provides transparent, distributed highly confidential, and immutable transactional distributed databases and is responsible for storing valid transaction information committed by each peer in a peer-to-peer system network. An intelligent transportation system (ITS) is an advanced integrated system of smart vehicles and diverse vehicular networks to develop an enhanced environment for a range of transport and traffic management services. A blockchain-assisted decentralized system for the Internet of vehicles (IoV) enables secure and efficient data sharing. We proposed a robust, scalable, fast, and decentralized fog computing architecture to accomplish and control the vehicular network efficiently and effectively in terms of vehicular positioning, information transferring and sharing. The fog computing technology and Internet of things (IoT) are necessary to the existence of transportation systems by creating intelligent and better decisions in a smart ecosystem. The fog computing technology extends the data processing from the cloud to the fog nodes (FNs), significantly improving the quality of service by reducing the network traffic that is sent to the cloud server.

**Keywords** Blockchain · Intelligent transportation system · Fog computing

---

U. Sakthi

Department of Computer Science and Engineering, Saveetha School of Engineering, SIMATS, Chennai, Tamil Nadu 602105, India  
e-mail: [sakthiu.sse@saveetha.com](mailto:sakthiu.sse@saveetha.com)

J. Dafni Rose

Department of Computer Science and Engineering, St. Joseph's Institute of Technology, Chennai, Tamil Nadu 600119, India

D. Sam

Department of Information Technology, Dr. MGR Educational and Research Institute, Chennai, Tamil Nadu 600095, India  
e-mail: [dahliasam@drmgrdu.ac.in](mailto:dahliasam@drmgrdu.ac.in)

M. K. Kirubakaran

Department of Information Technology, St. Joseph's Institute of Technology, Chennai, Tamil Nadu 600119, India  
e-mail: [kiruba23@gmail.com](mailto:kiruba23@gmail.com)

## 1 Introduction

Fog computing reduces network latency and saves network bandwidth by computing some portion of the data internally instead of sending it to the cloud decentralized database server. The main purpose of the fog is to reduce the information traffic that is transferred to the remote cloud data server during data processing by enabling processing, storage, and data management activities on fog nodes. In this research work, we propose a decentralized, privacy-preserving, secured, and efficient framework using blockchain and fog computing. Fog node locally processes the sensing data to provide real-time location-based services to drivers and passengers with minimal delay [1–3]. Internet of vehicles is a distributed wireless network of smart vehicles embedded with Web applications, sensors, and other components for information exchange [4]. In IoV intelligent ecosystem, the vehicles are not permanent, changing place from one location to another location frequently which leads to several security problems. The development of blockchain is a potent solution to the issues on the Internet of vehicles.

Due to the development of computing and communication technologies, smart vehicles are linked with the Internet, implanted computing systems, smart intelligent transport systems, and other vehicles nearby. The Internet of vehicles will play a crucial role in newly developed smart cities to solve road safety and traffic problems. The traditional privacy and security methods are not efficient for sharing information and communication in smart vehicles. To overcome these issues, a decentralized blockchain framework is suggested to enhance the performance of intelligent transport systems by providing security, efficiency, and integrity. The performance of the proposed intelligent transport system can be measured in terms of computing time, immutability, security, and availability of data in-vehicle communication. The rest of this paper is organized as follows. Section 2 includes the survey of fog computing-based smart transportation on the Internet of vehicles using blockchain technology. Section 3 explains the workflow of a blockchain-based intelligent transportation system. Section 4 explains the operational functionality of different layers in the architecture of the proposed system. Section 5 includes the performance analysis results to prove the security and effectiveness of the intelligent transportation system based on the blockchain concepts. In the last section, we conclude some findings and considerations for future work.

## 2 Related Work

With the rapid development of vehicular information systems in IoV, the authentication, security, and privacy preservation of smart vehicles can be accomplished by blockchain technology [5–7]. The quality of the vehicular system is affected by the availability of the IoV ecosystem. The IoV represents a network of smart physical devices like mobile and smart watches integrated with vehicles. Every device in the

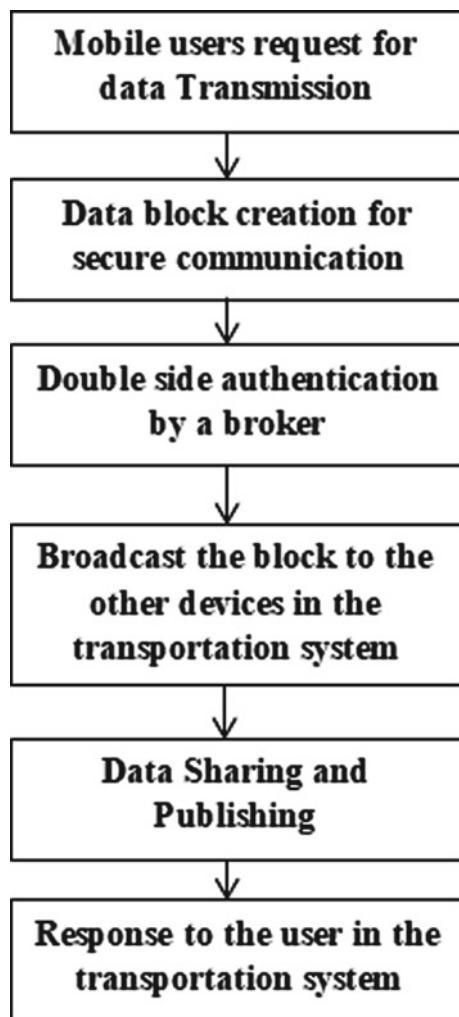
intelligent transportation system has unique identification in the vehicular network. The vehicular ad hoc networks (VANETs) are replaced by the introduction of the Internet of vehicles. The real-time application data can be shared between the smart vehicles and application framework and between the smart vehicles and smart physical devices. In future, nearly 30 billion mobile devices and smart vehicles will be connected to the Internet for real-time services and navigation-based applications. Smart vehicles are connected to the Internet, network applications, and other vehicles nearby in the vehicular network. Smart vehicles are integrated with IoT devices, which leads to the insecurity of data transmitted over the network. It needs additional security privacy policies and authentication of entities in the IoV networks [8–10].

Blockchain technology becomes a basis for many applications like healthcare, transport navigation system, supply chain management, and payment processing [11, 16, 19]. A decentralized global currency cryptosystem was proposed based on the blockchain technology called Bitcoin [1, 2, 12]. The programmable blockchain supports the creation of distributed applications as Ethereum, and the program code is called a small contract, which allows the secure transmission of sensitive data [13]. It ensures secure data exchange between the entities on the Internet of vehicular network. A public ledger called blockchain contains information about all the transactions, available to all the users in the network. All transactions are represented as blocks with time-stamped and cannot be updated.

### 3 Blockchain-Assisted Intelligent Transportation System

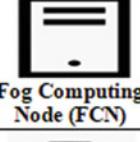
The blockchain concept has been implemented to establish the trust between the smart vehicle and the vehicular fog node using the smart contract of the blockchain. Figure 1 shows the system workflow of the blockchain-assisted intelligent a smart transportation system. The secure data block is added to the message, and the authentication operation is performed by a broker. After publishing the data block in the network and response will be send to the mobile user. The mobile user can request for data transmission in the network, and the data block is created for secure data transmission. The vehicles can perform data communication with the fog node using the smart contract, which ensures security and privacy. The communication between the unauthorized vehicles and the fog nodes is prevented by smart contracts to improve data security. The transaction consists of the data to be transferred between the sender and the recipient and the Ethereum recipient address. The sender sends a signed data package with the signature created using the private key to the recipient in the IoV for enabling secure data communication. The proof of work (PoW) algorithm is deployed in the blockchain network to validate the transactions and add the new block to the existing chain of blocks. It increases security and prevents attacks by adding the current block hash address to the existing hash function.

**Fig. 1** System workflow a smart transportation system



#### 4 The Architecture of the Blockchain Supported Intelligent Transportation System Elated Work

We proposed the IoV cloud fog-based framework that enables the security and privacy of smart vehicles in the intelligent transportation system. Unlike conventional cloud systems, the fog node interacts with the cloud server using a processing node in the roadside units (RSUs) to provide authenticated service to the Internet users in the vehicular network. Figure 2 shows the architecture of a secured IoV service system with fog nodes. The architecture consists of six layers; physical layer, fog layer, network layer, Etherum layer, blockchain service layer, and cloud layer.

<b>Application layer and Cloud Layer</b>				
<b>Blockchain Application Services</b>				
<b>Ethereum Public Blockchain</b>				
<b>Fog Node Layer</b>				
<b>Physical/User Layer (Smart Devices)</b>				

**Fig. 2** Architecture of secured intelligent transportation system

#### 4.1 Physical Layer

It is called a physical layer or device layer that contains android applications, which are responsible for data collection from the different physical devices in the IoV systems. It includes vehicles, roadside units, fog nodes, cloud nodes, sensors, and actuators. The sensors and actuators produce data and vehicles. To provide privacy, each vehicle has local storage to store private data. The RSUs are used as blockchain storage nodes for storing block data. Roadside fog computing nodes, different kinds of vehicles, and other applications are connected to process and maintain blockchain data.

#### 4.2 Fog Layer

This layer has different kinks of edge and fog processing nodes which are in the form of computing nodes and storage devices to perform blockchain-based authentication

by implementing the data privacy services. The edge computing end layer is accountable to perform dynamic data filtering operations to minimize the amount of data transferred to the centralized cloud, and data processing is performed close to the data source. It supports the farmer to precede real-time decision-making depending on the knowledge rule generated by the distributed fog node. The fog processing layer performs the following three functions such as (1) data acquisition comprises amassing data from the vehicular network, (2) stored data pre-processing for data exploration, and (3) knowledge patterns generation to the farmers or researchers and scientists or product developer or other users. The main advantages of Fog computing node are as follows:

**Low Latency:** Some portion of the data is processed in the fog computing node at the end of the network very nearer to the physical devices, which reduces the latency and increases the speed of the process.

**Security:** Integration of blockchain and fog node does not have to allow the transmission of data to the cloud server, which increases privacy and security. The sensitive data are analyzed in the cloud instead of transferring to the cloud, which increases the security level of the data.

**Data Availability:** The smart vehicle user can request the data locally available in the fog node instead send a request to the remote decentralized cloud server. The processing fog node is protected using blockchain technology.

### **4.3 Network Layer**

This layer transfers physical layer device data to the fog layer computing node through various communication technologies like wireless networks, Wi-Fi, Bluetooth, ZigBee, and data transmission hypertext transmission protocol. It connects IoT sensor devices with network devices and transmits data to the fog server for data processing and analysis.

### **4.4 Ethereum Layer**

It provides distributed computing environment for the application running on the Ethereum virtual machine (EVM) based on blockchain. The users can create a new application called a smart contract to run different programs in various applications.

### **4.5 Blockchain Service Layer**

Blockchain has been adopted as the back-end application in the Internet of things for content distribution and broadcasting between the devices through vehicle to

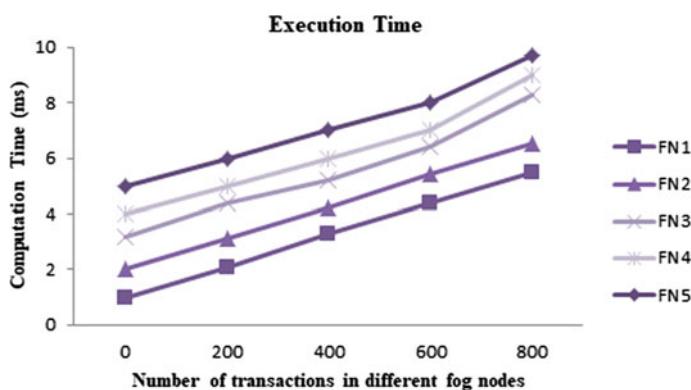
vehicle (V2V) and vehicle to infrastructure (V2I) communications. This layer has the power to increase service availability by avoiding malicious attacks and failure of operations. The event-driven message (EDM) procedure is used with blockchain, fog computing, and 5G for vehicular applications.

#### **4.6 Cloud Layer**

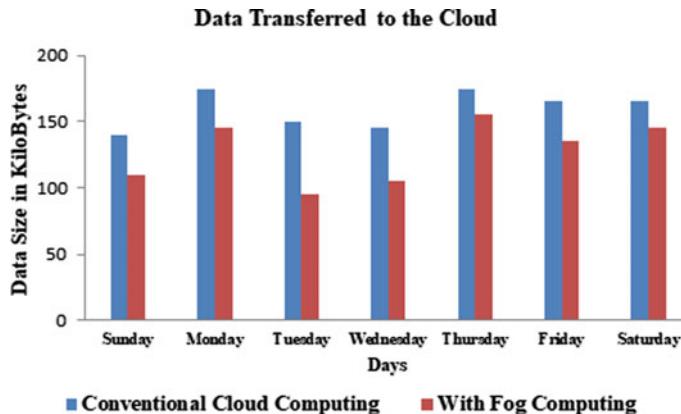
This layer consists of a high computing centralized cloud server and blockchain security services. Blockchain application provides secure data communication between the IoT devices and other users in the intelligent transportation system. The cloud server contains a large database for data analyzing and data processing of IoT applications. Fog computing is applied to execute EDM locally to reduce response time in the vehicular network. 5G communication provides services to the end-user with low latency time and security.

### **5 Performance Evaluation**

In this paper, we have analyzed the performance of decentralized intelligent transportation systems based on various parameters execution time, data availability, and data security. The experiments have been performed on Ethereum private blockchain, fog node, and a cloud server with 64 GB RAM. The proposed framework is analyzed to measure the performance of the proposed smart system depending on the execution time. The execution time is calculated as the time reserved by the proposed system to transfer messages between the smart vehicles to the other user in the network environment based on blockchain technology. Figure 3 shows that the execution time of



**Fig. 3** Execution time for user requests using fog node



**Fig. 4** Amount of data transferred to the cloud environment

the fog node is in the range of 0.8–8 ms when the fog nodes get a maximum of 800 smart vehicle user requests. The execution time has been observed in five fog nodes (FN) numbered from FN1 to FN5 for a fixed number of mobile user application requests. Therefore, it is noticed that the proposed system performance improved by reducing the execution time in several fog nodes. The time complexity of the proposed blockchain-based smart transportation system is the order of  $\log_2 n$ . In this system analysis, ThinkSpeak is used to implement the cloud environment.

Figure 4 shows the system effectiveness based on the amount of data transmitted to the cloud environment with distributed fog computing node. We have compared the performance of conventional cloud computing with the fog-reinforced cloud server with the one-week data size. The amount of data transferred to the cloud server is reduced by introducing the fog node for the local storage and local data processing. Some portion of the data is stored in the local computing node in the fog layer to improve the data availability and reduce the network traffic. The communication latency in the intelligent transportation system is reduced by introducing the fog node in the smart vehicle environment.

## References

- Shi, W., Zhang, X., Wang, Y., Zhang, Q.: Edge computing: state-of-the-art and future directions. *J. Comput. Res. Dev.* **56**(1), 69–89 (2019)
- Jonathan, J., Ryden, M., Oh, K., Chandra, A., Weissman, J.: Nebula: distributed edge cloud for data intensive computing. *IEEE Trans. Parallel Distrib. Syst.* **28**(11), 3229–3242 (2017)
- Noghabi, S.A., Agarwal, C.S., Ananthanarayanan, G.: The emerging landscape of edge computing. *GetMobile Mobile Comput. Commun.* **23**(4), 11–20 (2020)
- Li, Z., Yang, Z., Xie, S.: Computing resource trading for edge cloud-assisted internet of things. *IEEE Trans. Industr. Inf.* **15**(6), 3661–3669 (2019)

5. Leng, J., Ruan, G., Jiang, P., Xu, K., Liu, Q., Zhou, X., Liu, C.: Blockchain-empowered sustainable manufacturing and product lifecycle management in industry 4.0: a survey. *Renew. Sustain. Energy Rev.* **132**, 110112 (2020)
6. Niyato, D., Kim D., Kang J., Xiong, Z.: Incentivizing secure block verification by contract theory in blockchain-enabled vehicular networks. In: ICC 2019–2019 IEEE International Conference on Communications (ICC), pp. 1–7. IEEE (2019)
7. Lu, Z., Liu, W., Wang, Q., Qu, G., Liu, Z.: A privacy-preserving trust model based on blockchain for vanets. *IEEE Access* **6**, 45655–45664 (2018)
8. Xu, C., Liu, H., Li, P., Wang, P.: A remote attestation security model based on privacy-preserving blockchain for v2x. *IEEE Access* **6**, 67809–67818 (2018)
9. Tschorsh, F., Scheuermann, B.: Bitcoin and beyond: a technical survey on decentralized digital currencies. *IEEE Commun. Surv. Tutor.* **18**(3), 2084–2123 (2016)
10. Cao, B., Li, Y., Zhang, L., Zhang, L., Mumtaz, S., Zhou, Z., Peng, M.: When internet of things meets blockchain: challenges in distributed consensus. *IEEE Netw.* **33**(6), 133–139 (2019)
11. Luo, J., Chen, Q., Yu, F.-R., Tang, L.: Blockchain-enabled software defined industrial internet of things with deep reinforcement learning. *IEEE Internet Things J.* **7**(6), 5466–5480 (2020)
12. Xie, J., Tang, H., Huang, T., Yu, F.-R., Xie, R., Liu, J., Liu, Y.: A survey of blockchain technology applied to smart cities: research issues and challenges. *IEEE Commun. Surv. Tutor.* **21**(3), 2794–2830 (2019)
13. Nkenyeraye, L., Adhi Tama, B., Shahzad M.-K., Choi, Y.-H.: Secure and blockchain-based emergency driven message protocol for 5g enabled vehicular edge computing. *Sensors* **20**(1), 154 (2020)

# Smart, Safe, and Secure Shopping Experience Using Beacons



J. K. Lakshmi Divya, R. Iswarya, and V. S. Felix Enigo

**Abstract** We are living in an era where IoT is embraced in every walk of life and retail is no exception to this. With the popularity of the online shopping experience, it is a huge challenge for the retailers to make the in-store shopping experience more effective and interactive. The constantly evolving infrastructure of the IoT devices has made it possible to adapt and benefit from the proximity sensing beacon technology, paving the way to a variety of creative business models. IoT beacons send proximity BLE signals to mobile devices. These small yet powerful devices are all set to take over the retail industry soon to another level of customer engagement, retail marketing, on-demand digital advertising, and profitable experience. With all these advancements, there is always a growing fear for the privacy of customer data and how it is used and analyzed by retailers for understanding customer buying patterns. In this paper, our focus is to use beacon technology to make the shopping experience interactive, safe, and intelligent. This is achieved by personalizing customers' experience using data mining algorithms and giving useful insights to the retailers while preserving the customers' data privacy, thus making it a safe and secure intelligent shopping experience.

**Keywords** Internet of things (IoT) · Smart computing architecture · Proximity sensing beacons · Bluetooth BLE beacons · IoT beacons · Smart shopping · Smart cities · Recommendation · Implicit ratings · k-anonymity · Proximity marketing · Retail · IoT security

---

Supported by Department of Computer Science, SSN College of Engineering.

J. K. Lakshmi Divya (✉) · R. Iswarya  
SSN College of Engineering, Anna University, Kalavakkam, India  
e-mail: [lakshmidivyajk@gmail.com](mailto:lakshmidivyajk@gmail.com)

R. Iswarya  
e-mail: [ishuram@gmail.com](mailto:ishuram@gmail.com)

V. S. Felix Enigo  
Department of Computer Science, SSNCE, Anna University, Kalavakkam, India  
e-mail: [felixvs@ssn.edu.in](mailto:felixvs@ssn.edu.in)

## 1 Introduction

IoT evolution has given us the ability to automate every small thing, allowing us to experience the benefits of technology on a daily basis. IoT helps us communicate with devices (and/or network of devices) and also teaches us to coexist with those devices in the modern world. The BLE beacon technology has found its applications in a variety of scenarios like sports' avenues, airports to help passengers navigate between the gates, museums to give special historical information about the place, public transport, and retail.

Beacon is a Bluetooth radio transmitter. It broadcasts a radio signal that is made up of a combination of letters and numbers transmitted on a regular interval of approximately 1/10th of a second. A Bluetooth mounted device like the smartphone can detect a beacon once the phone is in range. Its model is as shown in Fig. 1.

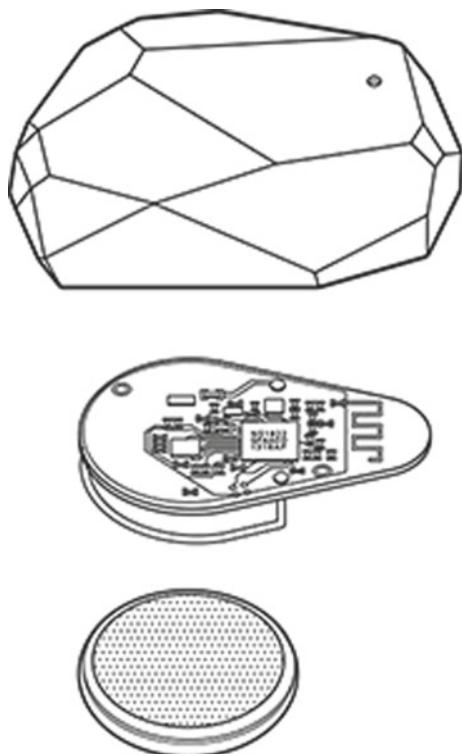
A beacon-based system consists of distributed fixed-location transmitter nodes and portable receivers (in most cases these are smartphones). A sensor-based system uses stationary receivers (connected to an intelligent control point) and portable transmitters (smartphones). Depending on the problem, we can choose beacons or sensors. Since the goal is to give the customers in-store, contact-less, and ambient intelligent shopping experience, we go with the beacons as in [1].

In this paper, we discuss the use of IoT beacons in the shopping environment. We place the beacons at the starting of each aisle in a clothing shop. Whenever a customer comes in proximity of an aisle, the corresponding products from that section are displayed on her mobile application. To personalize the shopping experience, the application recommends items which the customer is more likely to view. On the retailers' end, we use a secure algorithm to mine purchase patterns of customers. We give meaningful insights to the retailer along with preserving the privacy of the customers' data—thus making it a secure personalized experience for customers and also a profitable solution for retailers.

## 2 Related Works

Only two levels of Roberto Pierdicca et al. in [2] focused on helping retailers understand consumer behavior and product improvement. It showed how embedded localization systems can be enhanced. Ambient intelligence is defined and discussed. The work is mainly a preliminary investigation and needs further work. Overall, this paper gives us a great insight on how beacons can be used for retail.

The BeaSmart system in [3] starts by explaining about the emergence of IoT and IBM's Watson IoT. It further elucidates the use of Bluetooth Low Energy (BLE) beacons and discusses its challenges and use cases. A clear understanding of BLE beacons is established by reading this paper. Also, use cases in other industries like health care, mechanical, and pervasive computing [4, 5] show how ubiquitous computing changes the experience of the users of IoT.

**Fig. 1** IoT beacon

Some of the other researches on beacons were carried out by A. Zaslavsky as in [6]. Later came the IBM Watson in [7], which proposed the great vision (cited in [8]) of making the planet smarter by using IoT. The research in this area of IoT was continued by H. Green who proposed the [9]. By then, the IoT beacons became more commercially popular with start-ups like Estimote as in [10], Kontakt.io etc.

All these researches inspired a Bengaluru-based start-up (as reported in [11, 12]) to launch Mobmerry which deploys 1000 beacons on Bengaluru's 100-feet road in Indira Nagar to make shopping an interactive experience.

[13] This paper extensively talks about the BLE beacons, their architecture, android AltBeacon library, AltBeacon protocol, RSSI, and configuration of Gimbal beacon. It further gives a retail or shopping scenario where beacons can be elaborately used. The paper also gives a comparative study among various types of beacons.

Smart shopping using beacons [14] proposes an android application for an intelligent shopping experience by using a machine learning algorithm. It uses proximity sensing in the beacons to recommend products to the customers when they pick up or move a product. There is a PhP application server that sends the required response to the app. It mainly focuses on reducing the number of notifications by intelligently pushing only notifications of the products which the customer will be interested in.

Breadware [15], this web reference from 2019 reports that Forbes had claimed in a survey that by 2021 nearly 90 percent of the retailers would be using beacon technology and it will be a game changer for the retail marketing. It extensively elaborates on how IoT is embraced by the retail industry and mentions about the IoT in retail coming into existence with the Amazon Go store in Seattle. It also mentions the possible threats of data privacy invasion due to these advancements in IoT.

This article by Rajas and Pund [16] tries to track the issue of physical browsing while indoors in retail shops. It tries to draw comparisons between online shopping and physical browsing using IoT BLE beacon technology. While shopping online, the customer will narrow down the search to specific pages of interest. The article discusses this pattern of interest in physical browsing, where a customer could walk to certain aisles and gaze at specific items while in the store.

Jeon et al. [17] were a holistic overview of different types of beacons, use cases, and specifications. This paper talks about the potential of BLE beacons. It talks in detail about many concepts of beacons' infrastructure like localization active sensing, proximity detection, and interaction. It also elaborates on the challenges in the beacon BLE hardware, software, and the system. It also suggests that research on beacon sustainability will help the beacon infrastructure consume less resources. It studies beacons' other limitations like how to make its infrastructure more robust, and how to leverage and harvest its benefits to a greater extent.

This paper by Jeon et al. [18] proposes a sustainable way of managing the beacons. The paper describes that IoT beacons are capable of providing contextual and localization information to its customers. In the recent times, the size of wireless networks has grown so complex that finite batteries have become a challenge. Hence, it suggests an energy harvesting method and proposes using solar panels for powering the beacons, making them sustainable.

[19] Her blog talks extensively about how to find the implicit ratings for a binary stream of data. This is the idea we adapted for our customer's side recommendation algorithm. Manoharan et al. [20] give a good insight on how to integrate predicting/recommender systems with IoT-based systems.

Barkha Kasab and Ubale [21] talk about the privacy protection mechanisms used to preserve the unique identity of an individual by concealing the distinctly identifiable information of a person using some algorithms called k-anonymity and l-diversity. It also provides an in-depth information on different techniques and heuristics for anonymity algorithms.

El Emam [22] Protecting Privacy Using k-Anonymity: This article mainly tells us how k-anonymity algorithm is used in health care, where the data collected is used to zero down to any identifiable person easily. In order to do data mining, they adapted the k-anonymity algorithm to anonymize the large data sets, before doing the analysis.

### 3 Existing Systems

Most of the existing systems use IoT beacons to send or push contextual information. This has been used in a variety of places like airports, museums, bookstores, sports avenues, and retail. There are some systems which do help the customer with her shopping experience. They also use machine learning algorithms to filter the notifications that are sent to the customer when she is at the store—leading to a different experience from online shopping, where all the information on all the products of the store is available. However, these are only one-sided. They only focus on the customer’s experience. They do not provide many benefits to the retailer. Some provide useful insights—but in such systems, there is no mention of customer privacy. It is very much possible that privacy is violated, thereby lessening the willingness of customers to use the system.

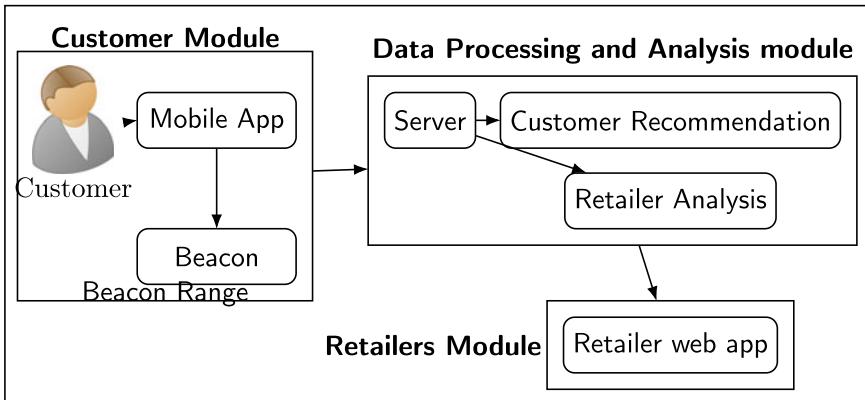
### 4 Proposed System

The proposed system tries to overcome the above limitations and tries to present an effective shopping-cum-retail environment, which is intelligent, safe, and more secure than its predecessors. This has been developed with a view to model a win-win situation for both the customers and the retailers, while preserving the privacy of the customers. The system is intelligent as it gives tailored experience to each and every customer using a data mining algorithm. It also provides insights to the retailers only after first applying a privacy filter.

Our intention in this paper is to show how we could create a powerful, smart, secure, and complete IoT application, by leveraging the benefits of IoT BLE beacons—making the experience intelligent (by adding a data mining algorithm) and secure (by adding a privacy preserving algorithm).

#### 4.1 Architecture

The proposed system as in Fig. 2 consists of a database server. It contains all the details of the customers and products. There are customer-interface modules which interact with the mobile application. Based on the proximity sensing of the BLE beacons and the movement of customers within the shop, recommendations are produced. Simultaneously, based on all these actions, necessary input is given to data analysis processing modules after preserving the privacy of the customers’ data using k-anonymity.



**Fig. 2** Architecture of proposed system

#### 4.2 Assumptions

- In this project, we have assumed that every customer who has visited the shop has a smartphone(Android) which is capable of receiving the BLE signals.
- The BLE beacons do not run out of power and are in a healthy state.
- Each aisle of the store is assumed to hold a unique category of items or products. Say in a clothing shop, when we mention women's aisle, it contains everything related to women and no other category (like men, or kids, or others) is included.

#### 4.3 Modules

**Customer-Interface Module** This module mainly deals with the customers' side of the experience, which can be divided into two sub-modules:

- Beacon Android application module:
  - IoT BLE Beacons: They are placed at the entrance of each aisle of the shop. They keep transmitting the BLE signals at regular time intervals. The range of transmission and the rate of transmission are adjusted according to the requirement.
  - Android Application: The Android application has been developed using the Estimote SDK. All the customers register themselves in the mobile application. The data is stored on the DB server. This application is capable of constantly scanning for the Estimote beacon's telemetry packets. Once it receives the telemetry packet, it extracts the UUID of the beacon and sends it to the recommendation module to personalize the experience of the customers.
  - The communication happens in the JSON format, which is decoded and displayed in a friendly UI format on the Android application using XML.

- Recommendation module:

- It uses machine learning and a data mining algorithm to send the recommendations. A user-based collaborative filtering with no direct rating has been used. We do not ask the customer to rate a product, as it is expensive and it requires a conscious non-negative effort from the customer to rate the product. Here, we try to get implicit ratings from the way the customer views the product. We assign every item viewed by the customer an itemStrength (using  $S(x)$  as the itemStrength of item  $x$ ) as in Algorithm 1 in the algorithm section. Later, with the implicit ratings (calculated from itemStrength), the user-based collaborative filtering is applied where the similarity between the customers is found using cosine similarity. Using the cosine similarity, we intend to predict the likelihood of an item that will be viewed by the customer (e.g., say customer X has viewed A, B, and C; customer Y has viewed A and B; it is more likely that customer Y would like to view product C). It is calculated as in Formula 1

$$\cos \theta = X \cdot Y / \|X\| \|Y\| \quad (1)$$

$$\text{Sim}(A, B) = \frac{\sum_{i=1}^n A_i \cdot B_i}{\sqrt{\sum_{i=1}^n A_i^2} \cdot \sqrt{\sum_{i=1}^n B_i^2}}$$

**Data Processing and Analysis Module** This module mainly deals with the retailers' side, which can be divided into two major sub-modules:

- Privacy Preservation Mechanism (PPM) module:

- Here, we do the important step of preserving the identity of the individual before analyzing the data collected. So, we anonymize the data using k-anonymity or l-diversity algorithm. With k-anonymity, the original data set containing the personally identifiable information of the application users is transformed, making it almost impossible for any unauthorized person/system to recognize the individuals in the data set. In k-anonymity algorithm, we anonymize the data such that at least  $k-1$  records are similar to each other. There are two methods to do this:

1. Suppression: We make an attribute value as “\*” or “\_” for all records, making the particular attribute not very usable.
  2. Generalization: In this method, we make the records anonymous by replacing the values using a category/range.
- The data we intend to anonymize depends on the data we want to protect: Here, we do not want the retailer to get the complete phone number of the person; we just want her to get useful insights from the beacons like (i) when is the shop crowded, (ii) which area of the shop is most frequented, and (iii) how many users are online.

- We used both the methods of k-anonymity to anonymize the user data. We then fed it to the analysis module to allow the retailer to see useful insights in a secure way.
- Analysis module:
  - After the data set is made anonymous, it is fed into the analyzing systems which find the customers' purchase patterns and help the retailer modify the store layout or product positions.
  - There is a Web app which would help the retailer see insights like the number of customers in the store, times when the store is crowded, etc.

#### **4.4 Constraints**

- The maximum range of transmission and the battery life depend on the type of the IoT beacons used.
- When objects come in the way of the IoT beacons, they cause blockages for the signals while transmitting.
- Frequent monitoring of the beacons' battery life is needed. It is cost-effective when large number of beacons are deployed compared to one or two.
- A good Internet connectivity is necessary.

#### **4.5 Algorithm**

The algorithms for itemStrength (Algo. 1), Recommendation (Algo. 2), anonymizeData (Algo. 3) are as follows:

---

##### **Algorithm 1** Procedure calculateItemStrength(T, Phone P, Item X)

---

**Result:** calculateItemStrength

**Input:** T is the database with tables seenItems, and Items

```
User with Phone P relates to an Item x via a mobileAppScreenEvent E switch(E){  
case "VIEW_ITEM": S(x) = 1  
case "LIKE_ITEM": S(x) = 2  
case "FOLLOW_ITEM": S(x) = 3  
case "ADD_TO_CART": S(x) = 4  
case "BUY_NOW": S(x) = 5  
}
```

---

---

**Algorithm 2** Procedure reco(T, Phone, item x)**Result:** Top 5 recommended items**Input:** T is the database with tables seenItems, and Items

Phone is the target phone number.

1. Predict the itemStrength S(T,P,x), for every item x as in Algorithm. 1
  2. Calculate the Cosine similarity as in 1
  3. OrderBy(similarity)
  4. Find the customer(phone P1) with highest similarity
  5. Select all items from P1 and  $\neg(P$  item history) //select items from P1-P
  6. Return itemIDs
- 

---

**Algorithm 3** Procedure anonymiseDataForRetailers(D)**Result:** AnonymizeUserDataForRetailers**Input:** D is the database with Tables Users, seenItems, Items

Phone is primary key in all the tables in Database D.

1. Make a Table retailersDisplayUsers
  2. `retailersDisplayUsers.Age = mapAge(Users.Age)` //k-anonymity generalization
  3. `retailersDisplayUsers.phone = suppressPhone(Users.Phone)` //k-anonymity suppression
  4. `retailersDisplayUsers.ItemsBought = Users.ItemsBought`
  5. `retailersDisplayUsers.BeaconsVisitedTime = Users.BeaconsVisitedTime`
- 

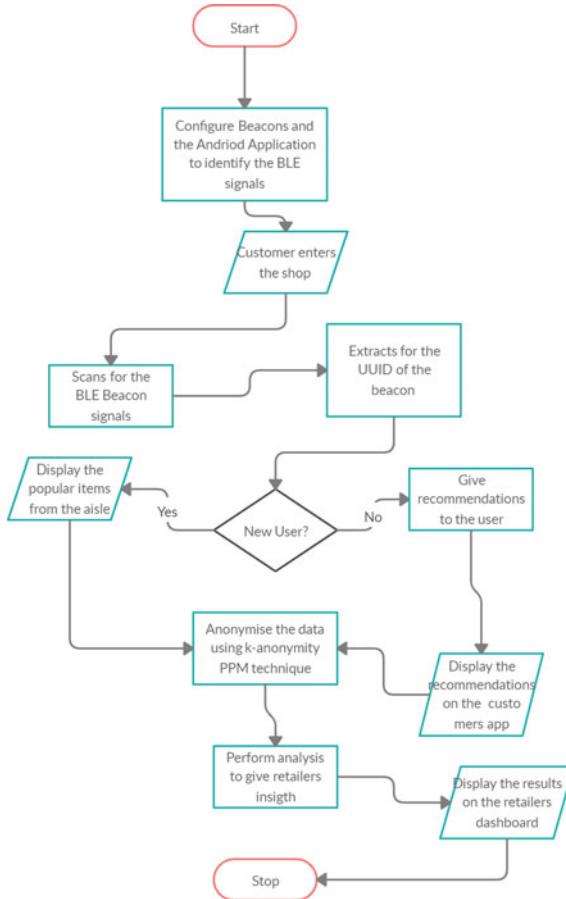
## **4.6 Overall System in a Flowchart**

The overall flow of the system is shown in Fig. 3.

## **4.7 Implementation**

The system has been implemented with all the above modules. The customers have to register themselves in the store's application (which has the beacon listening capability). The customer should be logged in on their mobile whenever they enter the shop. They would receive notifications whenever they enter a particular aisle. The notifications show recommendations of the products in that aisle, as depicted in the Architecture 2. Also, the retailer will be getting insights from the data set that is collected from the customer beacon interaction, periodically. The privacy preserving mechanism is applied to give secure insights to the retailer, hence making it a smart, secure, and safe retail environment.

Based on Algorithm 2, for the input in Table 1, the cosine similarity for the user with the phone number  $P1$  is as in Table 2. Algorithm 2 is also used to calculate the cosine similarity for users with phone number  $P2$  and phone number  $P3$ . After applying Formula 1, we found that  $\text{Sim}(P1, P2)$  is 0.85 and  $\text{Sim}(P1, P3)$  is 0.39. Hence,  $P1$  is more likely to view products very similar to customer  $P2$ , which is also evident from the itemStrengths or implicit ratings.



**Fig. 3** Flowchart depicting the overall system flow

**Table 1** Reco algo: sample user-item table

Phone	ItemBought	ItemStrength
P1	I1	4
P1	I2	2
P1	I3	1
P2	I1	5
P2	I2	5
P2	I4	3
P3	I1	1
P3	I5	2

**Table 2** Reco algo: cosine similarity matrix for target phone  $P_1$ 

Phone	I1	I2	I3	I4	I5	CosineSimilarity
$P_1$	4	2	1	0	0	–
$P_2$	5	5	0	3	0	0.852
$P_3$	1	0	0	0	2	0.39

**Table 3** k-anonymity algo: sample user details table

Phone	Name	Age	ItemsBought
1111111110	Harry Potter	25	$I_1, I_2, I_3$
2222222221	Hermione Granger	53	$I_1, I_2, I_4$
3333333332	Ron Weasley	15	$I_4, I_5$

**Table 4** k-anonymity algo: anonymized data for retailers analysis

Phone	Age	ItemsBought
xxxxxx1110	20–30	$I_1, I_2, I_3$
xxxxxx2221	50–60	$I_1, I_2, I_4$
xxxxxx3332	10–15	$I_4, I_5$

Table 3 is obtained after applying Algorithm 3 by suppressing the phone number, dropping the name of the person and mapping the age using a generalization technique. This data is then given to retailers who can use it to analyze and find out things like what age group prefers which kind of items. Hence, the user's personally identifiable information is protected for privacy (Table 4).

## 5 Results

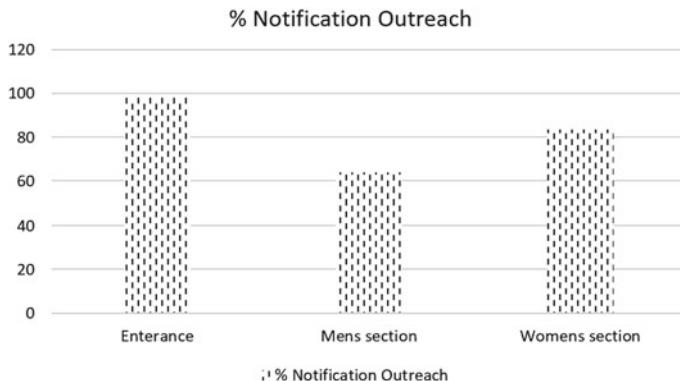
The intention of this system is to create an intelligent, contact-less, and secure shopping experience for the customers. On the retailers' side, we would like to see how well the beacons were able to reach the customers:

1. Number of notifications received by the customers
2. Number of notifications viewed (click-through-rate—CTR)
3. Number of items viewed from inaccessible areas of the store.

We conducted an experiment by deploying this application in a local clothing store. Two hundred people volunteered in this experiment, installed the customer side .apk file, and visited the store. We recorded the following data for a period spanning a couple of weeks.

**Table 5** Number of notifications sent

Location of beacons	Average #Notifications sent	% Notification outreach
Entrance	196	98
Men section	128	64
Women section	167	83.5

**Fig. 4** Beacon notification outreach**Table 6** Click-through-rate

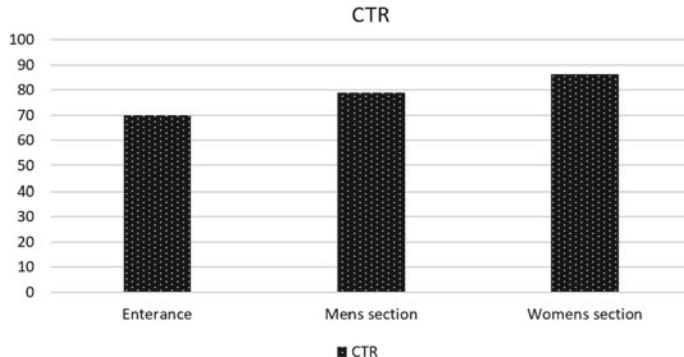
Location of beacons	Average #notifications viewed	Click-through-rate CTR (%)
Entrance	137	69.9
Men section	101	78.9
Women section	144	86.2

### 5.1 Notifications Sent

The observations are as in Table 5 and Fig. 4. We can see that the notifications hit almost every customer who had the .apk file. We can also see how many people walked through different sections. The average outreach in every location of the beacon definitely proved how much proximity marketing could be leveraged.

### 5.2 Click-Through-Rate

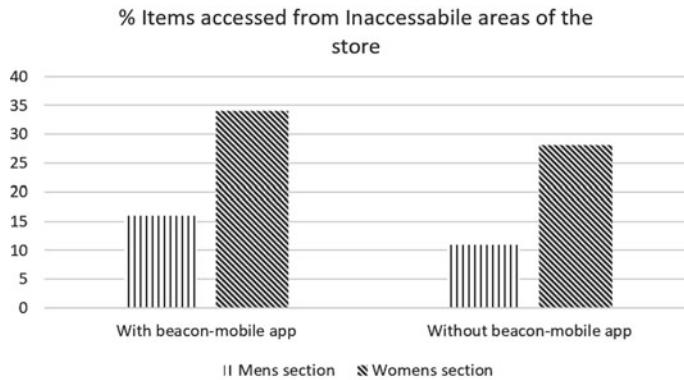
The observation is as in Table 6 and Fig. 5. This metric shows the likeliness of a customer to want to view similar items by clicking through the notifications and wanting to use the recommendations provided by our recommendation algorithm.



**Fig. 5** Beacon notification click-through-ratio CTR

**Table 7** Cost

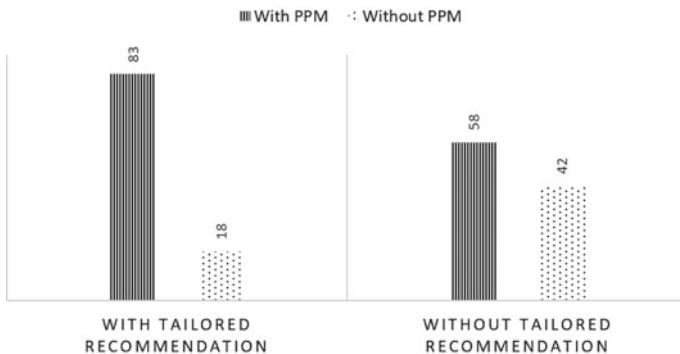
Location of the beacons	With beacon mobile app	Without beacon mobile app
Men's section	16	11
Women's section	34	28



**Fig. 6** Increase of items viewed from inaccessible corners of the store

### 5.3 Percentage Items Accessed from the Inaccessible Areas of the Store

The observations are recorded in Table 7 and Fig. 6. To understand the efficiency of the system, data was collected from the local shop owner, before installing beacons. The results also showed that the sales of items in inaccessible corners of the store increased.



**Fig. 7** Testing survey analysis

#### 5.4 PPM Survey

To understand the efficacy of the application, we conducted a survey on how much the customers were comfortable using an application incorporating privacy preservation mechanisms. We also surveyed if they like the tailored recommendations. Both these results are depicted in the graph in Fig. 7 which concludes that tailored recommendations and privacy preserving mechanisms were preferred by around 83% of the customers.

With the above results, we found that this application of using beacons to provide a contact-less, intelligent, and secure shopping experience was definitely a success, as it gave profits to retailers, and a safe and secure experience to customers. Thus, we gave a win-win solution for both customers and retailers.

## 6 Conclusion

We listed the features of IoT beacons. We also briefly discussed existing works on similar intelligent retail environments, and their drawbacks of being less secure and of being only customer-sided. We explored new ways of overcoming the drawbacks found in these previous smart systems, to provide a more secure, intelligent, smart, and interactive shopping experience.

Looking at all the above algorithms using IoT proximity beacon technology, we find that preserving the privacy of customers' data along with tailoring their experience will pave way to powerful, intelligent, and secure retail systems.

## 7 Future Work

There are so many dimensions we could see this technology progress toward. One major dimension is sustainability: As the batteries in the beacons are not reusable, if they are replaced by sustainable components (like harvesting energy from the renewable resources), then the IoT BLE beacon technology will also be green, sustainable, and cost-effective. With such changes, the IoT beacons will definitely change the retail industry in an eco-friendly way.

Another dimension involves new types of privacy preserving mechanisms and data mining algorithms.

There are several fields other than retail, where this technology could be potentially adapted.

## References

1. <https://electronics.stackexchange.com/users/41777/undertherainbow>, U. Beacons versus sensor nodes. Electrical Engineering Stack Exchange. <https://electronics.stackexchange.com/q/258104> (version: 2017-04-13)
2. Pierdicca, R., Liciotti, D., Contigiani, M., Frontoni, E., Mancini, A., Zingaretti, P.: Low cost embedded system for increasing retail environment intelligence. In: 2015 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), pp. 1–6. IEEE (2015)
3. Akinsiku, A., Jadav, D.: Beasmart: A beacon enabled smarter workplace. In: NOMS 2016-2016 IEEE/IFIP Network Operations and Management Symposium, pp. 1269–1272. IEEE (2016)
4. Chen, J.I.Z., Yeh, L.-T.: Analysis of the impact of mechanical deformation on strawberries harvested from the farm. J. ISMAC 3, 166–172 (2020)
5. Shakya, S., Nepal, L.: Computational enhancements of wearable healthcare devices on pervasive computing system. J. Ubiquitous Comput. Commun. Technol. (UCCT) 2(02), 98–108 (2020)
6. Zaslavsky, A.: Internet of things and ubiquitous sensing. Accessed Aug. 2019 (2013). <http://www.computer.org/web/computingnow/archive/september2013/>
7. Team, I.W.: Ibm-bluemix. Accessed Aug. 2019 (2014). <https://console.ng.bluemix.net/docs/services/IoT/index.html#gettingstarted/>
8. Team, I.W.: Ibm-smarterplanet. Accessed Aug. 2019 (2014). <http://www.ibm.com/smarterplanet/us/en/ibmwatson/>
9. Green, H.: Cognitive Iot: making the internet of things deliver for all of us. Accessed Aug. 2019 (2015). <http://www.ibm.com/blogs/think/2015/12/15/cognitive-iot-making-the-internet-of-things-deliver-for-all-of-us/>
10. Santos, A.: How beacons deliver content to mobile devices? Accessed Aug. 2019 (2016) <https://community.estimote.com/hc/en-us/articles/200848086-Delivering-Content-to-a-Mobile-Device>
11. Online, D.: Bangalore startup deploys beacons to create iot enabled shopping district. Accessed Aug. 2019 (2015). <http://www.dqindia.com/bangalore-startup/>
12. Team, N.O.: Mobmerry to deploy 1000 beacons in bangalore stores to connect retailers with potential customers. Accessed Aug. 2016 (2015). <https://news.nextbigwhat.com/mobmerry-to-deploy-1000-beacons-in-bangalore-stores/>
13. Shyam, J.S.: A simple smart shopping application using android based bluetooth beacons (Iot). Adv. Wirel. Mobile Commun. 10, 885–890 (2017)
14. Maitri, P., Megha, S., D.V.-P.R.K.: Smart shopping using beacons. Accessed May 2020 (2018). <http://troindia.in/journal/jcesr/vol5iss4part12/1-6.pdf>

15. Breadware.: How retail is using IoT solutions. Accessed May 2020 (2018). <https://breadware.com/blog/iot-in-retail/>
16. Rajas, M., Pund, M.A.: The internet of things—improving customer shopping experience at stores. *IRA-Int. J. Technol. Eng.* **7**(2(S)), 248–256. ISSN 2455-4480 (2017)
17. Jeon, K.E., She, J., Soonsawad, P., Ng, P.C.: Ble beacons for internet of things applications: survey, challenges, and opportunities. *IEEE Internet Things J.* **5**(2), 811–828 (2018)
18. Kang Eun, J., Tong, T., She, J.: Preliminary design for sustainable ble beacons powered by solar panels. In: 2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 103–109 (2016)
19. <https://towardsdatascience.com/building-a-collaborative-filtering-recommender-system-with-clickstream-datadfffc86c8c65>, S. L. Building a collaborative filtering recommender system with clickstream data. towardsdatascience.com
20. Manoharan, S., et al.: Early diagnosis of lung cancer with probability of malignancy calculation and automatic segmentation of lung ct scan images. *J. Innov. Image Process. (JIIP)* **2**(04), 175–186 (2020)
21. Barkha Kasab1, V.P., Ubale, S.: Enabling privacy preservation technique to protect sensitive data with access control mechanism using anonymity. Accessed May 2020 (2015). [https://www.ijcseonline.org/pub\\_paper/13-IJCSE-01356.pdf/](https://www.ijcseonline.org/pub_paper/13-IJCSE-01356.pdf/)
22. El Emam, K., Dankar, F.: Protecting privacy using k-anonymity. Accessed May 2020 (2008)

# Android Game for Amblyopia Treatment: A Prospective Study



Sarah AlGhamdi, Sadiqa Alghawas, and Nazeeruddin Mohammad

**Abstract** Amblyopia (also known as lazy eye) is a kind of poor vision that affects only one eye. It happens when the brain and the eye do not cooperate together properly. In such case, the brain cannot identify the sight from the impacted eye, and it becomes increasingly reliant on the stronger eye over time. Consequently, the vision of the weaker eye deteriorates. It is a complex symptomatic of sensory and motor functional disorder. Its main manifestation includes disorders of central and peripheral vision, light and color perception, contrast, electrical sensitivity and liability, and accommodative ability. According to statistics, around 2–3% of the entire world population suffers from amblyopia, which is equivalent to 10 million people under the age of eight. In Saudi Arabia, the prevalence of amblyopia in most of the clinical studies is 9.5%. There are a wide variety of treatment modalities, the most common is to wear a patch over the strongest eye for weeks or months, forcing the lazier eye to do more work. Current treatments such as wearing an eye patch or using atropine eye drops may not be acceptable. Implementation of such treatments is challenging for parents as their children do not feel comfortable wearing the patch. Therefore, in amblyopia treatment, there is a need to actively influence the nature of eye movements, consistent with the sensory system. This project provides a design of a visual Android game, which aims to help in the treatment of amblyopia or “lazy eye.” It demands the transmission of information to both eyes in order to operate cooperatively. Short-term results show improvements in some patients who followed this treatment method consistently

**Keywords** Amblyopia · Android · Game · Treatment

---

S. AlGhamdi (✉) · N. Mohammad

Department of Computer Science, Prince Mohammad Bin Fahd University,  
Dammam, Saudi Arabia

e-mail: [Saraalghamdi497@gmail.com](mailto:Saraalghamdi497@gmail.com)

N. Mohammad

e-mail: [nmohammad@pmu.edu.sa](mailto:nmohammad@pmu.edu.sa)

S. Alghawas

Department of Ophthalmology, Dammam Medical Complex, Dammam, Saudi Arabia

## 1 Introduction

In medical practice, smartphones are becoming a more common and fast evolving instrument. A plethora of applications are now available in the market to assist clinicians in performing various activities related to medical care (See [1] and references therein). Such applications become essential in our life due to COVID-19 [2]. Specifically, for ophthalmology-related care, activities such as measurement of visual acuity, glaucoma, and amblyopia can now be achieved using smartphone applications. Amblyopia has been a major problem to many patients [3, 4]. Indeed, around 2–3% of the entire world population is affected by it, which is equivalent to ten million children under the age of eight. Hence, introducing and addressing new ideas to lower its effect are considered very important. Despite the progress achieved in the treatment of amblyopia, the problem of improving visual acuity remains very relevant due to the high rate of amblyopia. Traditional methods of restoring visual functions in amblyopia mainly aim at stimulating the sensory component, which does not exclude an indirect effect on the motor component. Consequently, in the treatment of amblyopia, there is a need to actively influence the nature of eye movements, consistent with the sensory system. In [5], the effectiveness of playing computer game is assessed for children with amblyopia. The logMAR visual acuity (VA) has been evaluated before and after the treatments. Based on that, an improvement in the mean VA has been observed after seven weeks. However, this improvement was not statistically significant for which the sample size needs to be increased. In [6], different treatment protocols were compared to determine the adequate treatment time. These protocols result in almost similar improvement in the visual acuity. A mobile game called Space Vision has been developed in [7] for visual acuity test and home-based monitoring. According to this study, it is found that the behavioral and aesthetic customizations are crucial to enhance the resonance between the patient and the game..

In most clinical investigations in Saudi Arabia, the prevalence of amblyopia is 9.5%, whereas the prevalence in population studies is believed to be 1.6–3.6%. Furthermore, 68.5% of amblyopic children in the study are unilateral, whereas 31.5% are bilateral [8]. This is considered very critical and demands serious attention, for the purpose of ensuring the safety and well-being of children. However, if a child was not treated from early age, he or she will develop with one dominant eye, and the other one will be considered a lazy eye. Professionals treat lazy eyes with an eye patch, in order to cover the stronger eye, for the purpose of enforcing the lazy eye to function (Table 1). The treatment usually takes several months, in order to let a patient to recover from the sickness. Such form of treatment has been annoying for many children and has affected their self-esteem, as they look different among their friends with an eye patch. It is worth mentioning that the patching treatment will eventually fail if wearing of the patch is not continued [9]. This also limits a child's activities at school, having a negative impact on his or her well-being. As a result, orthoptists and ophthalmologists are always on the lookout for a more suitable solution to the problem, i.e., a viable treatment that is also well-accepted and hence truly

**Table 1** Amblyopia treatments

Traditional treatment	Developed treatment
Eye patch, eye drops, clinic therapy's	Games, designed to treat lazy eye
Both of these treatments take larger period of time, consume a lot of money, and affect patient self-esteem	This treatment consumes less cost and time, develops motivational environment among patients, encourages patients to score more, does not affect patient self-esteem

effective [10]. One of the developed treatments is known as Wow Vision Therapy, in which professionals assist their patients to overcome their obstacle through intensive office-based vision therapy [4]. Another achievement has been made by Amblyotech Company, which has managed to develop games, designed to assist lazy eye patients, and these games have been approved by professional doctors [11]. In this study, we design and develop an Android game [12], using Android Studio IDE, for amblyopia treatment. Specifically, the game is developed to help the children who suffer from amblyopia without visiting the clinic for special treatment, or wearing the eye patch. Also, we have created a Web site to explain the goal of the game and the way of playing it [13]. Likewise, the Web site includes a frequently asked questions section, through which parents can contact us and receive our consultation and recommendations. In addition, the Web site is written in Arabic, in order to provide assistance to the numerous Arabic-speaking parents, who have a child with amblyopia, and to encourage the children to play the game [14]. We want to make the Web site in Arabic, in order to make it easier for them to understand the idea of the game. Furthermore, a patient can view their score in the game, and depending on the score, he or she can move to the next level. A patient is not able to increase his or her level on the first few days, since it requires both eyes to work together. Since the patients have one lazy eye, they can move to the next level after much practicing in the game. Hence, it is possible to make a weekly report to compare the results of each patient, in order to estimate the improvement in each of the age groups.

### 1.1 Playing Therapy

Playing therapy is a strategy of meeting and adapting to children's health needs, and it is widely recognized by specialists as a successful and appropriate intervention in dealing with the development of children's brain [15]. The developed form of this game and the Web site will allow involved people such as doctors, children's parent to connect with patients more quickly and dynamically, in which they can follow up on their health status. This game is designed for a smartphone, tablet, with Android OS, and the Web site allows doctors and parents to access information when and where they need it, such as the children score before and after playing, comparing their result with other children in the same age. This process can be helpful in which

the doctors will be up to date with the patient health status. On the other hand, the developer can improve his game according to the children statistical results and the doctor's instructions. The game provides familiar objects to the children such as shapes, squares and music. It also provides an easy instruction to the children about how to play such as arrows (left—right—up) and the score he gets on each level. This score is the key that doctors can determine the improvement process on the children medical status.

## ***1.2 Study Objectives***

The main objective of this game is to help in treatment the children with amblyopia. Besides, playing allows children to express their thoughts and feelings using their most comfortable method (playing). So, playing therapy will give children a chance to process their therapy in enjoyable way. Game will allow the doctor to adjust the required levels of exposure to each eye of a patient, leading the work of the left and right eyes to a synchronous position.

## **2 Literature Review**

The development of effective methods for the treatment of this pathology is one of the most important tasks of pediatric ophthalmology. Amblyopia is the second most frequent (up to 6%) illness after myopia, which is the reason of the decrease of children's visual acuity in preschool and school age. In connection with the late appointment of an appropriate correction of refractive errors, the development of refractive amblyopia is observed in 33–98.4% of children. The prevalence of refractive amblyopia with hyperopia reaches 70%. By its nature, amblyopia is a functional pathology of the higher parts of the central nervous system, and its pathophysiological basis is persistent cortical inhibition of the function of central vision, which develops as the result of sensory deprivation in early childhood. In fact, the development of amblyopia is associated with the disruption of interneuron interactions at various levels of the visual system—from the sensory retina to the external cranial bodies and central regions in the occipital lobe of the cerebral cortex. Amblyopia is a complex symptomatic complex of sensory and motor functional disorders. Its main manifestation was previously considered a decrease in visual acuity. However, as the pathogenesis and clinical picture of this disease was studied, a number of other disorders of central and peripheral vision, light and color perception, contrast, electrical sensitivity and lability, along with accommodative ability, were detected. According to modern concepts, the main goal of treating refractive amblyopia is to achieve the maximum and consistently high visual acuity (0.4 and higher). The first simple and traditional methods of treating amblyopia are penalization and direct occlusion. The principle of these methods is to turn OFF the better Seeing Eye from

the act of vision. Penalization is most effective in 2–3-year-old children (97–98%) in the period, when there are no serious sensory impairments in the visual system. Yet, its effectiveness is significantly reduced (14.3–31.2%) at the older age. The disadvantages of this method are the duration of treatment (from 1 to 2.5 years) and the need for long-term mydriasis. Recently, various computer-based stimulation methods have been useful in the treatment of amblyopia. Computer programs increase the efficiency of the defective part of the visual analyzer due to a patient's sensible solving of visual problems. They contribute to the activation of brain neurons and the restoration of interneuron connections at all the levels of the visual system. This method has several advantages. Due to the capabilities of computer programs, there is a gradual complication of stimuli that are adequate for various channels and levels of the visual analyzer. The computer graphics arsenal provides tremendous opportunities for creating a variety of treatment programs, in which both automatic process control and accurate recording of the results of each session are provided. In all computer programs, medical procedures are provided in a game form with a patient's active participation, which greatly increases his or her interest and thereby shortens the treatment time. The ability to widely vary and dose the effects, along with changing the settings and the size of the stimulus, allows a researcher to individually select the treatment.

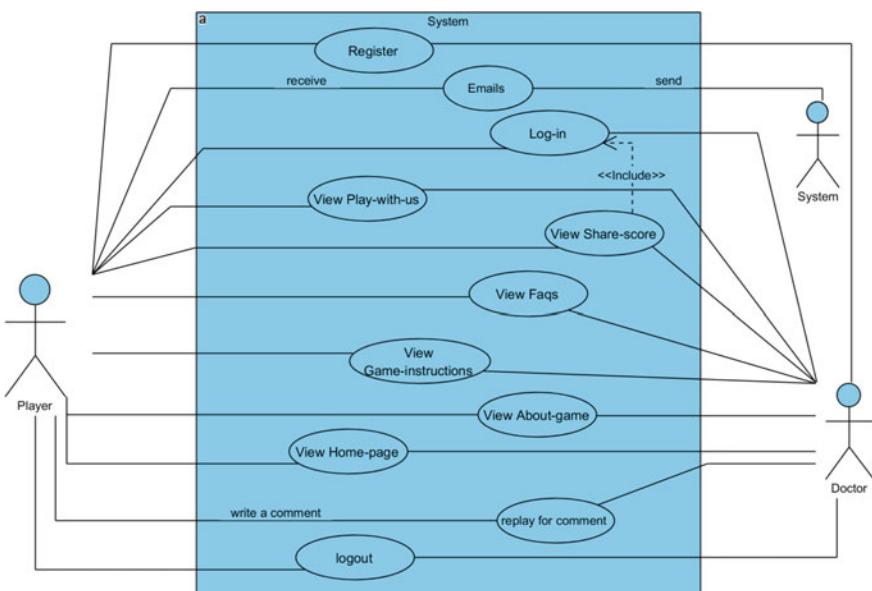
### 3 Methodology

The methodology of this project based on the game Dig Rush that Ubisoft was developed jointly with Amblyotech. The gameplay is designed to help in treating amblyopia or “lazy eye,” a disease, where one eye cannot work synchronously with the other one and the brain. The doctor can adjust the required levels of exposure to each eye of a patient, leading the work of the left and right eyes to a synchronous position. Additionally, different load allows training the mobility and speed of the problem eye on nerve impulses. The game approach to correcting this problem is appropriate due to the ineffectiveness of the existing treatment methods. The classic approach is to increase the load on the affected eye, which usually requires putting a bandage on the healthy one. However, the main disadvantage of the method is that in order to achieve the desirable result, one has to do the dressing for several hours every day. The bandage imposes certain restrictions, as a child's development is slowed down, since he or she has to know the world and learn to evaluate the perspective and distance, which is difficult to do with one open eye. This will simply lead to the need to re-go through the cognition process later. In addition, the bandage creates discomfort and makes a child feel sick. Moreover, if one goes to school, most likely, he or she will not cover his or her eye. The game, in turn, stimulates a child to go through the whole process of treatment very quickly. In the process of developing the

game, the studies of ophthalmologists from McGill University have been considered. The game is designed for tablets and requires the use of 3D glasses. Moreover, it implements an adjustable level of contrast of red and blue, which are refracted by 3D glasses.

### 3.1 Study Design

The game involves wearing 3D glasses with red-green glasses and directing the red square to hit the blue one to collect score, so it depends on two colors (red and blue). In order to see different colors, the patients are required to wear 3D glasses. The information must be transmitted to both eyes enabling them to cooperatively work together. As a result of the increased plasticity in the brain, the amblyopic brain will be able to relearn. The game is now available on Google Play, and patients can download it on their phone, write comments, and view other parents' comments. The use-case diagram of the model is illustrated in Fig. 1. In order to record users' scores in the game, we contacted with players weekly for the purpose of comparing patients' scores. If the score is too low, comparing to the other patients of same age, the patient should stop playing the game because that means low or no response at all. Otherwise, when the received score is high, the patient can continue playing the game.



**Fig. 1** Use case diagram

### ***3.2 Sample Size***

The sample was seven amblyopic patients from Dammam Medical Complex.

### ***3.3 Data Collection***

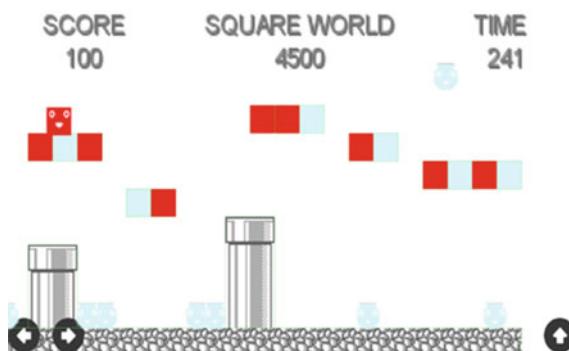
We visited the hospital and met patients and their parents, to explain the idea of the game and how they should play the game. Also, we gave the patients 3D glasses with red-green lenses to wear while they are playing the game. The patients were divided into groups depending on their age, and the measurements were recorded. We made sure that patients' identities remain secret. We met the patients weekly to get feedback. Then, and after one month, we met them again to do eye measurement tests and recorded the results.

### ***3.4 Data Management***

Patients played a game and received score for each completed level proceeding to the next level. The Web site collected scores and provided a report to compare the results of each patient every week, in order to estimate the improvement in each of the groups.

## **4 Results**

In this section, the obtained results for amplyopia treatment using Dig Rush game (shown in Fig. 2) are presented. Seven children with amblyopia from Dammam Hospital have participated for testing the proposed treatment method. The average age of these participants is seven years old. They are divided into two groups. One group is supposed to use the glasses for one month, five days a week, and an hour a day. Meanwhile, the other group received a closed bandage for the “lazy eye.” Among four children of the first group, the conditions of two participants have been improved. One of them did not achieve any improvement as he/she was not committing to use the glasses. However, the participants in the second group who wear eye patch on their “lazy eye” do not show any improvements. The child’s parent gets benefit from the Web site which explains the idea of the glasses and benefit of the glasses over other treatments. Table 2 shows the children information before using the glasses and their results after the proposed treatment. It is evident from the obtained results the advantages of the proposed solution (game and Web site) for children with “lazy eye.”



**Fig. 2** Dig rush game used in ambylopia treatment

**Table 2** Summarization of children treatment score

#	Child	Visual acuity	Visual acuity after using the glasses
1.	Patient 1	<b>Rt:</b> -7.50 -3.5 5 <b>Lt:</b> +3.00 -1.75 175	Improvement 0.3
2.	Patient 2	<b>Rt:</b> +8.50 -1.25 170 <b>Lt:</b> +8.00 -1.25 175	Improvement 0.4

**Fig. 3** Mohammed, a 4-years-old ambyopic patient is using Dig Rush game as a treatment



With the binocular treatment, the average amblyopic eye visual acuity (VA) letter score is enhanced from baseline by 1.3 (2-sided 95% CI: 0.1–2.6; logMAR 0.026). However, with continued spectacle correction solely, it is enhanced by 1.7 (2-sided 95% CI: 0.4–3.0; logMAR 0.034). After the adjustment of baseline visual acuity (VA), the difference of letter score across groups (binocular minus control) is obtained as -0.3 (95% CI: -2.2–1.5,  $P = 0.71$ , difference logMAR: -0.006). According to the data acquired from the iPad, it is observed that more than half of the partici-

pants in the binocular group have completed more than 75% of the recommended treatment. These include 58% and 56% of the participants in four and eight weeks visits, respectively. It is worth mentioning that significant improvements have been observed especially for one case of the seven patients (shown in Fig. 3). Unfortunately, other participants withdraw as their parents were not convinced in this type of treatment.

## 5 Conclusion

In this study, we have introduced the idea of designing and developing an Android game, using Android Studio IDE. The methodology of the project is based on the game Dig Rush that Ubisoft has jointly developed with Amblyotech. The game involves wearing 3D glasses with red–green glasses and directing the red square to hit the blue one to collect score. The game depends on two colors, which are mainly red and blue. In order to see different colors, a patient is required to wear 3D glasses. The information must be transmitted to both eyes enabling them to cooperatively work together. This way of treatment shows significant improvement, and more studies and evaluation should be considered in order to improve its efficacy. This study distinguished from the previous trials in being the first one that replaces the plastic glasses with VR Box Reality 3D Glasses, which obligate the children to see by their both eyes together and prevent seeing below the lower glasses border, besides these developed glasses put away the need to hold large devices preventing later complications on the neck. We hope to improve this game more by adding face recognition feature that ensures that the same patient continues the game.

**Acknowledgements** We would like to thank the patients who accepted to be involved in this study. Besides, we appreciate the Dammam Hospital staff for being cooperative and for facilitating our research.

## References

1. Hogarty, D.T., Hogarty, J.P., Hewitt, A.W.: Smartphone use in ophthalmology: what is their place in clinical practice? *Surv. Ophthalmol.* **65**(2), 250–262 (2020)
2. Silverstein, E., Williams, J.S., Brown, J.R., Bylykbashi, E., Stinnett, S.S.: Teleophthalmology: evaluation of phone-based visual acuity in a pediatric population. *Am. J. Ophthalmol.* **221**, 199–206 (2021)
3. Amblyopia (Lazy Eye) National Eye Institute (NEI).: (2020). <https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/amblyopia-lazy-eye>. Last Accessed 1 Mar. 2020
4. Amblyopia/Lazy Eye—No Patches, No Drops—Wow Vision Therapy.: (2020). <https://wowvision.net/amblyopia-lazy-eye-no-patches-no-drops/>. Last Accessed 1 Mar. 2020
5. Jukes, C., Bjerre, A., Coupe, J., Gibson, J.: Pilot study evaluating the feasibility of comparing computer game play with close work during occlusion in children aged 2–7 Years with Amblyopia. *Br. Irish Orthoptic J.* **15**(1), 115 (2019)

6. Gray, S.I., Bevan, C., Campbell, S., Cater, K.: Space vision: developing a game for vision screening and home-based monitoring. Proc. ACM Hum. Comput. Interact. **5**(CHI PLAY), 1–27 (2021)
7. Jost, R.M., Kelly, K.R., Hunter, J.S., Stager, Jr, D.R., Luu, B., Leffler, J.N., Dao, L., Beauchamp, C.L., Birch, E.E.: A randomized clinical trial of contrast increment protocols for binocular amblyopia treatment. J. Am. Assoc. Pediatr. Ophthalmol. Strabismus **24**(5), 282–e1 (2020)
8. Al-Tamimi, E.R., Shakeel, A., Yassin, S.A., Ali, S.I., Khan, U.A.: A clinic-based study of refractive errors, strabismus, and amblyopia in pediatric age-group. J. Family Community Med. **22**(3), 158–162 (2015). <https://doi.org/10.4103/2230-8229.163031>
9. Newsham, D.: Parental non-concordance with occlusion therapy. Br. J. Ophthalmol. **84**(9), 957–962 (2000). <https://doi.org/10.1136/bjo.84.9.957>
10. Gregson, R.: Why are we so bad at treating amblyopia?. Eye **16**, 461–462 (2002). <https://doi.org/10.1038/sj.eye.6700102>
11. Lazy Eye and Amblyopia Treatment with Lazy Eye Games (2020) (online) <http://lazyeyegames.com/>. Last Accessed 1 Mar. 2020
12. Access the game: <https://play.google.com/store/apps/details?id=com.sara.squareworld>
13. Access the website: <http://amblyopiagame.com>
14. Arabic Speaking Population in the World: How many people speak Arabic? (2020). <https://www.protranslate.net/blog/en/arabic-speaking-population-in-the-world-2/>. Last Accessed 2 Mar. 2020
15. Play therapy (2020). [https://en.wikipedia.org/wiki/Play\\_therapy](https://en.wikipedia.org/wiki/Play_therapy). Last Accessed 1 Mar. 2020

# A Hybrid Intrusion Detection Approach Based on Deep Learning Techniques



Diego F. Rueda, Juan C. Caviedes, and Wilmar Yesid Campo Muñoz

**Abstract** Intrusion detection systems (IDS) are designed to protect the networks from computer attacks. Through the constant monitoring of the network traffic, it is possible to identify anomaly behaviors that infer a likelihood of security threats. However, the growth of network traffic, the development of new techniques and algorithms to perform attacks, and the need to guarantee the security policies, force the community to research and develop novel intrusion detection models that are able to detect threats through anomalies in the traffic behavior. In this work, a novel intrusion detection model based on image recognition and classification algorithms is presented. In this proposal, each data record in the dataset is first converted into an image, and then convolutional neural networks (CNN) are used to perform feature extraction. Then, a support vector machine (SVM)-based algorithm is applied to identify the type of attack. The proposed hybrid model is trained and tested with the CIC-IDS2017 dataset. Experimental results evidence that our model is capable of detecting several intrusion threats with high accuracy and for some attack types our model outperforms the related work.

**Keywords** CNN · Deep learning · Hybrid intrusion detection · Network security · Network attacks · CIC-IDS2017 · SVM

---

D. F. Rueda (✉)

Institute of Informatics and Applications, Universitat de Girona, Girona, Spain  
e-mail: [u1930599@campus.udg.edu](mailto:u1930599@campus.udg.edu)

J. C. Caviedes

Department of Systems and Industrial Engineering, Universidad Nacional de Colombia, Bogotá, D.C., Colombia  
e-mail: [jcaviedesv@unal.edu.co](mailto:jcaviedesv@unal.edu.co)

W. Y. C. Muñoz

Department of Electronic Engineering, Universidad del Quindío, Armenia, Colombia  
e-mail: [wycampo@uniquindio.edu.co](mailto:wycampo@uniquindio.edu.co)

## 1 Introduction

Network security is a key aspect of any company's data policy nowadays. Roughly, these policies contemplate from application deployments, transfer of sensitive information, until the implementation of a communications network to offer external and internal services [11]. The security violations such as unauthorized accesses or intrusions can put at risk some of the established policies which, in summary, are related to availability, integrity, and confidentiality of the company information. The National Institute of Standards and Technology (NIST) defines an intrusion as the attempt to create a threat on the security policies or jump security mechanisms in networks or hosts [12]. These intrusion threats are fought by intrusion detection systems (IDS).

Previously, intrusion detection models were based on a catalog of threats that were updated periodically and protected only some parts of the network, such as centralized nodes or priority hosts. Thus, the system exposed other parts of the same network which the attackers use to invade a segment of interest. Accordingly, the trend of designing and implementing new methodologies and approaches to detect intrusion attacks is being highly influenced by the inclusion of machine learning-based methods [18].

Both, the threats and the traffic volume in networks, have increased at an exponential rate in the last years. For instance, only in the last ten years, traffic volume in mobile networks went from gigabytes to exabytes in monthly measures [8]. On the other hand, it is estimated that cyber-attacks can generate losses in the order of trillions of dollars in businesses [14]. Accordingly, the implementation of an intrusion detection model based on a static threat catalog is not relevant. Instead, it is better to choose a dynamic option to continuously monitor and classify traffic threats. Many types of attacks can dramatically affect a network, some of them are denial of service (DoS), distributed denial of service (DDoS), user to root (U2R), web attack, infiltration, or probing and remote-to-local (R2L) attacks.

Typically, IDS can be classified into two types: host-based and network-based. In host-based IDS, some software is installed on a particular host (e.g., antivirus plugins) and its operation is reactive; i.e., it waits for a potential attack to enter its domain and then neutralizes it. A network-based IDS is similar, only that it is implemented in network elements such as firewalls, allowing the analysis of network traffic coming from or going to multiple hosts [13]. At a functional level, the intrusion detection models can detect anomalies based on normal traffic behavior or directly classify attacks based on previous training. The advantage of detection based on normal traffic patterns is that intrusions that are unknown (i.e., they are not known in the training stage) can be detected. However, this can produce a decrease in the efficiency of the model due to the increase of false positives when anomalous behavior is generated by normal traffic. Despite this, the classification of attacks according to a training bench is ineffective against unknown attacks [24].

In order to address the problems presented in traditional intrusion detection models, machine learning, as well as deep learning techniques, is very effective in detecting such attacks [5, 6]. Support vector machine (SVM) [7, 9], neural networks

[15, 19, 25], and clustering algorithms are widely studied techniques in this field. The combination of intrusion detection techniques in data preparation, data processing, and data classification is considered emergent and has many potentialities. Although there are several proposals for intrusion detection, most of them have been tested on old datasets such as NLS-KDD [20] that do not consider the diversity of contemporary attacks as well current changes in traffic behavior.

This work aims to implement a hybrid intrusion detection model using deep learning and other traditional machine learning technique. Thus, the major contributions of this paper are the use of image recognition based on convolutional neural networks (CNN) to perform feature extraction of traffic patterns, and image classification using an SVM to identify the type of attack. Furthermore, the model has been trained and tested by using a modern dataset called CIC-IDS2017 [16] that contains several types of attack and allows its performance to be measured in order to compare the proposed model with previous approaches.

The rest of this paper is structured as follows: Sect. 2 contains a general review of previous work. The proposed intrusion detection model is described in Sect. 3. In Sect. 4, dataset description and preparation are provided. Performance analysis and detection results are also discussed. Finally, the conclusions and future work are presented in Sect. 5.

## 2 Related Work

Traditional network intrusion detection methods are rule-based, ignore contextual information, and because the size of the data is extremely high, these methods introduce complexity and reduce detection accuracy. In order to deal with these problems in the literature, several works have implemented intrusion detection methods based on machine learning [5] or deep learning techniques [6]. However, deep learning techniques are being widely used to improve the accuracy of intrusion predictions instead of traditional machine learning-based methods. The intrusion detection method presented in Ludwig [11] combines multiple assembled classifiers in a way that the individual results are merged in favor of multi-class classification. In this work, the authors used various deep neural networks (DNN) to distinguish normal behaviors from attacks. While the results show that a precision above 95% can be achieved applied to NLS-KDD dataset, the authors suggest that by including more techniques in the model, overall precision can be improved.

A hybrid model, based on the auto-encoder network (AN) to feature dimensionality reduction and long short-time memory (LSTM) to predict intrusion detection types, is addressed in [25]. With this proposal, the accuracy of this method is improved by 2% on average when compared with classical IDS. A semantic re-encoding and deep learning model (SRDLM) is proposed in Wu et al. [22] for intrusion detection. The SRDLM model re-encodes the semantics of network traffic (i.e., transform traffic data to words), increments the distinguishability of traffic, and improves the general-

ization by using deep learning techniques (i.e., ResNet network architecture). Results showed that the SRDLM method achieves more than 99% of accuracy to detect the Web character injection attack.

A deep learning network is implemented in [4] to automatically build a smart intrusion detection model. The authors rely on hybrid optimization framework (IGASAA) based on improved genetic algorithm (IGA) and a simulated annealing algorithm (SAA). This approach, which is called machine learning IDS (MLIDS), uses the IGASAA to find the optimal combination of the most relevant values that will serve as parameters in the construction of the IDS using a DNN. The values to consider are the input features, the data normalization, the activation function for the neural model, the learning rate, and the momentum. Optimal selection ensures a high efficiency of the model in terms of its hit rate, precision, and false-positive rate. The results obtained with this technique show an accuracy greater than 99.8%, which exceeds other approaches compared by authors.

A technique that shows promising results to improve the accuracy, false alarm rate, and timeliness of traditional intrusion detection algorithms is to convert the traffic data of an incoming network into images to transform the intrusion detection in an image classification problem. According to Xiao et al. [23], the use of CNN provides a method to automatically extract the features of the dimensionality reduction data, and the supervised learning to extract more effective information for intrusion identification. In this approach, the computational cost is reduced by converting the original traffic vector format into an image. The simulation results on NLS-KDD dataset indicated that the model reaches a detection accuracy of 94.0%, but for U2R and R2L attacks, the detection rates are significantly low at 20.61% and 18.96%, respectively. In [19], the traffic data of an incoming network is represented in grayscale images, thus transforming the anomaly detection problem to an image processing problem where texture is the key for detection. The authors also use the NSL-KDD dataset for model implementation, training, and validation. The results showed that the performance reaches a precision higher than 97.8%. However, the execution time of the model is very long when the number of layers of the CNN increases. Consequently, the authors mentioned the need for more work to improve the proposed model.

In more recent work, network traffic feature (NTF) is transformed into four-channel Red, Green, Blue, and Alpha (RGBA) images [21]. In [21], a multistage deep learning image recognition system (ResNet50) employing transfer learning is proposed to detect contemporary malicious behavior (network attack) and to recognize the attack type. Empirical quantification of the attack type recognition allowed to achieve 99.8% in detection accuracy of the generic attack on the UNSW-NB15 dataset, and 99.7% in detection accuracy of the DDoS attack on the BOUN DDoS dataset. In [17], authors proposed a hybrid model which combines an LSTM for feature extraction and a CNN for intrusion detection. The model validation in the UNSW-NB15 dataset showed a detection accuracy of 98% which improves the performance of RNN-based intrusion detection models.

Note that most of the literature reviewed focused on external intrusions. In fact, there are attacks such as operating system scripts that can be represented as internal

**Table 1** Summary of related work

References	Strategy	Dataset	Performance (%)
Ludwig [11]	AN, DBN, DNN, ELM	NLS-KDD	95.0
Zhang et al. [25]	AN-LSTM	NLS-KDD	99.6
Wu et al. [22]	SRDLM	NLS-KDD	99
Chiba et al. [4]	MLIDS	CICIDS2017, NSL-KDD, CIDDS-001	99.8
Xiao et al. [23]	CNN	NLS-KDD	94.0
Tao et al. [19]	CNN	NLS-KDD	97.8
Toldinas et al. [21]	CNN	UNSW-NB15, BOUN DDoS	99.7
Syms et al. [17]	CNN-LSTM	UNSW-NB15	99.7
Rhode et al. [15]	RNN	VirusTotal	95.0

threat per host. In [15], authors have studied the possibility of predicting whether a script executed in the operating system may be malicious based on a short sample of the data it manipulates. Consequently, the authors propose a recurrent neural network (RNN) to predict malicious behavior based on data from the operating system. The scope was to study the ability of the model for detecting malware families and variants that have not been previously stored known as zero-day attacks. In training the model, around 3000 malware samples are used, reaching 95% accuracy when one second of malicious code execution has passed.

Regardless of the type of technique used for the implementation of intrusion detection systems, in Table 1 can be seen that there is a trend toward the use of deep learning techniques to enhance the models accuracy compared to traditional methods. On the other hand, the possibility of implementing techniques based on image processing is highlighted to transform each data record of the dataset into an image. Most of the works reviewed can provide a guide to design an intrusion detection model applying deep learning techniques. But, to evaluate the performance model, it is relevant to use a dataset that includes modern attack types and traffic patterns rather than just data contained in legacy dataset such as NSL-KDD.

In this paper, the proposed hybrid method for intrusion detection is supported on a CNN to perform feature extraction of traffic patterns. Furthermore, in order to provide the capability to detect several types of attack, a SVM-based classifier is incorporated into the proposed model, because SVM has proven to be effective in intrusion classification problems [7, 9, 23]. Therefore, unlike previous works, a hybrid intrusion detection model using deep learning and a classification algorithm is proposed to address the detection of several contemporary types of attack. To train and test the proposed hybrid model, the CIC-IDS2017 dataset is considered. Note that this dataset is cleaned and normalized, eliminating out-of-range data, and using a common data scale for used features.

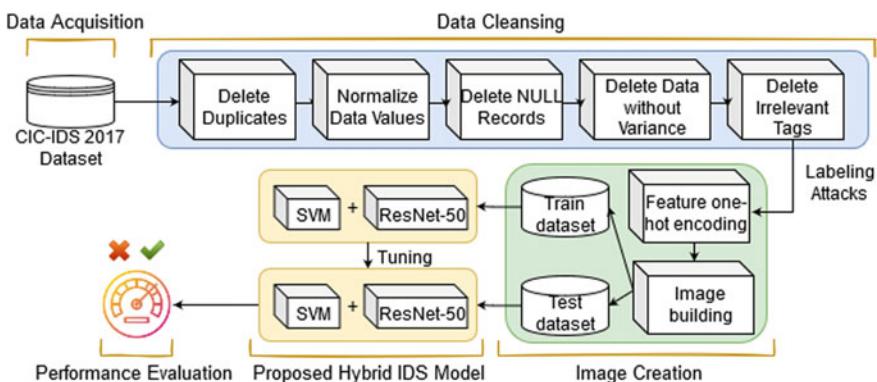
### 3 Proposed Method

In order to take advantage of deep learning and traditional machine learning techniques for intrusion detection, a hybrid model is proposed. The key to the implementation of this model arises in image recognition and a classifier algorithm. In this section, the hybrid model implementation process is described.

#### 3.1 Hybrid Intrusion Detection Model Description

Figure 1 summarizes the overall framework used to detect and classify attacks using the proposed hybrid model. The model implementation considers six fundamental steps:

- **Step 1. Data cleansing:** To generate a normalized, balanced and diverse dataset for training the proposed model. The key is to preserve the features with high variability while a dimension reduction is applied.
- **Step 2. Labeling attacks:** To categorize attacks of the same type in a unique label in order to group attacks with similar effects in the network behavior or damage.
- **Step 3. Feature normalization and image creations:** To balance the dataset and generate a bank of images through the transformation of each traffic record in an  $8 * 8$  image with 8-bit depth.
- **Step 4. Image dataset consolidation:** To train the model based on image recognition and test the classification performance. The dataset is randomly divided into 70% records for training and 30% for testing. Next, an image scaling technique and a summer color map are applied to convert it into a  $224 * 224$  RGB image.



**Fig. 1** Process flow of the proposed hybrid IDS model

- **Step 5. Hybrid intrusion detection model:** To implement the hybrid intrusion detection model combining CNN and SVM algorithms: CNN for feature extraction and image recognition, and SVM for attacks classification.
- **Step 6. Performance evaluation:** To analyze the detection accuracy by applying the hybrid model in the testing dataset.

### 3.2 Implementation Process

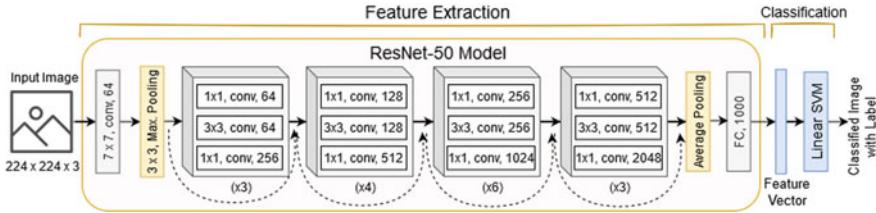
A convolutional neural network (CNN) processes data with a grid pattern, such as images, and to automatically and adaptively learn spatial hierarchies of features, from low-to high-level patterns. The CNN is built as a set of three layers between the input and output layers: a convolutional layer, a pooling layer, and a fully connected layer. The number of convolution layers as well as the number of pooling layers defines the CNNs structure. Suppose the CNN input feature is feature map of the layer  $i$  is  $M_i$  ( $M_0 = X$ ). Then, Eq. (1) expresses the convolution process as [23]:

$$M_i = f(M_{i-1} * W_i + b_i) \quad (1)$$

where  $W_i$  corresponds to the convolution kernel weight vector of the  $i$  layer;  $*$  represents the convolution operation;  $b_i$  corresponds to the offset vector of the  $i$  layer; and  $f(x)$  is the activation function and corresponds to the *ReLU* function. This activation function is widely used in CNN to replace all negative values in the feature map to zero. By specifying different window values, distinct feature information is extracted from the  $M_{i-1}$  data matrix in the convolutional layer, and through different convolution kernels, distinct features  $M_i$  in the data are extracted. The same parameters (weight and offset) are shared in the same convolutional kernel of the convolutional operation, which substantially reduces the number of parameters in the whole CNN. In the pooling layer, the feature map is mapped according to different sampling rules after the convolutional layer. The feature dimension is reduced in the pooling layer, and as a consequence, the influence of redundant features in the model is also decreased.

In the proposed hybrid model, the advantages of a CNN are exploited. Therefore, in order to provide the appropriated input to the proposed hybrid intrusion detection model, as part of the implementation, each record of the cleansed and normalized dataset is transformed in an  $8 * 8$  image with 8-bit depth. This image bank is the input of a Residual Network 50 (*ResNet50* [1]) CNN for image feature extraction. It receives a  $224 * 224$  image of three layers (i.e., RGB), so the constructed images are transformed using a nearest-neighbor interpolation image scaling technique and grayscale to color conversion using a *summer* color map. Figure 2 is presented the implementation details of the *ResNet50*.

A *ResNet50* architecture has demonstrated high accuracy for intrusion detection [21, 22]. On the other hand, as shown in [3], a ResNet architecture, particularly *ResNet50*, has less computational complexity than other CNN models such as VGG



**Fig. 2** Implementation details of the proposed hybrid IDS model

or AlexNet. Likewise, it is more accurate than models like GoogleNet or ShuffleNet. While the *ResNet101* or *ResNet152* improve the accuracy of the model, they increase its complexity to a greater extent compared to the *ResNet50*. For these reasons, the *ResNet50* was chosen as CNN because it maintains a balance between precision and computational complexity compared to other models.

Additionally, the capability to classify various types of attacks in the proposed hybrid model is provided by a support vector machine (SVM). This is because SVM has been shown to be an effective method to train the base learners in intrusion classifiers to detect several types of attack [7, 23], and the combination CNN-SVM achieves a better performance than FCN or the use of other classifiers [2]. Thus, the second last output of the *ResNet50* is considered to train an SVM classifier, and its input is a vector with 1000 features of the image built to represent the intrusion attack and being traffic. In the SVM classifier, this input vector is first mapped into a higher-dimensional feature space where the optimal separation hyperplane is obtained [4].

Furthermore, a decision boundary, which is basically the separation hyperplane, is defined by support vectors rather than all training samples, and thus, the SVM provides high robustness to outliers. In the proposed model, a linear SVM is implemented that encodes the input samples using a one vs all encoder, which consists of dividing the multi-class classification problem into multiple binary classification problems. In this way, a new entry only satisfies the condition imposed by the region of the closest class. Finally, the output of the SVM is the classification of the input vectors into the attack classes and being traffic as learned from the input labels. The complexity of the proposed model is given by the ResNet50, i.e.,  $O(n^4)$ .

## 4 Simulation and Results

In order to test and evaluate the accuracy of the proposed hybrid model for intrusion detection, we have implemented the model in MATLABR2020b using a machine with 4 CPU, 16 GB of RAM, and 1 TB of storage. This section describes the data preparation and normalization processes and the image creation to consolidate the image bank. Finally, the performance analysis of the hybrid model is presented.

## 4.1 Dataset Preparation

Sharafaldin et al. [16] proposed the CIC-IDS2017 dataset with the eleven most significant characteristics required by the Canadian Institute for Cybersecurity (CIC): attack diversity, anonymity, available protocols, full capture, full interaction, full network configuration, full traffic, feature set, heterogeneity tagging, and metadata. Compliance with these characteristics makes the dataset contains 13 up-to-date attacks that resemble data from networks deployed in reality. In addition, it has records for benign traffic, and all of them are labeled [16]. The network architecture used to collect data is based on two networks: attack and victim [16].

The selected dataset consists of 78 columns, an additional column labeled the type of attack, and 2.8 millions of records. In this work, all the fields in the dataset are not used for the analysis of our hybrid approach because they have no relevance to the intrusion detection case study. For this reason, columns 1 and 44–51 of the dataset were removed, leaving a total of 69 columns. The cleaning process (presented in Fig. 1) has the main objective to preserve the features in the dataset that generate more variability, delete duplicated records and keep the attacks with more diversity of records. It is important to note how the features that represent attributes of the size of the packets (either sent or received), duration of the traffic session during attacks and other time variables, such as inter-arrival time, have high variability in the dataset. The result is a dataset with 24 features and about 2.5 million records that can be grouped 4 types of measurements:

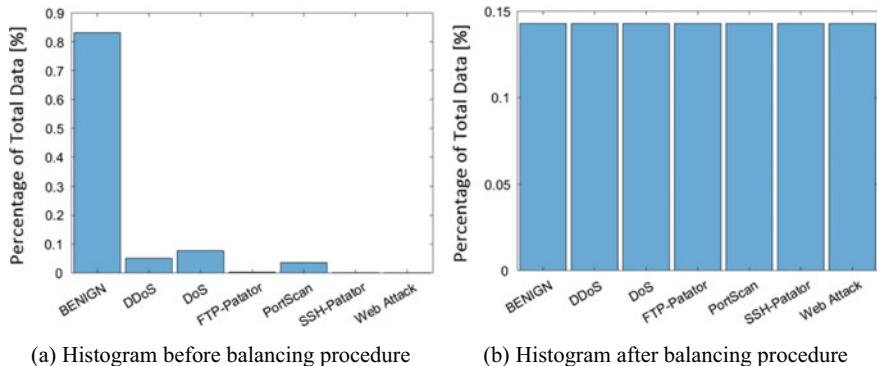
1. **Four measures of traffic for total packets and their lengths:** *Total Fwd/Bwd Packets* and *Total Length of Fwd/Bwd Packets*
2. **Eight measures of forwarding and backwarding packages involved in communication:** *Fwd Packet Length Max/Min/Mean/Std* and *Bwd Packet Length Max/Min/Mean/Std*
3. **Four measures of duration time of the flows in communication:** *Flow Duration*, *Flow Bytes\_s*, *Flow Packets\_s*
4. **Eight measures of inter-arrival time for the communication flows and the forwarding packets:** *Flow IAT Mean/Std/Max/Min* and *Fwd IAT Total/Mean/Std/Max/Min*

At this point, the dataset is still labeled with eleven different attacks. However, analyzing the dataset it can observe that there exist some attacks with several subtypes, but globally these represent a single attack, e.g., DoS Hulk, DoS Goldeneye, and DoS Slowloris can be represented as a DoS attack. The difference between them is the script that generates them. Moreover, some attacks such as Heartbleed, Infiltration, and Botnet have irrelevant representation in the dataset as they have less than 0.01% of total dataset size so the associated records are excluded from the analysis of this work. After filtering and grouping procedure, the attacks were thus classified into the seven classes as shown in Table 2.

When a histogram is generated to see how many records belong to each label, it becomes evident that the data is unbalanced as shown in Fig. 3a. In order to balance

**Table 2** Attacks classification in new classes

New label	Dataset label
Benign ( <i>B</i> )	Benign
DDoS ( <i>A1</i> )	DDoS
DoS ( <i>A2</i> )	DoS goldeneye, DoS hulk, DoS slowhttptest, DoS slowloris
FTP patator ( <i>A3</i> )	FTP patator
PortScan ( <i>A4</i> )	PortScan
SSH patator ( <i>A5</i> )	SSH patator
Web attack ( <i>A6</i> )	SQL injection, XSS, brute force

**Fig. 3** Data balancing result per each traffic label

the amount of data per attack, the type with the fewest number of records is taken into account as a reference to others. Thus, Web Attack with a total of 2.1 thousand records is selected to limit the number of data records per attack class. Then, from each class, a random sample of the same amount of records is taken, so a new dataset is built with near 15 thousand records which contain the six types of attack and an additional class for benign traffic. Balancing the data allows avoiding bias in the training of a neural network, in this case, the *ResNet-50* CNN. The result of balancing the number of records per attack is shown in Fig. 3b.

## 4.2 Dataset Normalization

For dataset normalization, it is considered that the range between maximum and minimum values in some features is too large and needs some preprocessing. First, we apply a logarithmic function to shorten the range. Note that all features are in the positive domain because their measures are related to lengths, time, or quantities.

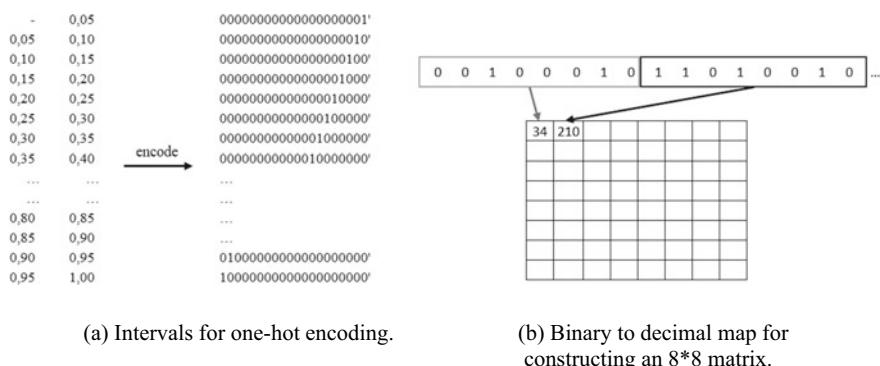
However, zero value is possible. Therefore, before applying the logarithmic function, a unit is added to all values. The next step is to perform a linear normalization using Eq. (2).

$$x_i'' = \frac{x_i' - \min(x_i')}{\max(x_i') - \min(x_i')} \quad (2)$$

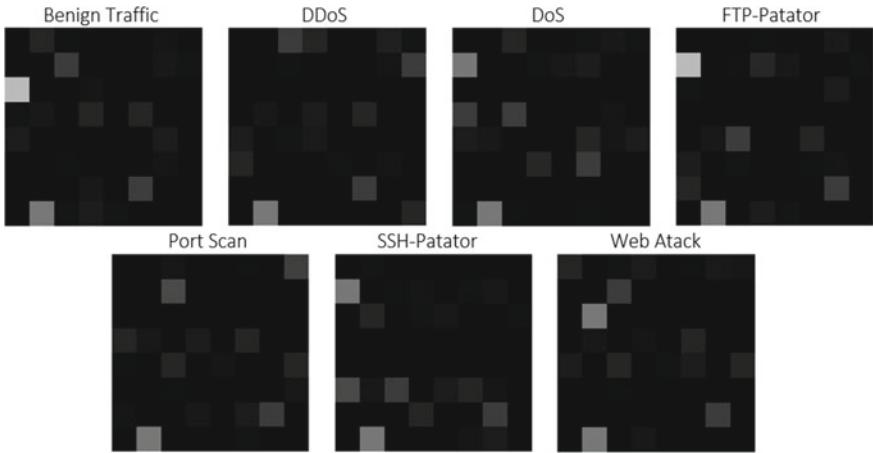
where  $x_i' = \ln(x_i + 1)$  and  $x_i''$  is the normalized value. With this normalization, all features in the dataset are compressed to a range from 0 to 1. When a descriptive analysis is performed separating the benign from the malign traffic (i.e., grouping all attacks in one unique class), the result shows that attacks typically have more packets sent but fewer packets received which is to be expected in attacks such as a DoS attack. Also, the total size of packets sent shows that in the attacks there is greater variability toward values closer to zero. These considerations are key for training the model and discussing the results

### 4.3 Image Creation

Each record of the cleansed and normalized dataset is converted into an  $8 * 8$  image with 8-bit depth in concordance with the method presented in Li et al. [10]. Thus, 20 intervals are set for encoding the values of the features using *one-hot encoding* as shown in Fig. 4a; i.e., each normalized value of 24 features or metrics is considered as a symbol that is encoded into a word of 20 bits. Once this is applied to each record of the normalized traffic dataset, the result is a new binary dataset with 480 columns. Considering the target dimensions of the image, 32 columns with zero values are added to dataset to complete 512 columns. After that, an  $8 * 8$  matrix is constructed using a binary to decimal conversion each 8-bits in the record as shown in Fig. 4b.



**Fig. 4** Encoding procedure performed on each data record



**Fig. 5** Images created for the traffic dataset records

It is important to have in mind that the same procedure is applied to all data records in the cleansed dataset to generate an image bank with six types of attack and also being traffic.

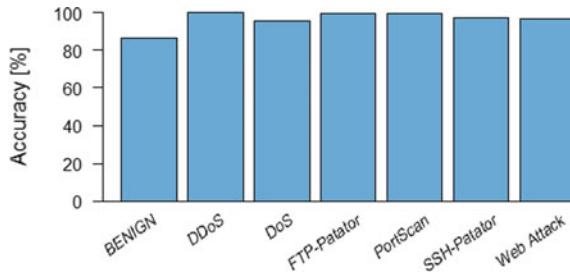
Then, the decimal values in the matrix are converted into an image using a grayscale conversion with 8 bits of depth. With this method, for each class of attack in the dataset, at least two thousand images are obtained to be used as input of the deep learning model. An example of the resulting images for some data records is shown in Fig. 5. Since the *ResNet50* has  $224 * 224$  RGB images as input, an image scaling method is applied, as well as a summer-type color map. This ensures that the images meet the *ResNet50* entry conditions.

#### 4.4 Performance Evaluation

Performance evaluation of the proposed hybrid model is carried out by considering the accuracy and the confusion matrix generated from the detection results as shown in Table 3. Attack classified correctly or incorrectly by the model is represented as *T* (True) or *F* (False), respectively. *P* (Positive) and *N* (Negative) symbolize the prediction results of the hybrid detection model as an attack or being traffic, respectively. In this sense, four groups (*TP*, *TN*, *FP*, and *FN*) are considered to categorize the output of the hybrid model. If the detection result of the hybrid model is an attack for testing data, and the detection result is correct, then the result is *TP*; i.e., the model has detected and classified appropriately the attack; *TN* indicates that the detection result of the model is positive and correct; i.e., benign traffic is not detected as an attack; *FP* means that the model predicts the data as an attack, but the detection result is incorrect; i.e., the benign traffic is detected as attack; *FN* indicates that the model

**Table 3** Confusion matrix using the proposed hybrid model

	<i>B</i> (%)	<i>A1</i> (%)	<i>A2</i> (%)	<i>A3</i> (%)	<i>A4</i> (%)	<i>A5</i> (%)	<i>A6</i> (%)
<i>B</i>	86.7	2.6	3.7	0	3.0	1.2	2.8
<i>A1</i>	0	100	0	0	0	0	0
<i>A2</i>	3.7	0	95.3	0.2	0	0.2	0.6
<i>A3</i>	0	0	0	99.7	0	0	0.3
<i>A4</i>	0	0	0	0	99.7	0	0.3
<i>A5</i>	0.8	0.5	0.3	0.3	0	97.5	0.6
<i>A6</i>	1.9	0	0.8	0	0	0.5	96.8

**Fig. 6** Detection accuracy results for the proposed hybrid model

predicts the data as benign traffic, but the detection result is erroneous; i.e., attack traffic is classified as benign traffic. As can be seen in Eq. (3), accuracy (AC) represents the probability that the samples are correctly classified by the hybrid model with respect to the total number of samples.

$$AC = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

Figure 6 shows that our model can achieve 86.7% detection accuracy in the case of benign traffic (*B*), 95.3% in the DoS (*A2*) case, 96.8% for Web Attack (*A6*), and 97.5% for SSH Patator (*A5*) when it is applied to a contemporary dataset such as CIC-IDS2017. Also, the accuracy obtained in the classification of DDoS (*A1*), FTP Patator (*A3*), and PortScan (*A4*) attacks is greater than 99%. One possible explanation for the relatively low accuracy in benign traffic is that benign traffic generated by applications such as bittorrent, online video games, or video conferencing is more likely to have been labeled as an attack than an attack as benign traffic. This is an important consideration when choosing a dataset because the dynamic behavior in modern services demand to generate multiple patterns in network traffic measures. In addition, the number of benign data records is the same that each attack data record, so in this work, the probability of benign data record being classified as an attack is higher than if all the data had been considered.

In order to compare the performance of the proposed hybrid model for intrusion detection with other related works, the average of the accuracy obtained for all traffic

classes is calculated. Thus, the proposed hybrid model has 96.53% accuracy in the average for intrusion detection. Other models like the one proposed in Wu et al. [22] can achieve a 94.03% accuracy when applying a semantic re-encoding and deep learning model but on the NSL-KDD dataset. Then, the proposed hybrid model outperforms the detection accuracy of this model. Regarding the model proposed in Zhang et al. [25], which is based on an auto-encoder and an LSTM, it achieved 97.6% accuracy for benign traffic classification and 95.3% for the DoS case on the NSL-KDD dataset. If compared to this last case, the proposed hybrid model is capable of detecting DoS attacks with better accuracy.

Similar to the approach in this paper, in [19] and [23] the authors have already proposed a method that converts the traffic data into an image and transforms the anomaly detection problem into an image processing problem. Despite these two works also considering the use of CNN as part of their models, they had not contemplated the use of an external classifier to detect the type of attack. In [23], principal component analysis (PCA) is considered data dimensionality reduction as part of data preprocessing in the entire IDS. Thus, detection accuracy using the IDS-CNN model on the NSL-KDD dataset is 94.0% [23]. Whereas in [19] the CNN is used for intrusion detection with any other technique for images classification of the NSL-KDD traffic patterns achieving a detection accuracy of 97.8%. Compared with these results, our hybrid approach is capable of detecting several intrusion threats with a similar accuracy but in a contemporary dataset.

## 5 Conclusions and Future Work

In this work, a hybrid intrusion detection model was implemented using a deep learning framework in combination with traditional machine learning techniques on a modern dataset. Our approach uses convolutional neural networks (CNN) to perform feature extraction of traffic patterns and classification using support vector machines (SVM) to identify the type of attack. Experimental results demonstrated that it is possible to use image processing techniques to characterize network traffic in order to detect anomalies related to intrusion attacks.

Moreover, with the proposed hybrid approach, it was possible to obtain a global accuracy of 96.53% and more than 99% accuracy in the recognition of attacks such as DDoS, FTP Patator, and PortScan. Compared with previous works, the global precision reaches similar values in the recognition of attacks with the advantage that it was tested to a contemporary dataset that contemplates several types of behavior in network traffic. Therefore, the combination of deep learning techniques can be considered an interesting strategy to improve the effectiveness of intrusion detection systems.

An intrusion detection that is not applied to a real environment does not show its true functionality to protect a network. For this reason, in future work the deployment of the proposed hybrid model in some network environment either real or simulated

will be carried out. Also, some attacks can be included in the approach such as zero-day attacks. The aim is to build a zero-day attack system and retrain the proposed approach to detect it. Last, new techniques to create images from the dataset (e.g., RGBA) and other classifiers models [e.g., K-nearest neighbors (KNN) or random forest (RF)] can be considered to address a new comparative study with the proposed model and other contemporary datasets (e.g., UNSW-NB15 or BOUN).

## References

1. Akiba, T., Suzuki, S., Fukuda, K.: Extremely large minibatch SGD: training ResNet-50 on ImageNet in 15 minutes. In: 2017 Conference on Neural Information Processing Systems NIPS (2017)
2. Basly, H., et al.: CNN-SVM learning approach based human activity recognition. In: Image and Signal Processing, pp. 271–281. Springer, Heidelberg (2020)
3. Bianco, S., et al.: Benchmark analysis of representative deep neural network architectures. IEEE Access **6**(4), 64270–64277 (2018)
4. Chiba, Z., et al.: Intelligent approach to build a deep neural network based IDS for cloud environment using combination of machine learning algorithms. Comput. Secur. **86**, 291–317 (2019)
5. Chih-Fong, T., et al.: Intrusion detection by machine learning: a review. Expert Syst. Appl. **36**, 11994–12000 (2009)
6. Ferrag, M.A., et al.: Deep learning for cyber security intrusion detection: approaches, datasets, and comparative study. J. Inf. Secur. Appl. **50**, 102419 (2020)
7. Gu, J., et al.: A novel approach to intrusion detection using SVM ensemble with feature augmentation. Comput. Secur. **86**, 53–62 (2019)
8. Jonsson, P., et al.: Ericsson mobility report. Technology Report. Ericsson (2021)
9. Kuang, F., Xu, W., Zhang, S.: A novel hybrid KPCA and SVM with GA model for intrusion detection. Appl. Soft Comput. **18**, 178–184 (2014)
10. Li, Z., et al.: Intrusion detection using convolutional neural networks for representation learning. In: Lecture Notes in Computer Science (LNCS), pp. 858–866. Springer (2017)
11. Ludwig, S.A.: Intrusion detection of multiple attack classes using a deep neural net ensemble. In: 2017 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1–7. IEEE (2017)
12. Mell, P., Grance, T.: The NIST definition of cloud computing. <https://csrc.nist.gov/publications/detail/sp/800-145/final> (2011)
13. Mohammed, M., Pathan, A.S.K.: Intrusion detection and prevention systems (IDPSs). In: Automatic Defense Against Zero-day Polymorphic Worms in Communication Networks, Chap. 3, 2nd edn., pp. 47–84. Auerbach Publications (2013)
14. Rafter, D.: Cyberthreat trends: 2019 cybersecurity threat review (2019)
15. Rhode, M., Burnap, P., Jones, K.: Early-stage malware prediction using recurrent neural networks. Comput. Secur. **77**, 578–594 (2018)
16. Sharafaldin, I., et al.: Toward generating a new intrusion detection dataset and intrusion traffic characterization. In: ICISSP 2018 4th International Conference on Information Systems Security and Privacy, pp. 108–116 (2018)
17. Smys, S., Basar, A., Wang, H.: Hybrid intrusion detection system for internet of things (Iot). J. of ISMAC **2**(4), 190–199 (2020)
18. Stalling, W.: Network Security Essentials: Applications and Standards, 6th edn., Pearson Education (2017)
19. Tao, W., et al.: A network intrusion detection model based on convolutional neural network. In: Security with Intelligent Computing and Big-data Services, pp. 771–783. Springer, Heidelberg (2020)

20. Tavallaei, M., et al.: A detailed analysis of the KDD CUP 99 data set. In: IEEE Symposium on Computational Intelligence for Security and Defense Applications, CISDA 2009, pp. 1–6. IEEE (2009)
21. Toldinas, J., et al.: A novel approach for network intrusion detection using multistage deep learning image recognition. *Electronics* **10**(15) (2021)
22. Wu, Z., Wang, J., Hu, L., Zhang, Z., Wu, H.: A network intrusion detection method based on semantic re-encoding and deep learning. *J. Netw. Comput. Appl.* **164**, 102688 (2020)
23. Xiao, Y., et al.: An intrusion detection model based on feature reduction and convolutional neural networks. *IEEE Access* **7**, 42210–42219 (2019)
24. Yin, C., et al.: A deep learning approach for intrusion detection using recurrent neural networks. *IEEE Access* **5**, 21954–21961 (2017)
25. Zhang, Y., et al.: A network intrusion detection method based on deep learning with higher accuracy. *Procedia Comput. Sci.* **174**, 50–54 (2020)

# Develop a Smart Data Warehouse for Auto Spare Parts Autonomous Dispensing and Rack Restoration by Using IoT with DDS Protocol



**R. Shiva Shankar, Ravibabu Devareddi, Gadiraju Mahesh,  
and V. MNSSVKR Gupta**

**Abstract** Nowadays, the concept of restoring and delivering auto spare parts is maintained in the warehouse with the help of cloud service facilities and IoT. These two technologies (IoT and cloud computing) can solve customer and warehouse managers' real-time problems. It is used to identify the exact spare part with acceptable period and avoidable delay by the traditional human-operated service instead of non-avoidable delay like customs. The tremendous growth of the Internet of Things (IoT) and cloud computing has provided a great solution to optimizing the delay time to deliver an item to the customer and retain the spare parts in the warehouse with accuracy. Hence, the out-of-stock issues are optimized. The main objective of this work is to design a system to organize the spare parts with passive UHF-RFID tags/stickers. A web server runs by a Raspberry Pi 3 controller, which issues commands to the ESP32 processor to enable the UHF-RFID reader/writer module. ESP32 will give an automatic robotic handler the message to pick the exact spare part by the identity of the passive RFID tag from the respective rack and place it on the standard conveyor belt to the delivery section, vice versa. The Raspberry Pi 3 module receives a request from GUI-based dashboard or mobile app. The DDS protocol can issue topics and store them as a publish–subscribe model to the IoT-cloud platform. Thus, entire system works as an intelligent retrieve and stores spare parts in the warehouse.

**Keywords** UHF-RFID tags · Data distribution service (DDS) · IoT · Cloud computing · ESP32 · Graphical user interface (GUI) · Publish–subscribe model · DDS protocol

---

R. Shiva Shankar (✉) · R. Devareddi · G. Mahesh · V. MNSSVKR Gupta  
Department of Computer Science Engineering, SRKR Engineering College, Bhimavaram, Andhra Pradesh, India  
e-mail: [shiva.shankar591@gmail.com](mailto:shiva.shankar591@gmail.com)

## 1 Introduction

Automatic systems with innovative intelligence and IoT support play a crucial role in industries [1], domestic purposes, and other organizations with various environments. Object recognition is a very vital part of IoT [2]. Warehousing is an essential aspect of spare parts distribution [3], with automakers balancing suppliers, IT systems, size, location, and picking technology. It is a very significant challenge to maintain the spare parts warehouse. Every automaker who understands the value of after-market support can easily retrieve and store in optimum time with accuracy [4]. For example, dealers want a car in a single night if it is not possible on the same day. Delivering a vehicle off the road implies a disgruntled dealer and a disgruntled customer, who may be selected for a different manufacturer and replace their vehicle. The ideal blend of complexity and synchronization, with the assembly of appropriate spare parts serving as the backbone, determines the character of a great car. Every participant in the car business must consider consumer preferences for factors like styling, comfort, and safety [5]. These needs may be satisfied if the correct replacement parts are available at the right time and in the right place. Therefore, carmakers are pushed to control the warehouse and the appropriate quantity of inventory [6] in the correct location. A great deal of thinking, analysis, planning, and care will provide after-market service and create and maintain the best facility.

There are seven recommended practices for managing spare parts inventories [7] to help manage more effectively. The first step is to identify all the components: As a result, some of these components may not be stocked and may not enter into the inventory control system as a part of the record. (2) Use the Bill of Materials (BOM) and manage it [8]: Ordering items and placing work orders will be considerably easier using BOMs. (3) Make the work order procedure more efficient: Have a solid yet straightforward procedure to follow the generated work orders for all the components circulation to correct inventory. (4) Put in place security measures: Ensure that the components in the warehouse have adequate security measures in place has to be examined. They maintain inventory accuracy, restrict parts' access, and adopt a policy stating that parts inventory reflects as "off bounds" to anyone. Physical security measures [9] such as parts counters restrict personnel from entering inventory stock locations, withdrawing necessary components, and allowing everyone to access the component inventory, which can rapidly lead to inaccuracies. Installing security camera systems [10] and implementing badge access to the entrance and exit points are the second technique to help this goal. (5) Consolidate and centralize portions: Centralize and integrate satellite parts inventories into the central parts warehouse wherever possible. Having all the parts centralized and consolidated improves security and makes it easier to monitor and maintain inventory accuracy for this potentially huge asset. (6) Use an inventory control system to manage the parts inventory rather than spreadsheets: Using ERP's warehouse [11] capability or a WMS [12] to manage the parts inventory will assure accuracy and simplicity of management. Using barcodes and scanning the capability in conjunction with the system will also substantially increase the efficiency of the parts warehouse procedures and inventory

accuracy [13]. (7) Assign each item to a specific stock location so that employees may find it quickly: Ensure that stock locations are produced with the least amount of detail that the systems will allow (usually a bin and slot location for each SKU [14]). When the location identifier is a shelf with many other parts, it might be difficult for staff to find a part. By implementing the guidelines greater control over the parts inventory is gained and faster work orders and requests are enabled. In this work a solution is presented that would improve the security of spare parts for improving audit accuracy and would replace the employee's manual access ability. To get to the precise rack of spare parts location, the RFID reader scans the tags equipped to each shelf in the warehouse. After determining the position, the ESP32 controller sends a message to the robotic handler and spare parts will deliver to the area through an automatic standard conveyor belt [15], where they are placed on a standard automated conveyor using a lift assistance robotic hand to pick [16]. The spare part will store in the corresponding rack based on the SKU ID with the respective RFID tag [17]. Each spare part or group of spare part is to be identified by the microcontroller (ESP32) [18] and issued as a topic to publish or subscribe-based model in IoT [19] with DDS protocol [20].

## 2 Literature Survey

Tejesh and Neeraja [21] presented a warehouse inventory management system participating vital role in the issue of good and management systems. They considered RFID as one of the wireless communication technologies that may fit the inventory management process identification. To open-source hardware through a wireless link via the Internet, they claimed that this emerged system yields a very cost-effective and dynamic process and is effective compared with the previously existing work for the current warehouse inventory management systems. Zhao et al. [22] introduced an intelligent warehouse management system (IWMS) that has sparked a lot of attention since it increases production and efficiency while reducing the number of human employees and mistakes. They discovered that a critical component of IWMS, along with warehouse portal intake and output registration, is now facing several problems. UHF-RFID readers will install throughout the portal. This intake and output registration will be carried out by calculating the precise RFID tags pasted on items. Ananthi et al. [23] discussed that most industries utilize the human resource-based inventory control process. They introduced an IoT-based inventory management system without the invention of human beings. A methodology based on RFID communication technology with the help of the Internet can design a principle to identify the stock details by the movement of the RFID tag attached to the goods. The Raspberry Pi module has a web service facility by running a GUI application.

By designing the approach collision resolving supported missing RFID tags, [24] introduced the RFID communication technology for identification tags within a certain period. In CR-MTI, several tags can react to different binary pattern-based collision slots. The reader may use bit tracking technology and a specially constructed

string to verify them together, substantially improving time efficiency. They investigate the best parameter choices for their proposed CR-MTI to enhance its performance. Xie et al. [25] proposed a tag collision identification mechanism. When tags collide, the reader cannot identify all colliding tags, significantly increasing the identification time. Gareis et al. [26] demonstrated a mobile robot platform with UHF-RFID capable of totally autonomous inventory taking by providing three-dimensional (3D) product maps and enabling the realization of the intelligent warehouse concept. Liu et al. [27] advocated utilizing RFID-enabled robots to tackle the problem of tag localization, which is essential for warehouse applications, including automated item retrieval and misplacement detection. They created an MRL system using off-the-shelf robots and RFID chips. When an RFID-enabled warehouse robot gets down a straight aisle, the reader uses two vertically deployed antennas ( $Z_1$  and  $Z_2$ ) to scan the target tag and sends tag phase data and timestamps to the server. They can also determine the relative location of the target tag about antenna  $Z_2$ 's trajectory. Finally, they estimate the target tag's position in 3D space using geometric connections between the target tag and antenna trajectories. They ran some tests to evaluate the MRL system's performance, and they were able to achieve high accuracy in both 2D and 3D localizations. Bu et al. [28] introduced RF-3DScan, a passive RFID-based system for 3D reconstruction, including package orientation and stacking on tagged products. With merely 2D scanning, RF-3DScan can determine fine-grained package stacking. They create an RF-3DScan prototype and put it through its faces in real-world settings. By using RF-3DScan, 92.5% accuracy and a rotation angle error of 4.08° on average were observed.

Cui et al. [29] used finite element analysis to assess the scissor's mechanism strength to optimize. Next, the AGV uses a multi-modal control mode to autonomously control and operate the entire vehicle using the robot operating system. The AGV uses magnetic stripe navigation, RFID site labels, ultrasonic sensors, and LiDAR as a safety system to achieve autonomous driving. Finally, the entire system was created and each module was tested. The results demonstrate that the whole system complies with the design specifications and completes specific work assignments.

Mishra and Mohapatra [30] presented an IoT cloud architecture for passive RFID tag-based real time. Stock keeping units (SKUs) tracking and machine learning algorithms were applied for predictive stock analysis in different supply chain segments such as inventory management. RFID tags are attached to SKUs, read at warehouse entry and exit, and real-time data are transmitted to a cloud server over the Internet. The data was analyzed using a machine learning-based software engine that utilizes categorization, training, and testing procedures. Gareis et al. [31] analyzed cost-effective and trustworthy technology. They used UHF-RFID for IoT applications. UHF-RFID may help to achieve 3D product maps for intelligent warehouses. They have examined synthetic aperture radar (SAR) trajectories in 2D and 1D. Bernardini et al. [32] present a robot used to move the reader antenna. The researchers look into two distinct estimation approaches. Gareis et al. [33] proposed a particle filter-based real-time localization estimation technique. The two approaches are labeling tags on objects and monitoring tag responses on created trajectories. Item locations are not

available online. The tags' location is offline after the entire SAR path has flowed, and all the data are available. They show how real-time signal processing influences tag counts and readings after evaluating the impact of trajectories depending on the lateral and radial resolution. Wang et al. [34] developed an RFID-based localization system. Due to the inability to distinguish distinct individuals, extreme object occlusion, and other reasons, present indoor localization technologies cannot satisfy the needs of situations such as retail shops, warehouses, and libraries, ShopSense. Based on the widespread use of RFID tags in scenarios like monitoring and analyzing human behavior and providing customized services. Zhang et al. [35] introduced an automatic collection of kiwi fruits to the cold storage system. It is critical to keep track of kiwi cold storage data. Cameron [36] discusses the characteristics unique to the ESP32 microcontroller regarding capabilities available to Bluetooth and BLE connections.

Sinodakis et al. [37] described an AS/RS for RFID system recognition to govern logistical processes. Two stepper motors for the  $x$ - and  $y$ -axes and a DC motor with a speed reducer for the  $z$ -axis for prototyping the three-axis AS/RS system. Irma et al. [38] discussed the technological development of automation in the industry and called the standard Industry 4.0 along with M-2-M communication. Szabó et al. [39] introduced a robotic arm control with color recognition built with a Raspberry Pi module. Colored bottle stoppers connected to the robotic arm's joints. Güleç and Orhun [40] proposed controlling a robot with a Raspberry Pi and an android-based application. Wi-Fi connects the android phone to the Raspberry Pi board and serves as the robot's command. The signals created by the android app were Python programming language on the Raspberry Pi.

Ananthi et al. [41] used RFID communication technology and IoT to develop stock inventory systems. The IoT-based industrial inventory control system has evolved to track the inventories linked to the card and stock data, dates, and hours for extra validation. Gündoğan et al. [42] investigated how networking protocols affect industrial communication services. It compares and contrasts the old MQTT-SN with the current IETF recommendation, the CoAP, and new ready-to-use ICN techniques. In multihop circumstances, publish–subscribe systems are more adaptable, whereas ICN approaches are more resilient. Al-Masri et al. [43] offered a thorough examination of existing communications protocols employed in the deployment of IoT devices. They discussed the protocols' unique methods in various IoT contexts and the problems. In the context of the IoT, the merits and drawbacks of different communications protocols were explored in this work. Kumar et al. [44] presented the study of existing IoT protocols and their applications to assist researchers and developers in selecting acceptable protocols for use in an IoT-based ecosystem. The architecture, its benefits, drawbacks, and application of various protocols in terms of dependability, security, interoperability, and power restrictions, among other things were explored in this paper.

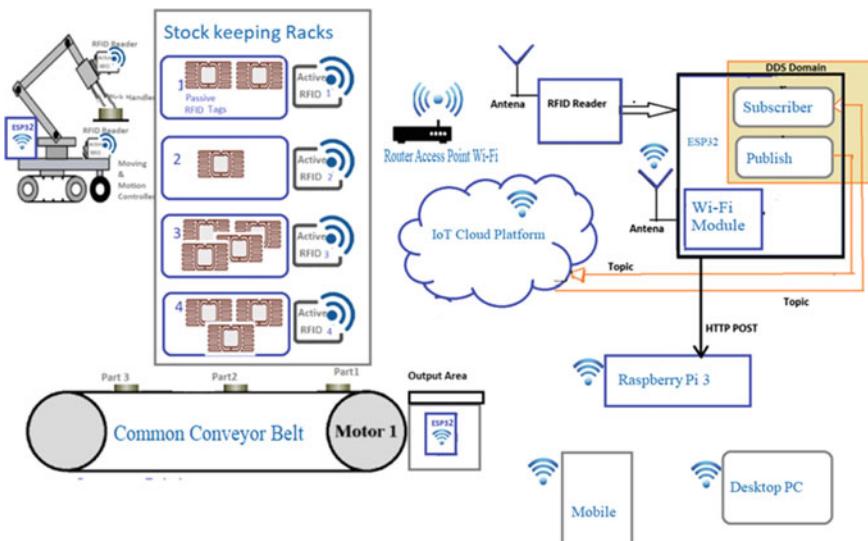
In this research, Abdulghani et al. [45] discovered security and privacy issues in several IoT protocols. In addition to assessing suggested solutions to specific problems, IoT has made our lives easier by fusing the real and virtual worlds. As a result, IoT is D2D connectivity that monitors and responds to environmental changes.

Abels et al. [46] used data distribution service to provide a platform for linked objects that incorporates composable semantics, security, reliability, and QoS. DDS's security, reliability, and QoS, high-level semantic improvements for backward compatible semantics. Finally, they recommended that more research into out-of-the-box composability and compatibility between standard IoT data models and compliance solutions is needed.

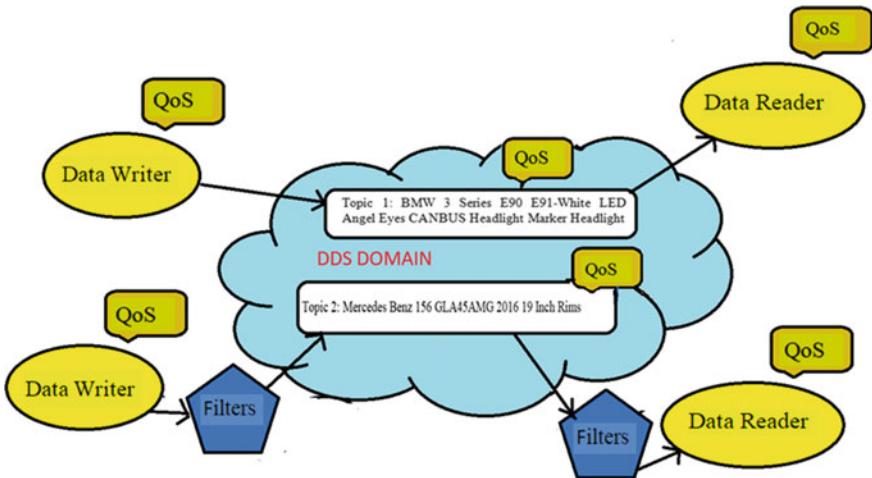
### 3 Proposed System

The publish–subscribe pattern integrates the RFID tag [47] in the IoT in our approach. DDS [48] is a middleware protocol, latency data communication, exceptional reliability, sophisticated security, and scalable architecture. The suggested system architecture for automated restoration and retrieval of spare parts utilizing RFID tags interfaced with IoT using DSS protocol and microcontroller nodes is illustrated in Fig. 1, with the OMG API standard for data-centric connectivity.

Middleware is the software layer interacting between the operating system and the applications in a distributed system. It allows multiple components to communicate and share data more efficiently. In addition, it makes distributed system development more accessible by enabling software developers to focus on the specific goal of their applications rather than the mechanics of data flow across apps. The application framework organizes transport and creates low-level data designs using the DDS



**Fig. 1** Typical system architecture for automatic restore and retrieve spare parts management process using IoT



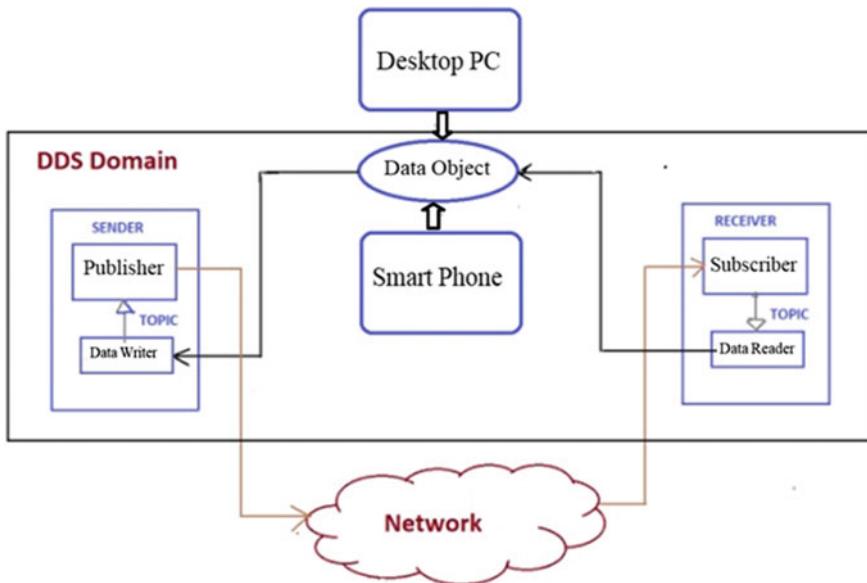
**Fig. 2** Data centricity in DDS domain with IoT

middleware. Different programming languages provide the same concepts and APIs, allowing programs to interchange data across operating systems, languages, and processor architectures [49].

There are several standards and solutions for communications middleware. As seen in Fig. 2, DDS is data-driven, making it ideal for IoT. The majority of middleware's job is to transport data between different applications and systems. Message-centric middleware requires programmers to build a code that transmits messages. Data-centric middleware [50] allows programmers to create a code that describes how and when data should be shared and then share data directly.

DDS's submit–subscribe approach enables scalable, real-time, reliable, high-overall performance, and interoperable data updates. In addition, DDS enables dynamic system augmentation orchestration and a variety of platforms, low-footprint devices to the cloud [51], and DLRL. The DLRL layer interfaces DCPs activities, allowing dispersed data exchange across IoT-enabled items. The DCPs layer provides facts to subscribers, whereas the DLRL layer displays DCPs features. A typical DDS domain architecture is in Fig. 3.

RFID is a technology that uses electromagnetic fields to identify and track tags attached to objects. An RFID system consists of a radio transponder, a radio receiver, and a transmitter. When triggered by an electromagnetic interrogation pulse from a nearby RFID reader device, the tag communicates digital data, usually an identifying inventory number, back to the reader. Active RFID tags [52] and passive RFID tags [53] are the two RFID tags. (1) Passive tags get energy from the RFID reader's probing radio waves. (2) Active tags are battery-powered to read from hundreds of meters distant RFID scanner. In addition, unlike a barcode, the tag does not need to be visible to the reader. RFID stands for radio frequency identification and data gathering (AIDC).



**Fig. 3** Typical architecture for DDS domain

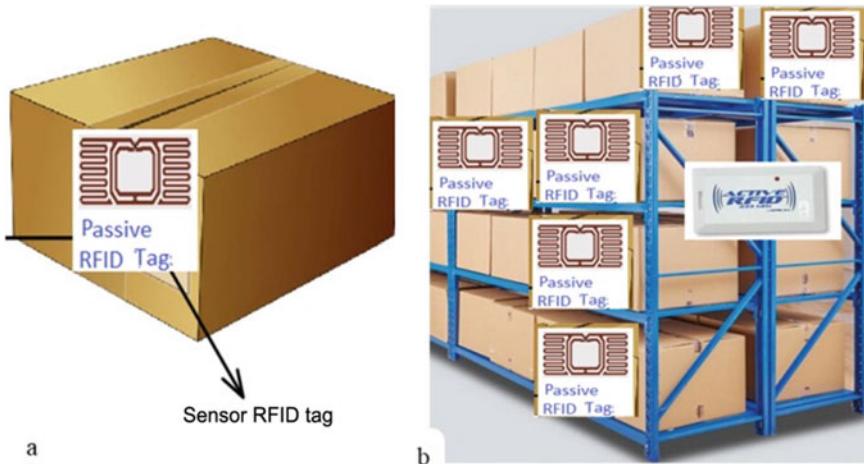
## 4 Methodology

Spare part inventory for maintenance organizations with automatization using IoT as automatic restore and delivery of spare parts is very helpful to improve their sales. Moreover, it will reduce the time for professionals and customers to get a particular spare part.

Spare parts are attached with a passive RFID sticker (which should not have any battery power, under a specific range of scope of RFID readers). Identical spare parts within the particular rack are related to stock keeping unit (SKU). Each SKU is placed in the frame and tagged with an active RFID tag (it contains embedded battery power. This tag can be identified from a long distance as nearly 100 m of space around).

Whenever a spare part is requested or restored to the rack using a smartphone, they will forward the request to the server. For example, the Raspberry Pi 3 [54] module shown in Fig. 4 will use the DDS protocol and the ESP32 microcontroller to broadcast this spare part's active RFID tag as a subject to the cloud platform.

The Esp32 [55] connects to the Raspberry Pi 3 through the HTTP POST mechanism. If we wish to restore components to a rack, the active RFID tag will publish to the cloud platform. Using the DDS protocol, the subscriber retrieves the active RFID tag as a topic from the cloud, while the RFID reader searches for the active RFID tag. The data will update in the cloud and send a message to the automatic robotic handling picker, which an ESP32 controller controls.

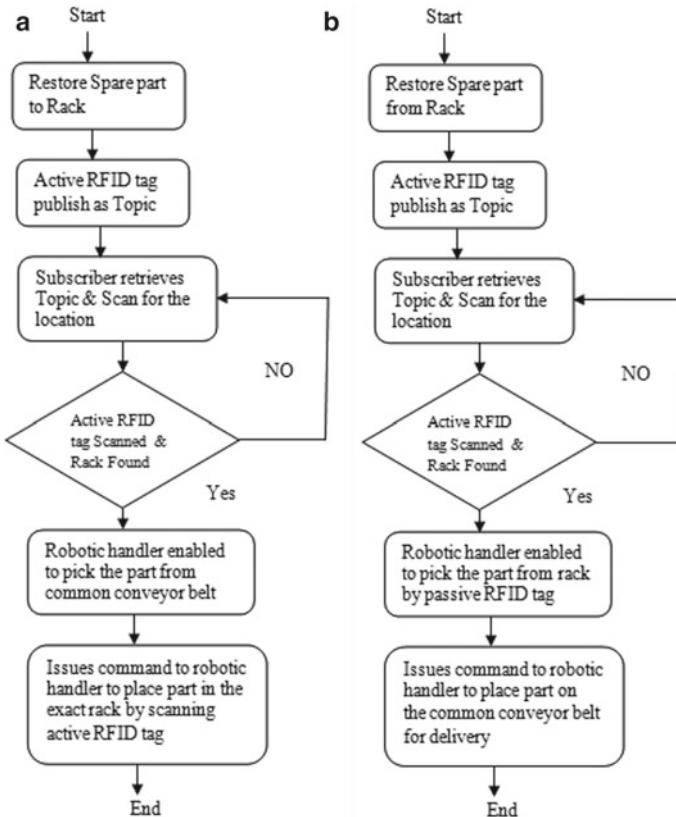


**Fig. 4** **a** Passive RFID tag on a spare component package and **b** spare parts rack with active RFID tag

It instructs the moving mechanism to pick the spare part placed on a standard conveyor belt and move it to the same rack to restore it using the active RFID tag. Figure 5a to retrieve the parts from the shelf in the opposite direction. In Fig. 5b, the data space in DDS arrange to be a local data store termed the global data space shown in Table 1.

The global data space appears as native memory space, and an API can access that data memory, and it seems to be local storage. On the other hand, DDS sends messages to remote nodes to update the relevant stores, regarded as a local stores. It uses dependability, system health (liveliness), and even security to customize data transmission behavior. Not every end-point in a simple system needs every item in your local store. This data space for this application is shown in Table 1. First, an RFID tag is used to the SKU of the product rack and a passive RFID tag is used for each spare part which is then placed in the appropriate frame to recognize the active RFID tag for the spare part, as shown in Fig. 6.

For example, in the first three rows of Table 1, there is an active RFID tag: 1111 1111 1111 0001, which is a binary string that indicates the product description as “BMW 3 Series E90 E91-White LED Angel CANBUS Headlight Marker Headlight” with S. No. 4430, and a passive RFID tag: 1000 0000 0000 0000 0001, as well as two more spare parts available with the same SKU: RS944RUR with product S. No along with a “1” status flag. It signifies that spare components are available, and null implies that they are not.



**Fig. 5** **a** Flowchart for replacing a spare item in the rack and **b** retrieve spare part from

## 5 Implementation

The warehouse automatically retrieves and restores spare parts automatically to strengthen the inventory management system. The most significant application for bright things with intelligence as IoT and communication technologies like RFID is helpful for processing controller units like ESP32 and Raspberry Pi 3 module. Here active and passive RFID stickers are used to relocate the rack and specific spare parts by using the mobile app or desktop pc dashboard.

ESP32 modules are programmed by Arduino C-program [56] to control the RFID tag readers and automatic conveyor belt. The protocol DDS uses a publish–subscribe model for the messages passing mechanism of the robotic picking handler. MIT app inventor [57] software is used to design mobile app to interface the web server and Raspberry Pi 3. Here Raspberry Pi 3 is programmed by Python [58] script. The active RFID readers are the most important for this system implementation because they can read the active RFID tags. The spare part rack and the passive

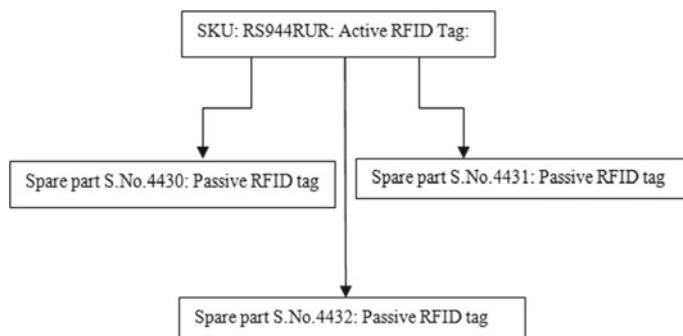
**Table 1** Products referrals SKU—specific rack data

Products referrals SKU-specific rack	Each spare part is identified by its product serial number				
Active RFID tag ID	Passive RFID tag ID	Stock keeping unit (SKU)	Product S. No.	Description	Status
1111 1111 1111 1111 0001	1000 0000 0000 0000 0001	KS944RUR	4430	BMW 3 seriesE90 E91-white LED angel CANBUS headlight marker headlight	1
1111 1111 1111 1111 0001	1001 0000 0000 0000 0101	KS944RUR	4431	BMW 3 series E90 E91-white LED angel CANBUS headlight marker headlight	1
1111 1111 1111 1111 0001	1010 0000 0000 0000 1001	KS944RUR	4432	BMW 3 series E90 E91-white LED angel CANBUS headlight marker headlight	1
1111 1111 1111 1111 0010	1011 0000 0000 0000 1101	RS944RUR	2315	Mercedes Benx W156 GLA45AMG 2016 19 inch-rims	1
1111 1111 1111 1111 0010	1100 0000 0000 0000 0101	RS944RUR	2316	Mercedes Benx W156 GLA45AMG 2016 19 inch-rims	1
1111 1111 1111 1111 0010	1110 0000 0000 0000 1001	RS944RUR	2317	Mercedes Benx W156 GLA45AMG 2016 19 inch-rims	1

(continued)

**Table 1** (continued)

Products referrals SKU-specific rack	Each spare part is identified by its product serial number				
Active RFID tag ID	Passive RFID tag ID	Stock keeping unit (SKU)	Product S. No.	Description	Status
1111 1111 1111 1111 0011	1001 0000 0000 0000 1101	PS944RUU	5566	Philips 12459SPC1 rally H4 headlight bulb (130/100 W)	1
1111 1111 1111 1111 0011	1011 0000 0000 0000 0101	PS944RUU	5567	Philips 12459SPC1 rally H4 headlight bulb (130/100 W)	1
1111 1111 1111 1111 0100	1011 0000 0000 0000 1001	MS944RUK	2310	PHILIPS XtremeVision G-force car headlight bulb (12 V, 55 W)	1
1111 1111 1111 1110 0011	NULL	SS944RUR	NULL	Philips WhiteVision Ultra H4 car headlight bulb, 4.200 K,	0
1111 1111 1111 0111 0100	NULL	KS944RUT	NULL	Philips H4 4300 k Car Crystal vision headlight bulbs	0

**Fig. 6** Typical tree structure for SKU with spare parts serial numbers. With the appropriate RFID tags

RFID readers are attached to the automatic robotic handler, which can read the specific spare part by reading the passive RFID tag. LF-RFID, HF-RFID, and UHF-RFID tags are available for LF-RFID [59]. Although there is less radio interference at low frequencies, the read time is also longer. A frequency range of 30–300 kHz is considered low frequency. HF-RFID has a frequency range of 3–30 MHz. The reading distance ranges from 10 cm to 1 m. The bulk of HF-RFID [60] devices operates at 13.56 MHz, which is relatively immune to radio interference [61]. They can work with read-only, write-only, and rewritable RFID tags. Up to 20 HF tags can be processed simultaneously by readers with memory capacities ranging from 64 bytes to 8 KB. The most sensitive to interference is UHF-RFID tags, which operate frequencies ranging from 300 MHz to 3 GHz.

## 6 Conclusion

Maintenance in organizations like inventory management is challenging to fulfill requirements in an acceptable period. Inaccurate machine spare parts inventories will shut down production equipment just as soon as not have the things needed to make finished products. Thus, we proposed an intelligent and automatic auto spare parts inventory organization system that can automatically store and retrieve the spare parts without human invention with accuracy and optimum time with the help of IoT with DDS protocol. It extends the system more efficiently and accurately for everyone that works with or purchases the spare parts that keep production online. The latest development in the delay releases the spare parts from the customs office, affecting timely services and delivery to consumers. There have been substantial delays in operating activity and product delivery. They may obtain items and components in various methods, and such a slowdown would encourage manufactured products and parts, a painful strand for a marketplace and taxpaying participants. This customs delay is unavoidable, but this proposed methodology can reduce the uncertainty in delivering the spare parts using this automated restore and retrieving an excess part process instead of the traditional manual process with human-operated service. This system ensures that the correct maintenance parts are in the storeroom when needed, gives an update on the components in stock and remaining needs to be arranged.

## References

1. Wu, H., et al.: The implementation of wireless industrial internet of things (IIoT) based upon IEEE 802.15.4-2015 TSCH access mode. In: 2019 IEEE Intl Conference on Dependable, Autonomic and Secure Computing, pp. 367–36 (2019). <https://doi.org/10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00075>
2. Srinivasan, K., Azhaguramya, V.R.: Internet of things (IoT) based object recognition technologies. In: 2019 Third International Conference on I-SMAC (IoT in Social, Mobile, Analytics

- and Cloud) (I-SMAC), pp. 216–220 (2019). <https://doi.org/10.1109/I-SMAC47947.2019.9032689>
- 3. He, M., Wei, X.: The model research of information automation system based on RFID in logistics business enterprise of warehouse. In: 2009 IEEE International Conference on Automation and Logistics, pp. 1727–1731 (2009). <https://doi.org/10.1109/ICAL.2009.5262661>
  - 4. Berenyi, Z., Charaf, H.: Retrieving frequent walks from tracking data in RFID-equipped warehouses. In: Conference on Human System Interactions, pp. 663–667 (2008)
  - 5. Zhang, L., Gao, F., Chen, F.: Based on set pair analysis of spare parts supply accurate security scheme decision. In: 2013 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering (QR2MSE), pp. 1253–1255 (2013). <https://doi.org/10.1109/QR2MSE.2013.6625796>
  - 6. Guirong, Z., Yuxin, M.: Study on auto enterprise inventory management. In: 2011 International Conference on Information Management, Innovation Management and Industrial Engineering, pp. 181–184 (2011). <https://doi.org/10.1109/ICIII.2011.191>
  - 7. Zhang, S., Huang, K., Yuan, Y.: Spare parts inventory management: a literature review. Sustainability (2021). <https://doi.org/10.3390/su13052460>
  - 8. Watts, F.B.: Bills of material. In: Engineering Documentation Control/Configuration Management Standards Manual. Wiley, Hoboken (2018)
  - 9. SchaubKen, J.L., Biery, D.: Physical security measures. In: The Ultimate Computer Security Survey (1995). <https://doi.org/10.1016/B978-0-7506-9692-0.50031-6>
  - 10. Krishna, A., Pendkar, N., Kasar, S., Mahind, U., Desai, S.: Advanced video surveillance system. In: 2021 3rd International Conference on Signal Processing and Communication (ICSPC), pp. 558–561 (2021). <https://doi.org/10.1109/ICSPC51351.2021.9451694>
  - 11. Qiyun, H.: ERP-based data warehouse model design. In: 2010 2nd International Conference on Computer Engineering and Technology, vol. 4, pp. 698–702 (2010). <https://doi.org/10.1109/ICCET.2010.5485323>
  - 12. Shen, L., Duan, W., Ren, Y., Yang, C.: Management service on WMS/WFS services. In: 2010 18th International Conference on Geoinformatics, pp. 1–5 (2010). <https://doi.org/10.1109/GEOINFORMATICS.2010.5567481>
  - 13. Jiang, C.: An integrated use of spreadsheets software in logistics education. In: 2009 4th International Conference on Computer Science and Education, pp. 1580–1584 (2009). <https://doi.org/10.1109/ICCSE.2009.5228302>
  - 14. Wang, F., Ng, H.Y., Ng, T.E.: Novel SKU classification approach for autonomous inventory planning. In: 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), pp. 1441–1445 (2018). <https://doi.org/10.1109/IEEM.2018.8607736>
  - 15. Aryal, G., Mapa, L., Camsarapalli, S.K.: Effect of variables and their interactions on RFID tag readability on a conveyor belt—Factorial analysis approach. In: 2010 IEEE International Conference on Electro/Information Technology, pp. 1–6 (2010). <https://doi.org/10.1109/EIT.2010.5612175>
  - 16. Yap, H.K., Ang, B.W.K., Lim, J.H., Goh, J.C.H., Yeow, C.: A fabric-regulated soft robotic glove with user intent detection using EMG and RFID for hand assistive application. In: 2016 IEEE International Conference on Robotics and Automation (ICRA), pp. 3537–3542 (2016). <https://doi.org/10.1109/ICRA.2016.7487535>
  - 17. Zhao, Y., Liu, X., Li, L., Zhao, X.: RF signal fluctuation based passive UHF RFID tag localization. In: 2021 International Conference on Communications, Information System and Computer Engineering (CISCE), pp. 168–172 (2021). <https://doi.org/10.1109/CISCE52179.2021.9445992>
  - 18. Gatial, E., Balogh, Z., Hluchý, L.: Concept of energy efficient ESP32 chip for industrial wireless sensor network. In: 2020 IEEE 24th International Conference on Intelligent Engineering Systems (INES), pp. 179–184 (2020). <https://doi.org/10.1109/INES49302.2020.9147189>
  - 19. Fox, J., Donnellan, A., Doumen, L.: The deployment of an IoT network infrastructure, as a localised regional service. In: 2019 IEEE 5th World Forum on Internet of Things (WF-IoT), pp. 319–324 (2019). <https://doi.org/10.1109/WF-IoT.2019.8767188>

20. Yoon, G., Choi, J., Park, H., Choi, H.: Topic naming service for DDS. In: 2016 International Conference on Information Networking (ICOIN), pp. 378–381 (2016). <https://doi.org/10.1109/ICOIN.2016.7427138>
21. Tejesh, B.S.S., Neeraja, S.: Warehouse inventory management system using IoT and open source framework. Alexandria Eng. J. **57**(4) 3817–3823 (2018). ISSN 1110–0168
22. Zhao, Y., Li, L., Zhao, X., Liu, X.: UHF RFID based warehouse portal in-out registration method. In: 2021 International Conference on Communications, Information System and Computer Engineering (CISCE), pp. 514–517 (2021). <https://doi.org/10.1109/CISCE52179.2021.9445950>
23. Ananthi, K., Rajavel, R., Sabarikannan, S., Srisaran, A., Sridhar, C.: Design and fabrication of IoT based inventory control system. In: 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)2021, pp. 1101–1104 (2021). <https://doi.org/10.1109/ICACCS51430.2021.9441701>
24. Su, J., Sheng, Z., Liu, A.X., Fu, Z., Huang, C.: An efficient missing tag identification approach in RFID collisions. IEEE Trans. Mob. Comput. <https://doi.org/10.1109/TMC.2021.3085820>
25. Xie, X., Liu, X., Qi, H., Guo, S., Li, L.: A tag-correlation-based approach to fast identification of group tags. IEEE Trans. Mob. Comput. <https://doi.org/10.1109/TMC.2021.3052572>
26. Gareis, M., et al.: Novel UHF-RFID listener hardware architecture and system concept for a mobile robot based MIMO SAR RFID localization. IEEE Access **9**, 497–510 (2021). <https://doi.org/10.1109/ACCESS.2020.3047122>
27. Liu, X., et al.: Accurate localization of tagged objects using mobile RFID-augmented robots. IEEE Trans. Mob. Comput. **20**(4) 1273–12841. <https://doi.org/10.1109/TMC.2019.2962129>
28. Bu, Y., et al.: RF-3DScan: RFID-based 3D reconstruction on tagged packages. IEEE Trans. Mob. Comput. **20**(2) 722–738 (2021). <https://doi.org/10.1109/TMC.2019.2943853>
29. Cui, Z., Xu, H., Chen, Z., Yang, H., Huang, S., Gong, M.: Design of a novel AGV with automatic pick-and-place system based on scissor lifting platform. Chin. Autom. Congr. (CAC) **2020**, 4435–4440 (2020). <https://doi.org/10.1109/CAC51589.2020.9327003>
30. Mishra, A., Mohapatro, M.: Real-time RFID-based item tracking using IoT & efficient inventory management using machine learning. In: 2020 IEEE 4th Conference on Information & Communication Technology (CICT), pp. 1–6 (2020). <https://doi.org/10.1109/CICT51604.2020.9312074>
31. Gareis, M., Carlowitz, C., Vossiek, M.: A MIMO UHF-RFID SAR 3D locating system for autonomous inventory robots. In: 2020 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM), pp. 1–4 (2020). <https://doi.org/10.1109/ICMIM48759.2020.9298989>
32. Bernardini, F., Motroni, A., Nepa, P., Buffi, A., Tellini, B.: SAR-based localization of UHF-RFID tags in smart warehouses. In: 2020 5th International Conference on Smart and Sustainable Technologies (SpliTech), pp. 1–6 (2020). <https://doi.org/10.23919/SpliTech49282.2020.9243755>
33. Gareis, M., Fenske, P., Carlowitz, C., Vossiek, M.: Particle filter-based SAR approach and trajectory optimization for real-time 3D UHF-RFID tag localization. IEEE Int. Conf. RFID (RFID) **2020**, 1–8 (2020). <https://doi.org/10.1109/RFID49298.2020.9244917>
34. Wang, P., Guo, B., Wang, Z., Yu, Z.: ShopSense:customer localization in multi-person scenario with passive RFID tags. IEEE Trans. Mob. Comput. (2020). <https://doi.org/10.1109/TMC.2020.3029833>
35. Zhang, T., Cao, C., Liu, Y., Yu, H.: Design and implementation of kiwi fruit cold storage management system based on RFID. Int. Wirel. Commun. Mob. Comput. (IWCNC) **2020**, 2194–2198 (2020). <https://doi.org/10.1109/IWCNC48107.2020.9148540>
36. Cameron, N.: ESP32 microcontroller features. In: Electronics Projects with the ESP8266 and ESP32 (2021). [https://doi.org/10.1007/978-1-4842-6336-5\\_22](https://doi.org/10.1007/978-1-4842-6336-5_22)
37. Sinidakis, G., Kostopoulos, V., Drakaki, M., Karnavas, Y.L., Tzionas, P.: An RFID-enabled automated storage and retrieval system via microcontroller stepper motor control. In: 2019 8th International Conference on Modern Circuits and Systems Technologies (MOCAST), pp. 1–4 (2019). <https://doi.org/10.1109/MOCAST.2019.8742069>

38. Ulices, C., Irma, M., Carlos, J.: Design of an electronic coupling to control speed of motor on a conveyor belt using IoT. In: 2020 International Conference on Mechatronics, Electronics and Automotive Engineering (ICMEAE), pp. 115–121 (2020). <https://doi.org/10.1109/ICMEAE51770.2020.00027>
39. Szabó, R., Gontean, A., Sfirač, A.: Robotic arm control in space with color recognition using a Raspberry Pi. In: 2016 39th International Conference on Telecommunications and Signal Processing (TSP), pp. 689–692 (2016). <https://doi.org/10.1109/TSP.2016.7760972>
40. Güleç, M., Orhun, M.: Android based WI-FI controlled robot using Raspberry Pi. In: 2017 International Conference on Computer Science and Engineering (UBMK), pp. 978–982 (2017). <https://doi.org/10.1109/UBMK.2017.8093402>
41. Ananthi, K., Rajavel, R., Sabarikannan, S., Srisaran, A., Sridhar, C.: Design and fabrication of IoT based inventory control system. In: 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 1101–1104 (2021). <https://doi.org/10.1109/ICACCS51430.2021.9441701>
42. Gündoğan, C., et al.: The impact of networking protocols on massive M2M communication in the industrial IoT. IEEE Trans. Netw. Serv. Manag. (2021). <https://doi.org/10.1109/TNSM.2021.3089549>
43. Al-Masri, E., et al.: Investigating messaging protocols for the internet of things (IoT). IEEE Access **8**, 94880–94911 (2020). <https://doi.org/10.1109/ACCESS.2020.2993363>
44. Kumar, N.V.R., Kumar, P.M.: Survey on state of art iot protocols and applications. In: 2020 International Conference on Computational Intelligence for Smart Power System and Sustainable Energy (CISPSSE), pp. 1–3 (2020). <https://doi.org/10.1109/CISPSSE49931.2020.9212227>
45. Abdulghani, R.M., Alrehili, M.M., Almuhanne, A.A., Alhazmi, O.H.: Vulnerabilities and security issues in iot protocols. In: 2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH), pp. 7–12 (2020). <https://doi.org/10.1109/SMARTTECH49988.2020.00020>
46. Abels, T., Khanna, R., Midkiff, K.: Future proof IoT: composable semantics, security, QoS and reliability. In: 2017 IEEE Topical Conference on Wireless Sensors and Sensor Networks (WiSNet), pp. 1–4 (2017). <https://doi.org/10.1109/WISNET.2017.7878740>
47. Makhijani, K.M., Maradia, K.G.: RFID tag impedance matching techniques: a survey. In: 2015 International Conference on Computational Intelligence and Communication Networks (CICN), pp. 44–45 (2015). <https://doi.org/10.1109/CICN.2015.17>
48. Schlesselman, J.M., Pardo-Castellote, G., Farabaugh, B.: OMG data-distribution service (DDS): architectural update. In: IEEE MILCOM 2004. Military Communications Conference, vol. 2, pp. 961–967 (2004). <https://doi.org/10.1109/MILCOM.2004.1494965>
49. Meng, Y., Xingmin, W., Wei, S.Z.: Design and implementation of emergency qos of publish-subscribe middleware based on DDS. In: 2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE), pp. 2024–2027 (2019). <https://doi.org/10.1109/EITCE47263.2019.9095082>
50. Valls, G., Basanta Val, P.: Usage of DDS data-centric middleware for remote monitoring and control laboratories. IEEE Trans. Ind. Inform. **9**(1) 567–574 (2000). <https://doi.org/10.1109/TII.2012.2211028>
51. Yuefeng, H.: Study on data transmission of DCPS publish-subscribe model. In: 2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), pp. 1–2172 (2018). <https://doi.org/10.1109/IMCEC.2018.8469351>
52. Zhang, T., Chen, Z., Ouyang, Y., Hao, J., Xiong, Z.: An improved RFID-based locating algorithm by eliminating diversity of active tags for indoor environment. Comput. J. **52**(8), 902–909 (2009). <https://doi.org/10.1093/comjnl/bxn039>
53. Santiago, A.G., Fernandes, C.A., Costa, J.R.: Broadband UHF RFID passive tag antenna for near-body operation. In: 2012 IEEE International Conference on RFID-Technologies and Applications (RFID-TA), pp. 271–274 (2012). <https://doi.org/10.1109/RFID-TA.2012.6404528>
54. Yamanoor, N.S., Yamanoor, S.: High quality, low cost education with the Raspberry Pi. In: 2017 IEEE Global Humanitarian Technology Conference (GHTC), pp. 1–5 (2017). <https://doi.org/10.1109/GHTC.2017.8239274>

55. Maier, A., Sharp, A., Vagapov, Y.: Comparative analysis and practical implementation of the ESP32 microcontroller module for the internet of things. In: 2017 Internet Technologies and Applications (ITA), pp. 143–148 (2017). <https://doi.org/10.1109/ITECHA.2017.8101926>
56. <https://docs.espressif.com/projects/esp-idf/en/latest/esp32/resources.html>
57. Appinventor : <https://appinventor.mit.edu/>
58. Raspberry Pi documentation : <https://www.raspberrypi.org/documentation/usage/python/>
59. Liubavin, K.D., Ermakov, I.V., Losevskoy, A.Y., Nuykin, A.V., Strakhov, A.S.: Low-power digital part design for a LF RFID tag in a double-poly 180 nm CMOS process. In: 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), pp. 150–153 (2021). <https://doi.org/10.1109/ElConRus51938.2021.9396585>
60. Farahani, H.S., Rezaee, B., Gadringer, M., Zöscher, L., Amtmann, F., Bösch, W.: An accurate de-embedding and characterization methodology for dual-band HF/UHF RFID chips and antennas. In: 2021 15th European Conference on Antennas and Propagation (EuCAP), pp. 1–5 (2021). <https://doi.org/10.23919/EuCAP51087.2021.9411164>
61. Wagih, M., Shi, J.: Wireless ice detection and monitoring using flexible UHF RFID tags. IEEE Sens. J. (2021). <https://doi.org/10.1109/JSEN.2021.3087326>

# Link Prediction in Paper Citation Network based on Deep Graph Convolutional Neural Network



Bui Thanh Hung

**Abstract** With regards to network structured data nowadays, link prediction stands out as a key problem. It has been observed that interactions between the entities act as the basic foundations of many applications in various fields such as chemistry, biology, or social networks. In the previous studies on this subject, most of the approaches used heuristics methods with an evaluation function to find the similarity index between entities, thereby predicting the possibility of links between them. However, heuristics methods are based on assumptions about the existence of associations between entities; thus its shortcoming is that when these assumptions are not accurate, the results of the algorithm would be reduced significantly. Taking those into considerations, in this paper, we propose a method to solve this problem based on deep graph convolutional neural network. Experimental results on paper citation network dataset have shown that particular and this method is promising.

**Keywords** Graph neural networks · Deep graph convolutional neural network · Link prediction · Paper citation network

## 1 Introduction

Generally, most of data in the real world exists in the form of links, such as the bonds between proteins in human cells. These links contain a lot of information such as entity properties, network topology, or network evolution.

Link prediction is the problem of predicting the existence of links between entities in a structured network over a period of time. These links fall into two main categories:

---

B. T. Hung (✉)

Faculty of Information Technology, Ton Duc Thang University, 19 Nguyen Huu Tho Street, Tan Phong Ward, District 7, Ho Chi Minh City, Vietnam

e-mail: [buithanhhung@tdtu.edu.vn](mailto:buithanhhung@tdtu.edu.vn)

- The link is lost (missing) in the event that the data have problems which needs to be corrected,
- New link in future (new) between two entities in the network.

Majority of applications in various fields such as biology, chemistry, medicine, materials science, or social networks use interactions between these entities. These can include recommending friends in social networks [1], recommending movies in Netflix [2], or predicting protein sequences [3]. Determining the existence of an association between these entities requires a considerable amount of experimental efforts and can even take an enormous amount of time. For that reason, instead of blindly testing all these links, it will be more effective to use the link prediction methods; then, the scientists and engineers could focus their concentration on the most likely links to save the testing cost significantly.

Over the past decade, it has been attracting attention from lots of research communities (as well as publications) on this subject ranging from algorithms, improvements, applications, challenges, and future development directions research. However, despite lots of in-depth studies as well as various publications, there has not been any method to predict the link which proves to be outstanding nor effective till now.

In this paper, with the advent of deep learning, we propose a method based on deep graph convolutional neural network for link prediction in paper citation network. The remaining of this paper is structured as follows: Related work is presented in Sect. 2, and Sect. 3 presents methodology; experiment of our approach is shown in Sect. 4, and Sect. 5 concludes our research as well as future direction.

## 2 Related Works

The link prediction problem can be described as follows: considering a network with the structure  $G = (V, E)$  where  $V$  is the set of graph vertices and  $E$  is the set of edges of the graph. From a set of  $V$  vertices and a subset of correct associations (links are certain to exist), the task of association prediction is to identify whether links between entities exist as observed or not.

In the past, several link predication researches have been proposed with the focus on heuristic methods. These methods are mainly based on a similarity evaluation algorithm, where each pair of vertices  $x$  and  $y$  will be assigned a score  $s_{xy}$  defined as the similarity, or similarity, between  $x$  and  $y$ . All unobserved links will be evaluated and ranked based on their  $s_{xy}$  scores, and those with connections with similar vertices will gain higher scores.

The similarity of vertices can be evaluated by the features of the vertices themselves. Two vertices are considered similar if they have many features in common [4]. However, the features of the vertices are often hidden, and thus, these methods focus on another similarity index that is structural similarity, based solely on the structure of the network. There are many similarity indices and are classified in different

ways such as local similarity that only considers the neighborhoods of the predicted vertices as a measure of similarity typically the common neighbors index (CN) [5] or the network-wide similarity index that considers the similarity of paths based on the structure of the entire network as a basis for finding similarity, typically the Katz, SimRank Index [6, 7].

We notice that this method mainly relies on evaluation indicators to determine the similarity scores between them. Although it might appear to be simple, the definition of similarity between nodes is not as simple at all. Indeed, the definition of similarity between nodes is a huge challenge if taking a closer deep-dive study. These similarity metrics can be very simple but can also turn out to be very complex: It might work well on one network but perform poorly on another. The CN index example assumes that two vertices are more likely to be connected if they have many common neighbors. This peak price appears to be correct in the friend recommendation network; however, in another different field of the protein network where two proteins with similar neighbors are less likely to bind [8], it appears to be incorrect. In conclusion, these methods remain to have several shortcomings.

Another approach is to use the method of using artificial neural networks, which shows more advantages in solving these network problems versus the above-mentioned ones. There have been a number of works applying graph structure to artificial neural networks such as recursive neural networks (RNNs) [9–14]; among those, the introduction of the graph neural network model (GNN) [15] has strongly promoted the research on the application of GNN model as an effective solution to link prediction problem. The GNN has emerged as a powerful model in learning graph structured data [15–17]. The main reason for this is because learning from a graph neural network will make learning the features of the nodes and learning the structure of the graph into a unified whole. In this way, GNN has shown outstanding performance in link prediction problems [18–21].

A graph neural network (GNN) is a general type of network for representations in the form of graphs. By representing the problem as a graph—encoding the information of individual elements as vertices and the relationship between them as edges—GNN learns patterns from the graph from which to predict the relationship between the vertices is unknown.

Therefore, all networks are represented as graphs, combining learning of the features of individual elements with their position in the graph will give more accurate results than other methods. Based on this advantage, we propose using GNN to solve this problem. What makes our research different with the others is that we integrate the deep graph convolutional neural network model into SEAL framework to solve link prediction problem. We apply this model in paper citation network domain and evaluate the proposed model with others including the SEAL models with GCN and SAGE.

### 3 Methodology

There are two main link prediction frameworks based on GNN which are:

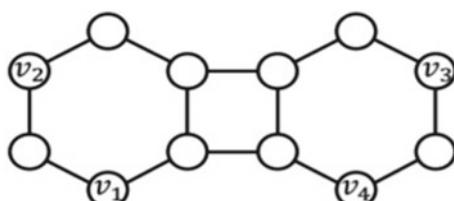
- Graph autoencoder (GAE): In this method, the first GNN will be applied to the entire network. Then, the information of the source and destination vertices will be aggregated to make a prediction about the link [18].
- SEAL: In this method, a local subgraph will be extracted around the link to be predicted. Then, the child vertices in this subgraph will be assigned different labels according to their distance to the source and destination vertices. Finally, GNN is applied to this subgraph to aggregate information and make predictions about the association [20].

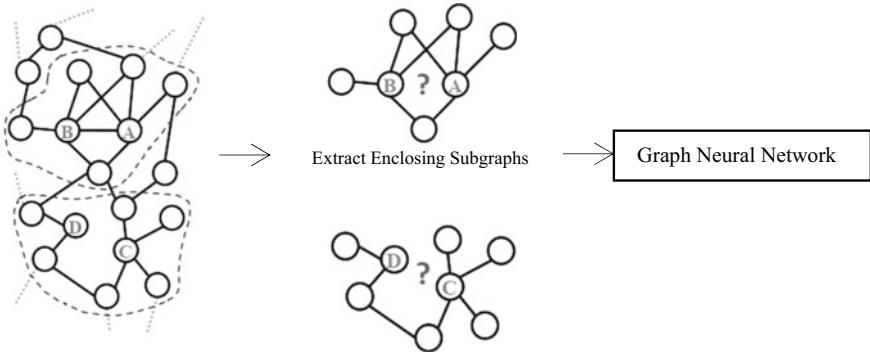
It could be observed that both GAE and SEAL works are quite similar when they both use GNN to learn the structures and features of vertices. However, in the association prediction problem, the SEAL method proves to be more superior to the GAE method [21] for two basic reasons:

- The SEAL method locally extracts a portion of the graph around the link to be predicted instead of having to learn the structure of the entire bearing. In the link prediction problem, it is seen that vertices located far away from the link to be predicted are often less influential than those located near it. Therefore, the subgraph extraction will save the training cost.
- The SEAL method incorporates labeling on the vertices in the subgraph before using GNN. This method is particularly effective because sometimes GNNs fail to distinguish between some structural and functional links different. From Fig. 1, it could be seen that the vertices  $(v_1, v_4)$  and  $(v_2, v_3)$  are isomorphic; so we can predict if the association  $(v_1, v_2)$  exists, then  $(v_3, v_4)$  will exist because they are isomorphic with each other. However, since 2 vertices  $(v_2, v_3)$  are isomorphic, GNN will also predict that the association  $(v_1, v_2)$  exists, then  $(v_1, v_3)$  will exist. Actually, this is not absolutely accurate; since it is obvious that  $v_1$  is much closer to  $v_2$  and shares more common neighbors than  $v_1$  and  $v_3$ . To solve this problem, SEAL proposed a labeling method based on the distance between vertices as an additional feature, so that GNN can learn more accurately.

Taking those into consideration, this paper proposes using SEAL model to solve the link prediction problem. Different with the original model, we integrate deep graph convolutional neural network model into training because of its advantages.

**Fig. 1** Example of GNN does not distinguish links





**Fig. 2** Architecture of SEAL [20]

### 3.1 *The Proposed Model*

The general model of the SEAL method to solve the problem is presented in Fig. 2. This model consists of three main parts:

- Represent the data as a graph and then extract the subgraphs around the association to be predicted.
- Building feature matrix of vertices.
- Training with GNN models.

The detail of this model is presented as follows:

### 3.2 *Extract Subgraphs*

The first step of the SEAL method is to extract the subgraphs around the vertices, including the vertices containing the link that the GNN needs to train to build the training dataset and the vertices containing the link that the GNN needs to predict to build the training dataset. Each pair of vertices can extract one or more subgraphs surrounding that pair of vertices for training. These subgraphs will be represented as adjacency matrix and treated as input of GNN.

### 3.3 *Constructing a Feature Matrix*

The second step in the SEAL method is to build a feature matrix of the vertices in the subgraph. This step is critical to train the GNN model to predict the association effectively. The feature information matrix in the SEAL method consists of three main components:

Label the vertices  
 Node embedding  
 Node attributes

### *Label the vertices*

The first component of the feature information matrix is the label of the vertices. Each label in the graph needs to be labeled, in which the labeling function is a function:  $f_1: V \rightarrow N$  will assign an integer label  $f_1(i)$  to the vertex  $i$ . The purpose of this is to mark the different roles of each vertex in the subgraph.

- The  $x$  and  $y$  vertices that we need to predict the connection are called central vertices.
- The vertices have different positions relative to the central position and have different structural importance than the predicted link.

Labeling the vertices plays an important role in training GNN. As mentioned in the previous section, if we do not assign labels, GNN will not be capable of distinguishing the importance of central nodes and may lose structural information when predicting link existence.

The SEAL method proposes the double-radius node labeling (DRNL) technique [20] as follows:

- The two central vertices  $x$  and  $y$  will always be labeled 1.
- Vertices  $i$  and  $j$  will have the same label if  $d(i, x) = d(j, x)$  and  $d(i, y) = d(j, y)$ . This criterion is because the position of the vertex inside the graph can be represented by its distance to the two central nodes, namely  $(d(i, x), d(j, x))$ . Therefore, we place the nodes in a trajectory with the same label from which we can reflect the relative position and structural importance of that vertex and the central vertices.

### *Node embedding and node attributes*

Besides the structural labels of the vertices, the feature information matrix also contains information about the implicit or explicit features of the vertices. By combining node embedding together with node attributes into the feature information matrix, GNN will learn more accurately.

SEAL proposes a method to generate node embedding and node attributes as follows: assuming we have a graph  $G(V, E)$ , and the set  $E_p$  is the sample set containing the vertices associated with  $E_p \subseteq E$ , and the set  $E_n$  is the sample set contains vertices with no connections where  $E \cap E_n = \emptyset$ . The proposed SEAL method is called as negative injection whereby instead of embedding nodes created directly from  $E$ , we will create the set  $E' = E \cup E_n$ . In this way, the feature information matrix will simultaneously have the same association existence information between the sample set containing connected vertices and the sample set containing unconnected vertices and will learn more accurately.

### 3.4 Training with Deep Graph Convolution Neural Network (DGCNN)

The DGCNN model is proposed by Zhang et al. [22]. The architecture of this model is shown in Fig. 3. This model includes three stages:

- Graph convolution layers extract features of subgraphs.
- Sort pooling layer sorts the characteristics of the previous vertices and synchronizes the size of the input.
- Traditional convolution layers and dense layers read the resulting graph representations and then switch back to the traditional CNN model to learn and make association predictions.

#### *Graph Convolution Layers*

Given a matrix  $A$  with data structure adjacency matrix and feature information matrix  $X$ , Graph convolution layer will have the following structure:

$$Z = f(\tilde{D}^{-1}\tilde{A}X\tilde{W}) \quad (1)$$

where

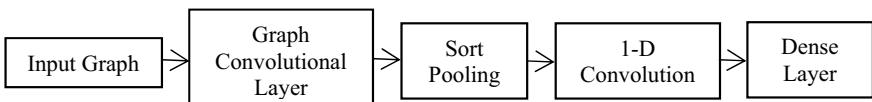
- $\tilde{D}$  is the order matrix.  
 $\tilde{A} = A + I$  ( $I$  is the unit matrix).  
 $W$  is the convolution matrix.  
 $F$  is the threshold function.

#### *Sort Pooling Layer*

DGCNN uses a sort pooling layer as a bridge between the graph convolution layers and the traditional convolution layers. This layer has the effect of rearranging the state of the last graph convolution layer into an ordered feature descriptor matrix. From this, it is possible to quickly identify isomorphic subgraphs of the graph. Then, use max- $k$  math to flatten it to one size before passing it through the traditional convolution layers.

#### *Traditional Convolution Layers*

The output of the sort pooling class is a  $Z$  tensor of size  $k \times \sum_1^h c_t$  where each row represents the number of vertices and each column represents the feature. To perform



**Fig. 3** Architecture of DGCNN [22]

CNN, firstly, it is necessary to normalize the dimension of tens  $Z$  to  $k \times \sum_1^h c_t \times 1$  vector. Then, it will perform 1D convolution layer convolution with size  $\sum_1^h c_t$  to apply filter sequentially to tensor  $Z$ . Then, some max pooling and 1D convolution layers will be added to understand local features. Finally, a fully connected class to aggregate results and make predictions.

## 4 Experiments

### 4.1 Dataset

We conducted experiments on the paper citation network dataset named CORA [23]. The CORA dataset includes 2708 scientific articles on machine learning classified into 7 classes. Where each vertex represents each paper and the edges represent citations between papers. Each article in the CORA dataset is described by a vector with values 0–1 representing the presence of the corresponding word in the dictionary. The dictionary includes 1433 unique words. The CORA dataset will be viewed as a structured network-oriented dataset.

The dataset will be divided into 8.5 parts for training the network, 0.5 parts to evaluate on the training set, and 1 part to test the results of the model.

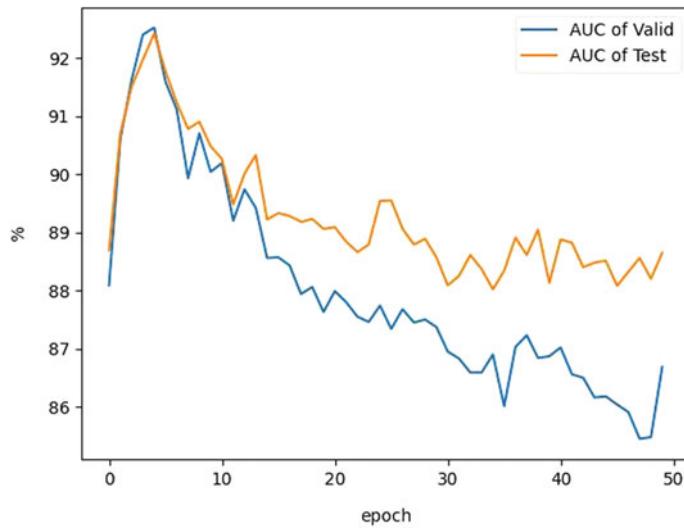
### 4.2 Results

We used the DGCNN model with 3 layers for the graph convolutions layer. With the traditional convolutions layer, we used two 1D convolution layers with 1 max pooling layer in between the two 1D convolution layers and finally, the fully connected layer using the sigmoid function. We used the binary cross entropy function, and optimization algorithm is Adam algorithm. We used the number of epochs of 50 and then selected the epochs with the best index to evaluate the results. The PyTorch\_geometric library is used to build all graph neural network models [24]. We did experiments using Python language, PyTorch library, and Google Colab.

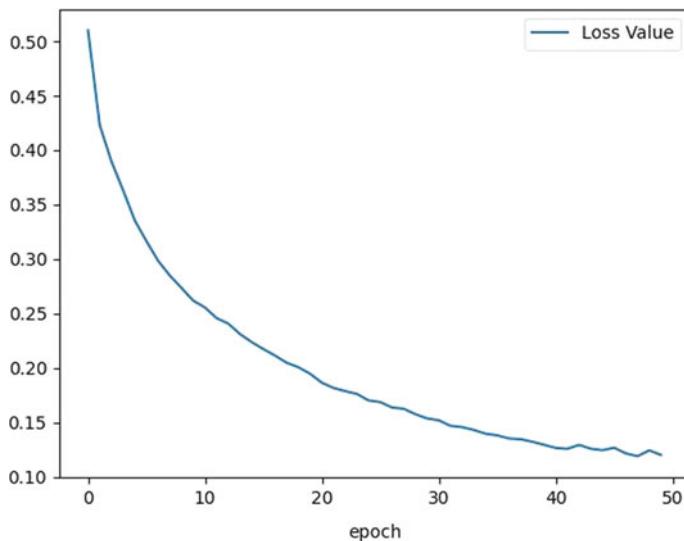
We used area under ROC curve (AUC) as a measure to evaluate the effectiveness of the proposed model. The AUC is an estimate of the probability that the classifier will rank the randomly selected positive case (where a link exists) over the negative case (no association exists). AUC has a value from 0 to 1. A completely wrong model has an AUC value of 0, and a completely correct model has an AUC value of 1.

We compared the results of the proposed model with various SEAL models (training with GCN and SAGE) and two heuristic methods: common neighbor (CN) and Adamic Adar (AA).

Figure 4 shows AUC results of valid and test datasets during training. The loss value in training process is shown in Fig. 5, and Table 1 compares the experimental



**Fig. 4** AUC of valid and test datasets during training



**Fig. 5** Loss value in training process

results of the proposed model with some heuristics methods and the SEAL with GCN and SAGE. As the result is shown in Table 1, the proposed model has much better performance than the SEAL models with GCN and SAGE as well as the heuristics methods.

**Table 1** The result

Model	AUC
SEAL with DGCNN	92.42
SEAL with GCN	90.25
SEAL with SAGE	91.14
Common neighbor (CN)	73.14
Adamic Adar (AA)	73.24

Italics indicates your proposed model's result

On a side note, it is also noticed that GNN models are generally sensitive to noisy data. If the data of the model are noisy (add/lose), edges can affect the entire training model. Therefore, if the training data are noisy due to malicious attacks or users, it can seriously affect the accuracy of the model. Most of the recent works are directed toward developing the power of GNN models. However, there are not many studies to develop the ability to represent vertices and edges in the GNN model. The SEAL method with only a simple labeling implementation has significantly improved the performance of the GNN model. And the proposed model, integrating DGCNN in the SEAL is better than the SEAL with GCN and SAGE.

## 5 Conclusion

With the aim to propose an effective method to solve the link prediction problem, in this paper, we used the SEAL model with DGCNN and did experiments on the CORA paper citation network dataset. We compared the proposed model with the SEAL with GCN and SAGE and two other heuristics algorithms. Experimental results show that the proposed model gives better results than the original SEAL and heuristics methods. In future, we expect to represent features of vertices and incorporate it into the GNN model and learn methods of denoising to preprocess data before entering GNN. We will try the combination of GNN models to find the optimal model for the link prediction problem.

## References

1. Adamic, L.A., Adar, E.: Friends and neighbors on the web. *Soc. Netw.* **25**(3), 211–230 (2003)
2. Bennett, J., Lanning, S., et al.: The netflix prize. In: Proceedings of KDD Cup and Workshop, vol. 2007, pp. 35 New York (2007)
3. Qi, Y., Bar-Joseph, Z., Klein-Seetharaman, J.: Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins Structure Funct. Bioinform.* **63**(3), 490–500 (2006)
4. Lin, D.: An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning. Morgan Kaufman, San Francisco (1998)

5. Barabasi, A.-L., Albert, R.: Emergence of scaling in random networks. *Science* **286**(5439), 509–512 (1999)
6. Katz, L.: A new status index derived from sociometric analysis. *Psychometrika* **18**, 39 (1953)
7. Lü, L., Zhou, T.: Link prediction in complex networks: a survey. *Phys. A* **390**(6), 1150–1170 (2011)
8. Kovács, I.A., Luck, K., Spirohn, K., Wang, Y., Pollis, C., Schlabach, S., Bian, W., Kim, D.K., Kishore, N., Hao, T., et al.: Network-based prediction of protein interactions. *BioRxiv*, 275529 (2018)
9. Frasconi, P., Gori, M., Sperduti, A.: A general framework for adaptive processing of data structures. *IEEE Trans. Neural Netw.* **9**(5), 768–786 (1998)
10. Hung, B.T.: Document classification by using hybrid deep learning approach. In: *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering—LNICST*, vol. 298, pp. 167–177 (2019)
11. Hung, B.T., Semwal, V.B., Gaud, N., Bijalwan, V.: Hybrid deep learning approach for aspect detection on reviews. In: *Proceedings of Integrated Intelligence Enable Networks and Computing*. Springer, Singapore (2021)
12. Hung, B.T.: Integrating diacritics restoration and question classification into Vietnamese question answering system. *Special Issue on Adv. Eng. Comput. Sci. J—ASTESJ*. **4**(5), 207–212 (2019)
13. Smys, S., Chen, J.I.Z., Shakya, S.: Survey on neural network architectures with deep learning. *J. Soft Comput. Paradigm (JSCP)* **2**(03), 186–194 (2020)
14. Mugunthan, S.R., Vijayakumar, T.: Design of improved version of sigmoidal function with biases for classification task in ELM domain. *J. Soft Comput. Paradigm (JSCP)* **3**(02), 70–82 (2021)
15. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. *IEEE Trans. Neural Netw.* **20**(1) (2009)
16. Bruna, J., Zaremba, W., Szlam, A., LeCun, Y.: Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv*, 1312–6203 (2013)
17. Li, Y., Tarlow, D., Brockschmidt, M., Zemel, R.: Gated graph sequence neural networks. *arXiv preprint arXiv*, 1511–05493 (2015)
18. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv*, 1609–02907 (2016)
19. Chami, I., Ying, Z., Re, C., Leskovec, J.: Hyperbolic graph convolutional neural networks. *Adv. Neural Inf. Process. Syst.* 4868–4879 (2019)
20. Zhang, M., Chen, Y.: Link prediction based on graph neural networks. *Adv. Neural Inf. Process. Syst.* 5165–5175 (2018)
21. Zhang, M., et al.: Revisiting graph neural networks for link prediction. *arXiv preprint arXiv*, 2010–16103. (2020)
22. Zhang, M., et al.: An end-to-end deep learning architecture for graph classification. In: *Proceedings of the AAAI Conference Artificial Intelligence*, vol. 32(1) (2018)
23. McCallum, A., Nigam, K., Rennie, J., Seymore, K.: Automating the construction of internet portals with machine learning. *Inf. Retrieval J.* (2000)
24. Fey, M., Lenssen, J.E.: Fast graph representation learning with PyTorch geometric. In: *ICLR Workshop on Representation Learning on Graphs and Manifolds* (2019)

# Traffic Event Reporting Framework Using Mobile Crowdsourcing and Blockchain



Abin Oommen Philip, RA. K. Saravanaguru, and P. A. Abhay

**Abstract** Timely detection of traffic events is of upmost importance in contributing towards traffic safety and ease of commute. A spatial mobile crowdsourcing framework is proposed, enabling road users to report traffic events such as accidents, traffic rule violation, bad road and environment conditions through their mobile devices using images, videos and mobile sensor data. The framework is designed as a decentralized application over blockchain and makes use of interplanetary file system for storage. As transportation infrastructures like CCTV and road side units become increasingly intelligent over time, they are expected to detect traffic events like accidents and traffic rule violations at the edge using machine learning techniques without human intervention. Development and training of traffic event detection models at the edge require training the system using labelled training instances. Crowdsourcing is proposed as a means of collecting training data and validation of event during the learning phase.

**Keywords** Spatial mobile crowdsourcing · Traffic event reporting · Blockchain · Interplanetary file system · Machine learning and smart contracts

## 1 Introduction

Road traffic injuries are the leading cause of death for children and young adults. Approximately, 1.3 million people die each year as a result of road traffic crashes and cost most countries 3% of their gross domestic product [1]. As vehicles on road

---

A. O. Philip (✉) · RA. K. Saravanaguru

School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India

e-mail: [abinphilip1987@gmail.com](mailto:abinphilip1987@gmail.com)

RA. K. Saravanaguru

e-mail: [saravanank@vit.ac.in](mailto:saravanank@vit.ac.in)

P. A. Abhay

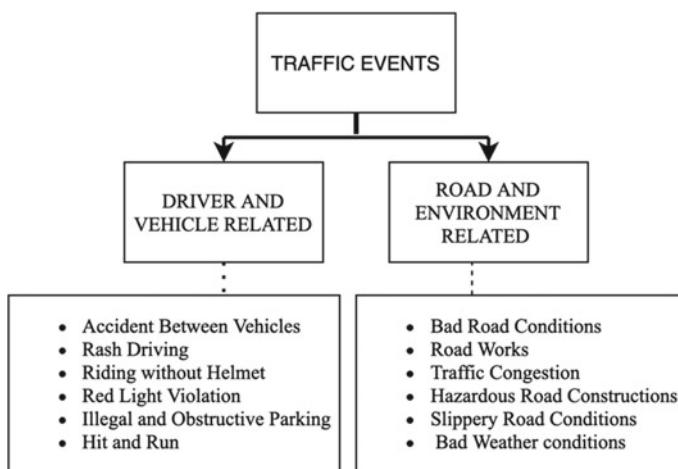
Upeka Venture Catalyst Pvt Ltd., Bangalore, India

increase, the rate of casualties also increase. The main reason being rash driving, non-adherence to traffic rules, road and environment conditions and, at times, negligence and improper infrastructure planning by the civic authorities.

Depending on traffic officials alone to report traffic incidents leads to majority of the incidents being left unreported. Responsible citizens can take up the role of cops in reporting traffic incidents with evidence using their mobile phones. Crowd-sourcing acts a means to gather information and getting various work done by citizens in a cooperative manner. Spatial mobile crowdsourcing involves using mobile device application for crowdsourcing applications that require location, mobility and contextual information [2]. As majority of road users have access to smartphones, they can report various traffic incidents using their mobile devices.

The main vision of intelligent transportation systems (ITS) is safety, reducing congestion and travel time for commute [3]. Detecting, reporting and validation of traffic events like accidents, violation of traffic rules and hazardous road conditions contribute to achieving the aim. The transportation department can utilize spatial mobile crowdsourcing to gather photo-, video- and sensor (accelerometer and gyroscope)-based evidence to validate various traffic events and road conditions. Crowd-sourcing can also be utilized to validate traffic events related to road and environment conditions. Figure 1 highlights some of the traffic events that can be reported and validated via crowdsourcing. Accident-related evidence can help the authorities in investigation and settlement of the cases; rule violation-related evidences can help in penalizing and charging the responsible. Road- and environment-related reports and validation can help in disseminating the information to other road users and help authorities identify hot spots to be rectified.

Eventually, all vehicles and infrastructures like CCTV, road side units would become intelligent, connected and self-capable of detecting incidents at the edge as



**Fig. 1** Types of traffic events reported and validated using crowdsourcing

part of ITS vision. Incidents like red light violation, vehicle collisions, rash driving etc. can be detected by the infrastructure and vehicles itself using computer vision and machine learning capabilities [4–9]. This requires training the system using huge volume of labelled data sets. Crowdsourcing can act as a means of obtaining labelled training data specific to regions and applications. Responsible citizens can take up the duty of reporting traffic incidents; these reports can act as evidence as well as contribute to training data. In the initial implementation phase, the smart edge infrastructure may be less accurate in detections of events and prone to false positives. Crowdsourcing can also be used initially to validate the incidents detected by vehicles and intelligent infrastructure reducing the chances of false positives and increasing the generalizing capabilities of the system.

Blockchain is proposed as a decentralized mechanism to overcome issues like single point of failure and trust associated with depending on a centralized third-party infrastructure [10]. Blockchain provides trust, privacy, immutability, auditability and automated exchange of incentives via smart contracts.

The paper presents an overall conceptual idea and initial results associated with the project in development of a conceptual framework for traffic event reporting, validation and incident specific data collection for training machine learning models to detect incidents from media based on spatial mobile crowdsourcing application over blockchain. The main contributions of the work are

- Propose a conceptual spatial mobile crowdsourcing framework to obtain labelled training instances of traffic events in addition to reporting and validation of traffic incidents by citizens over blockchain.
- Design a mobile application interface and API to be integrated with blockchain and inter planetary file system (IPFS) for the proposed framework.
- Discuss the evaluation and storage of various types of media files on IPFS.
- Design and evaluate the cost and time associated with uploading of evidence over blockchain using smart contracts.

The rest of the paper is organized as follows. Section 2 provides a brief background and survey on related state-of-the-art technologies related to the framework being proposed in Sects. 3. Section 4 discusses the partial results about implementation and evaluation of the proposed framework. Finally, we conclude and present scope for further extensions in Sect. 5.

## 2 Background and Related Work

A brief background and literature survey is conducted in the following areas to understand the state-of-the-art research and appreciate the conceptual framework presented in Sect. 3.

- Spatial mobile crowdsourcing for event reporting.
- Internet of vehicles (IOV) and intelligent transportation systems (ITS).
- Crowdsourcing as means of collecting labelled data for training machine learning models.
- Blockchain in crowdsourcing.

## 2.1 *Spatial Mobile Crowdsourcing*

Traditional crowdsourcing involved Web-based platforms in which humans were actively involved in the procedure of computing and providing data. With the popularity of mobile devices, crowdsourcing has achieved new dimensions like spatial crowdsourcing in which location, mobility and contextual information play a vital role.

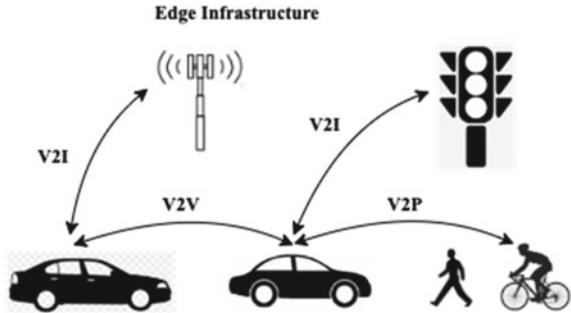
In [11], the authors proposed a photo-based spatial mobile crowdsourcing framework for event reporting that required workers to be present physically at a particular location for fulfilling a particular task using their mobile devices. The event reporting platform helped requesters recruit ideal reporters for their task; the reporters provided photo-based reports from which requestors selected highly relevant data from an evolving picture stream. The techniques used eliminated false submissions and ensured photo credibility and reduced information redundancy. There are mobile-based crowdsourcing applications specifically for reporting of traffic events named “echallan” and “mParivahan” that enable citizens to report traffic violations and road accidents [12]. These mobile crowdsourcing applications hosted on centralized systems suffers from the following drawbacks.

- Anonymity and privacy of information provider is at risk.
- Application suffers from single point of failure.
- These applications are limited to event reporting and does not extend to using the data for labelled training sets.
- Evaluation of trust and incentive mechanism to motivate contribution is not considered.

## 2.2 *Internet of Vehicles and Future of Intelligent Transportation Systems*

Transportation systems and vehicular infrastructures are becoming increasingly connected and intelligent over time. Intelligent transportation systems (ITS) are one of the prominent visions of smart cities. Vehicular Ad hoc networks (VANETS) [13, 14] are paving way to Internet of vehicles (IOV). Internet of vehicles is emerging as a domain from Internet of things [15] enabling all entities in ITS environment to be connected through vehicle to everything connectivity. All entities like vehicles, infrastructures like road side units (RSU's) and even pedestrians would be

**Fig. 2** V2X communication scenarios



connected to each other as shown in Fig. 2. The pedestrians can become part of the V2X connectivity using their mobile phone as communication devices [16]. The communications in V2X environment are enabled via dedicated short-range communication (DSRC) [17] and cellular V2X protocols (C-V2X) [18]. The advent of 5G is envisioned to boost the performance of C-V2X. There have been several other communication protocols suggested for data transmission in vehicular environments [19, 20]. These technologies would enable the infrastructures to become smarter and eventually self-capable of detecting incidents.

### 2.3 *Crowdsourcing and Machine Learning*

Supervised machine learning models are developed using huge volumes of labelled and trained data sets. Especially, training of computer vision-related models require huge volume of labelled image and video streams. Relaying on CCTV footages alone to collect traffic event-related data is not an efficient mechanism in terms of quality and quantity. Crowdsourcing of traffic event images and footages via mobile applications by responsible citizens for incentive can provide an efficient means of collecting event specific data set. Crowdsourcing has been suggested as means of video annotation mechanism in literature [21]. The authors in [22] used crowdsourcing data as a means of predicting real-time traffic accidents post-impact prediction. Computer vision and machine learning algorithms have been proposed to detect traffic-related events like rash driving, red light violation, accidents, helmet less driving etc. [4–6].

### 2.4 *Crowdsourcing Using Blockchain and IPFS*

The authors in [23] justify the need for using blockchain for crowdsourcing applications as a means of ensuring security, privacy and trust. In [24], the authors discussed assessing trustworthiness of crowdsourced data using smart contracts. Quality of data

collected is crucial and includes detection of intentional or unintentional manipulation, deception and spamming. Blockchain was proposed as means of eliminating single authority and increasing resilience to institutional data manipulation. Zebralancer [25] was proposed as private and anonymous crowdsourcing system over blockchain, overcoming two fundamental challenges of decentralizing crowdsourcing, data leakage and identity breach. The work also discussed how malicious entities could be prevented ensuring anonymity.

There have been frameworks designed for detecting and investigation of accidents over blockchain [26–29]. The proposed crowdsourcing application can act as an additional source of evidence in such cases. There are other works in literature that discuss trust calculation of crowd sourcing entities over blockchain and mechanisms for incentive calculation of contributors [30, 31].

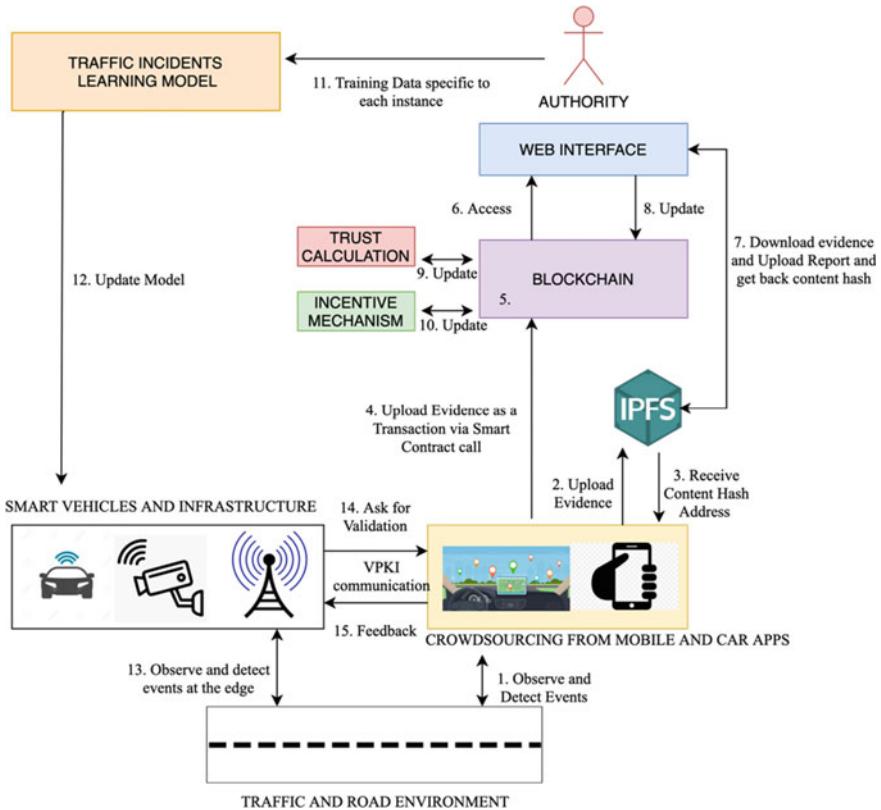
Blockchain as such is not suitable for storing large files like images, videos etc. Depending on cloud storage brings all the associated risks with centralized storage [32]. Interplanetary file system (IPFS) acts as a decentralized file storage mechanism ensuring integrity, immutability and availability [33]. TFCrowd [34] made use of IPFS as a solution for crowdsourcing contributors to upload solutions. IPFS returns a content address hash (CID) for the file uploaded to be accessed later.

## 2.5 Research Gap and Problem Formulation

From the studies conducted, it is observed that there exist several proposed independent systems to report traffic events via crowdsourcing. The possibility of detecting traffic events at the edge through CCTV has also been explored in literature. The automated traffic event detection through computer vision and machine learning can be made more efficient by providing labelled data sets that are context- and location-specific, enhancing the accuracy of the system. In this regard, the authors propose combining multiple independent approaches together as a conceptual framework in the specific context of traffic event reporting and validation. There are no frameworks proposed in literature to the best of our knowledge that combines all these mechanism into a single conceptual framework in the context of traffic event reporting and validation. We discuss how each individual systems can be combined and the flow of events in improving the overall process of traffic event detection and eventually making traffic event reporting systems to work on its own at the edge in future. The contributions made and novelty of the work in this regard has been discussed in Sect. 1.

## 3 Proposed Framework

The overall conceptual framework is presented in Fig. 3. The framework can be viewed as a combination of four subsystems consisting of



**Fig. 3** Overall framework

- Crowdsourcing from mobile application and vehicle, with flow of events explained as steps 1 to 4.
- Event validation by authority over blockchain and IPFS, explained as steps 4 to 10.
- Training machine learning models to detect traffic events using context and location specific data, explained as steps 11 and 12.
- Detection of incidents at the edge and validation via crowdsourcing in the initial phase, explained as steps 13 to 15.

It is assumed that the entities like CCTV, vehicles, RSUs and road users via mobile devices are part of the V2X communication environment. The procedure steps and flow of events happening in the framework are described in brief.

1. The road users observe traffic-related events and capture it using the crowd-sourcing mobile application. The vehicles can also contribute to the crowd-sourced data via dashcam recordings and on board diagnostic data [35].

2. Image-, video- and sensor-related evidence pertaining to traffic events are uploaded to IPFS.
3. The content hash (CID) that acts as a unique address identity of the file uploaded is returned to the crowdsourcing application.
4. The user through the crowdsourcing application uploads details like type of incident, description of event, timestamp, location and IPFS hash as a transaction via a smart contract call to blockchain network. The users can record and report event only via the application interface (discussed in Sect. 4.1); thus, it ensures the authenticity of location and time of event. The GPS location and timestamp are added to the event by the application.
5. The transaction gets added as part of a block and gets added to the blockchain via consensus mechanism.
6. The authorities can access the traffic events and evidence through a Web interface. The events are segregated based on location and each authority may get to see only the events within their jurisdiction.
7. The files related to the incident may be downloaded from the IPFS using the content hash address that form a part of the transaction. The contents in IPFS may be pinned using paid services for permanency. Once the contents are downloaded, the contents may be unpinned. The integrity of the content is still ensured as the hash remains as part of transaction in blockchain. The authorities add report and findings related to the incident reported to IPFS, and the content address hash is obtained.
8. The IPFS hash corresponding to investigation and settlement report along with details related to verification of event reported is added to the blockchain as transaction.
9. Based on the verification of the event by authority, the trust score of the evidence provider is updated. The trust score may be decremented if false and unrelated reports are submitted and such users may be prevented from uploading reports in future if their trust scores fall below a threshold. Blacklisted incident reporters are updated as transaction in the blockchain. There are works in literature that discuss calculation and updating of trust scores for entities in vehicular and crowdsourcing environment [36, 37].
10. The incentive mechanism module performs calculation related to incentive to be returned to data provider based on type of incident and their trust score. Dynamic incentive calculations techniques may be applied for deciding incentives. The incentives may be in form of cash or discount coupons and can be dispersed via smart contracts.
11. The authority after verification of image-, video- and sensor-related data may provide the data to traffic event specific training and learning module to learn new instance and generalize the event detection model with new instances. The module deals with training of models to detect specific traffic incidents using images, videos and sensor data. As the files provided are already categorized based on type of incident, it enables ease of obtaining new instances for labelled data and training the model. Few examples of kind of incidents that can be learned include detection of rash driving and red light violations from video

- streams, detection of helmet less driving from images, detection of potholes and sharp unscientific road curves from mobile accelerometer and gyroscopes values.
12. The updated traffic incident detection model is deployed on the edge devices like CCTV, RSU and vehicles and gets constantly updated as model learns based on new instances.
  13. The edge infrastructure devices like CCTVs, RSUs and vehicles are enabled with intelligence to detect traffic incidents at the edge.
  14. In the initial deployment phase, the detections made by edge devices may not always be accurate and prone to false positives. The intelligent edge infrastructures may ask the vehicles and road users via crowdsourced applications to validate the incident detected. For example, the applications may be asked to validate the observance of rash driving, red light violation etc. These communications can happen via vehicular public key infrastructure (VPKI) [38] assuming all the entities are part of the V2X environment.
  15. The vehicles and road users on receiving the validation request may provide feedback via the crowdsourcing application interface. Over the period, the infrastructure becomes self-reliant in detecting events at the edge without human intervention. The RSUs and CCTVs on detection of accidents may send a request notification to road users in the vicinity to provide additional evidence in the form of photos and videos. The road users may upload evidence similar to the procedure followed in step 1–5.

## 4 Implementation and Evaluation

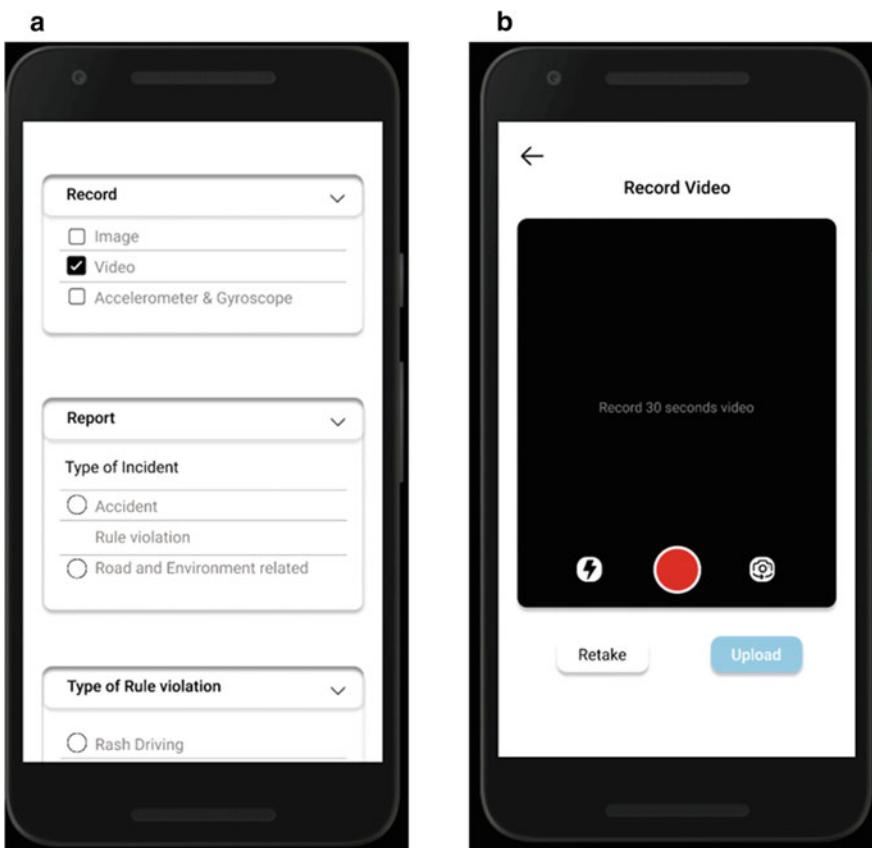
The project is part of ongoing research work by the authors and partial results such as mobile application interfaces, API calls to integrate IPFS and Ethereum blockchain and smart contract design justifying the feasibility of conceptual framework is discussed. The section discusses the implementation and evaluation of following

- Interface Design of crowdsourcing application.
- API calls to integrate mobile application with IPFS and Ethereum blockchain via Infura.
- Evaluate the time required to upload image, video and sensor data to IPFS and return unique content address hash.
- Design and evaluation of smart contract in terms of gas cost to upload event report via mobile application.

#### 4.1 Design of Mobile Interface for Crowdsourcing Application

We discuss the design of interface and flow for crowdsourcing mobile application. The user interface was designed using Figma [39] and exported to android studio [40] for building the application. The crowdsourcing-based event reporting mobile application interface has been designed keeping in mind ease of use. Next, we walk through the flow of events happening as part of uploading and reporting an event using the mobile application interface.

- The user can select the type of file to be recorded as shown in Fig. 4a Categories include image, video and mobile sensor values. Accelerometer and gyroscope values are recorded as part of mobile sensor values. The video and sensor values are recorded for 30 s maximum.

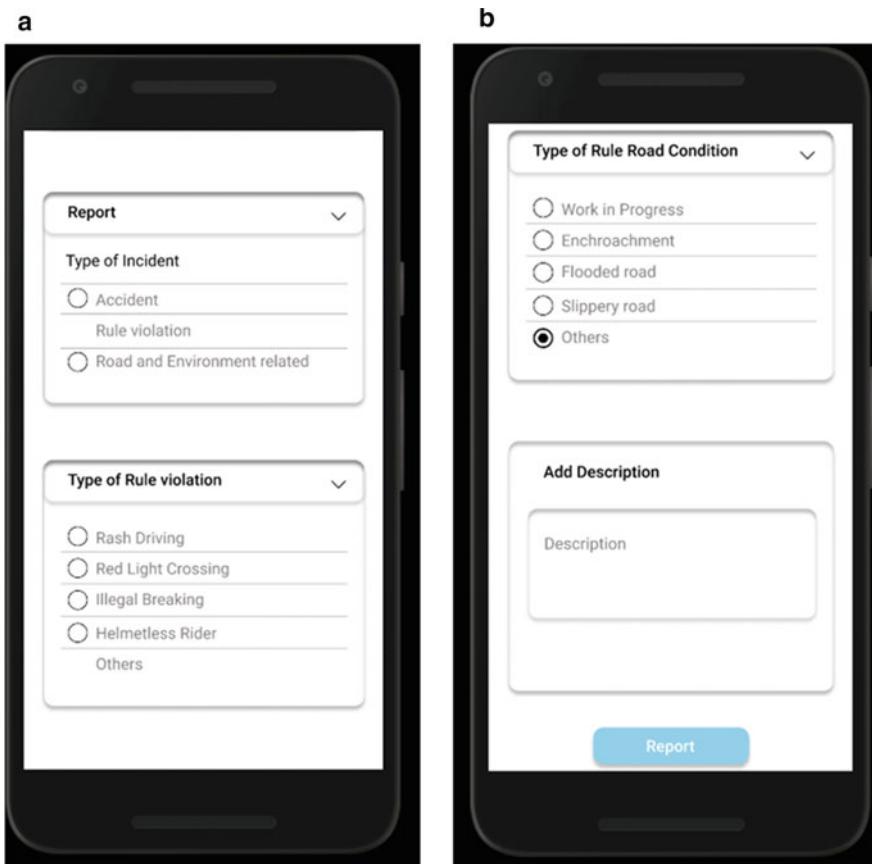


**Fig. 4** **a** Interface for selecting type of file to be recorded (left). **b** Interface for recording video (right)

- Video, image and sensor value can only be recorded live via the application as shown in Fig. 4b. Pre-recorded files cannot be used for upload. This ensures timestamped location- and context-specific data.
- The accelerometer and gyroscope sensor readings are stored and transmitted as csv files. A sample pattern of 30 s sensor reading corresponding to a user travelling through a road in bad condition with potholes was collected and is presented in Fig. 5.
- Once the file is recorded, the interface returns for the user to select type of incident to be reported as in Fig. 6a. The categories of event include accident, rule violation,

**Fig. 5** Accelerometer pattern corresponding to travelling through potholes



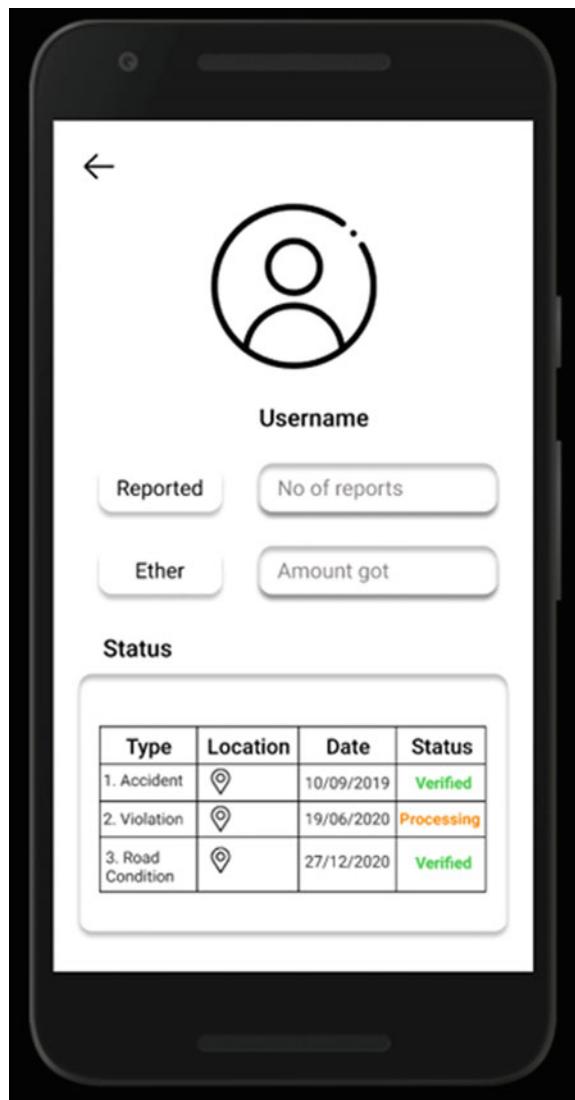


**Fig. 6** **a** Interface for selecting type of incident and subcategory. **b** Interface for uploading the event and adding description

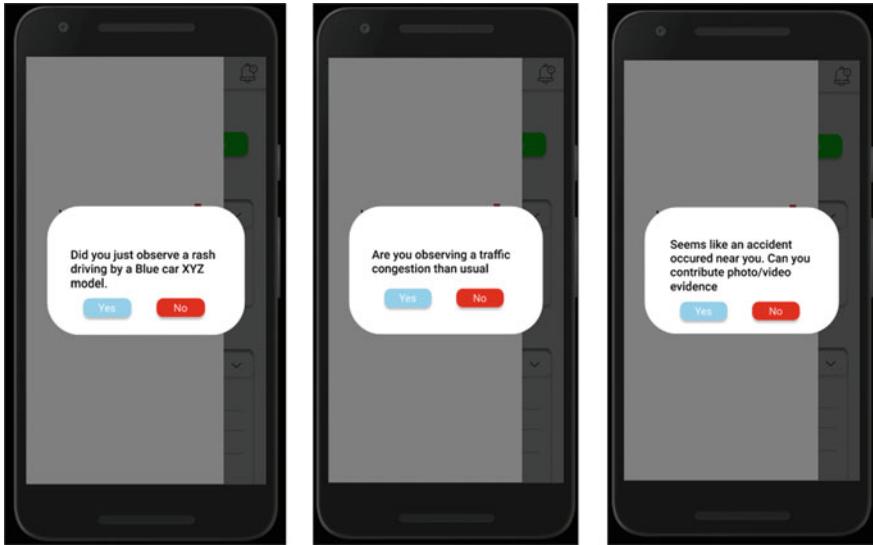
road and environment related. Further subcategory within type of event can be selected by further scrolling down.

- Additional description regarding the event can also be added as comment by the user as in Fig. 6b. On clicking Report button, the media file gets automatically uploaded to IPFS and content address hash is returned. The attributes Incident Id, Type of Incident, Sub Category within incident, Description, IPFS hash, Timestamp and Location gets written onto blockchain via smart contract call. All the upload and submissions happen in the background via API calls, and user is freed from the burden of additional technologies.
- The user can view status of each event reported along with the reward obtained as in Fig. 7.
- The smart infrastructure can validate events via crowdsourcing. Users may also be asked to provide additional evidence regarding an accident that might have

**Fig. 7** Interface to view status of events reported and rewards received



occurred. The validations happen as notification pop ups as shown in Fig. 8. It is assumed that the mobile phone is part of the V2X communication environment. The validation can also happen using similar notifications popping up on car users' dashboard.



**Fig. 8** Notification pop up to validate events detected by infrastructure

#### 4.2 Integration of the Application with IPFS and Ethereum Blockchain

- Infura [41] provides the required tools and infrastructure for the mobile application to access Ethereum and IPFS. The mobile application device need not run a blockchain node; it can access Ethereum and IPFS via Infura. As the first step of development, a project is created on Infura and credentials like PROJECT\_ID and PROJECT\_SECRET are obtained. These credentials are used to authenticate the API calls to Infura from the mobile application.
- As the users contribute media files via the mobile application, the backend code calls the API corresponding to uploading file to IPFS via Infura as in Fig. 9.
- The file would be pinned by default using /api/v0/add command. The purpose of pinning the file is for ensuring availability and permanency. The IPFS clears all unpinned files occasionally as part of maintenance and garbage collection. Once the authorities download the file using the content address hash, the file can be unpinned as in Fig. 10.
- The authorities would be interacting with blockchain via a Web interface as in Fig. 11. The Web interface would provide details of incidents reported based on location. The authorities can retrieve the evidence file from IPFS using the content address hash. After the investigation is completed, the authorities submit a report. The Web application can interact with Ethereum via Infura using the API in Fig. 12. The network used for our development was Ropsten test network [42].
- The authorities interact with IPFS to download the media file uploaded by users. The content address hash obtained from corresponding transaction is

```

curl -X POST -F file=@myfile \
-u "PROJECT_ID:PROJECT_SECRET" \
"https://ipfs.infura.io:5001/api/v0/add"

> {
    "Name": "ipfs_file_docs_getting_started_demo.txt",
    "Hash": "QmeGAVddnBSnKc1DLE7DLV9uuTqo5F7QbaveTjr45JUDQn",
    "Size": "44"
}

```

**Fig. 9** API to upload file to IPFS via Infura from mobile application

```

curl -X POST -u "PROJECT_ID:PROJECT_SECRET" \
"https://ipfs.infura.io:5001/api/v0/pin/rm?arg=QmeGAVddnBSnKc1DLE7DLV9uuTqo5F7QbaveTjr45JUDQn"

```

**Fig. 10** API to unpin the file in IPFS

**Transport Department**

Type	Date	Location	Description	Validate	IPFS Hash	PDF Report
1. Accident	10/09/2019	📍	The horrific high-speed car crash that killed seven—all in their 20s—in the early hours of Tuesday	Validated	QmNnR9kUXFmJEQ7joS..	
2. Violation	19/06/2020	📍	A speedy tempo, carrying milk to deliver in south Mumbai, dashed several people sleeping on the road ...	Processing	QmSz6594cCGXPAHoBxb..	
3. Road Condition	27/12/2020	📍	Huge holes formed after heavy rain could cause major accidents for drivers who are not aware of the hole when they are filled with rain water	Verified	QmSzTk2Lm3fzc86HWC..	

**Fig. 11** Web interface for authorities to interact with blockchain

```

curl https://<network>.infura.io/v3/YOUR-PROJECT-ID \
-X POST \
-H "Content-Type: application/json" \
-d '{"jsonrpc":"2.0","method":"eth_accounts","params":[],"id":1}'

```

**Fig. 12** API to connect Web application to Ethereum via Infura

```
curl -X POST -u "PROJECT_ID:PROJECT_SECRET" \
"https://ipfs.infura.io:5001/api/v0/cat?arg=QmeGAVddnBSnKc1DLE7DLV9uuTqo5F7QbaveTjr45JUdQn"
```

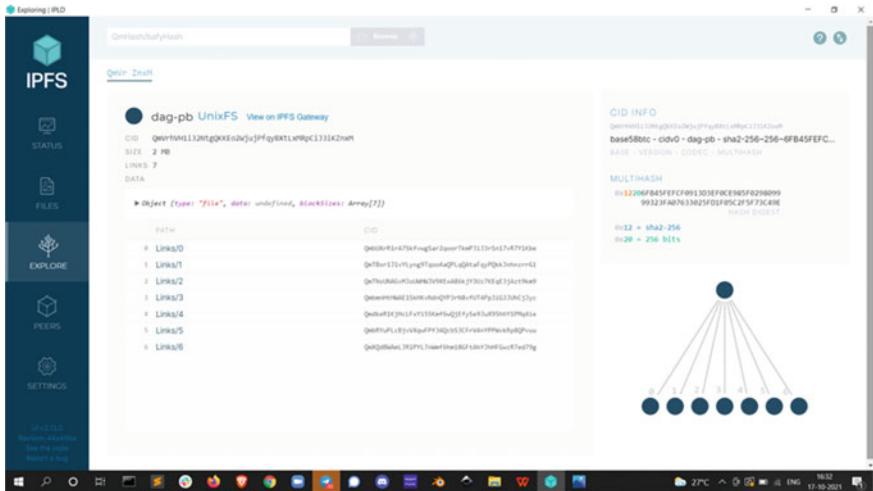
**Fig. 13** API to download file from IPFS using content address hash

used to download the file from IPFS using API call in Fig. 13. The credentials PROJECT\_ID and PROJECT\_SECRET are used for authentication.

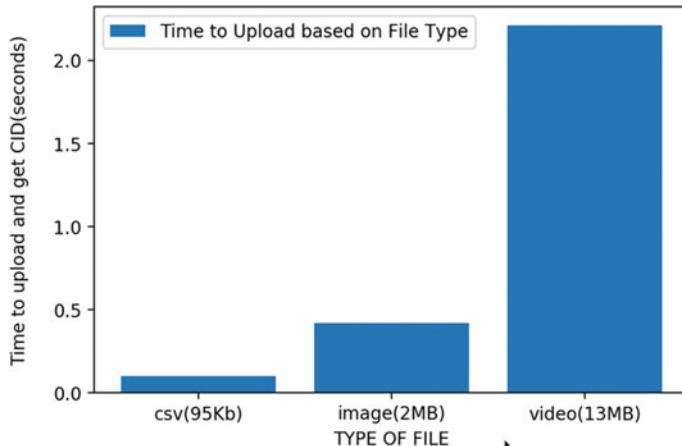
### 4.3 IPFS Upload and Evaluation

Next, we discuss details regarding uploading evidences files onto IPFS. An image, video and csv file corresponding to accelerometer and gyroscope sensor data is uploaded to IPFS and time to upload and retrieve back the corresponding content address hash is evaluated. The data gets split and is stored as chunks of 256 kb each on IPFS nodes. Details corresponding to uploading an image file from mobile application onto IPFS is presented in Fig. 14. The image file is split into seven chunks of 256 kb each. Each chunk generates a hash; they are combined to obtain the root hash which corresponds to the address (CID) to retrieve the file.

We compare the time in seconds taken to upload an image file of 2 MB, video file of 30 s duration of 13 MB and csv file of 95 kb corresponding to accelerometer and gyroscope data for 30 s in Fig. 15. The bandwidth during upload was on an average 3 MBps. The time taken to upload and retrieve content address hash justifies IPFS as a feasible means of decentralized storage ensuring integrity of data.



**Fig. 14** Details on uploading an image file to IPFS



**Fig. 15** Time comparison to upload various types of media files onto IPFS

#### 4.4 Smart Contract Design and Evaluation

We briefly discuss the implementation details of the smart contract. Ropsten test network has been used to implement Ethereum blockchain [43]. The smart contract is written in solidity and designed using Remix environment. The smart contract is used to perform transaction like uploading of event report by crowdsourcing application. The attributes that form part of the transaction is discussed in Table 1, and the smart contract code in solidity is presented in Fig. 16.

**Table 1** Attributes of transaction send to blockchain

Attribute	Description
Incident Id	Corresponding to each event report an ID is generated
Type of incident	The types of incidents include accident, rule violation, road and environment-related report. The category is selected by user using the mobile application interface
Subcategory	Each incident can be further categorized. Like rule violation can be further categorized into red light violation, illegal parking, riding without helmet etc. This category is also selected by user through the mobile application interface
Description	The user can further describe the event in their own words to provide additional information
Filehash	The IPFS hash of the evidence file that acts as content address hash to retrieve the file from IPFS ensuring integrity
Timestamp	Time of incident upload
Location	GPS coordinates of the incident
setCrowdData()	All the attributes are bundled together and passed to blockchain via smart contract call using this function

```

1 // SPDX-License-Identifier: MIT
2 pragma solidity ^0.5.0;
3
4 contract CReport{
5     constructor () public {
6         owner = msg.sender;
7     }
8
9     address public owner;
10
11 struct User{
12     string userType;
13     address userAddress;
14 }
15
16 mapping (address => User) public Users;
17
18 struct CrowdData{
19     uint256 id;
20     string location;
21     string timestamp;
22     string typeofincident;
23     string subcategory;
24     string description;
25     string reportHash;
26 }
27
28 mapping (uint => CrowdData) public crowdDatas;
29
30
31 function setCrowdData(uint256 id,string memory location,string memory timestamp,string memory typeofincident,
32 string memory subcategory,string memory description,string memory reportHash) public {
33     crowdDatas[id]= CrowdData(id,location,timestamp,typeofincident,subcategory,description,reportHash);
34 }
35 }
36

```

**Fig. 16** Smart contract code in solidity designed in Remix environment

**Table 2** Evaluation of smart contract

Action performed	Gas used	Cost (ether)	Time (s)
Contract deployment	905,891	0.000906 ETH	11.28
setCrowdData()	296,409	0.000296 ETH	13.83

The smart contract used to upload the event report is evaluated in terms of cost, and time and results are presented in Table 2. The contract deployment is a one-time process and happens initially. Each time user uploads an evidence report, the function setCrowdData() is called. Based on the cost of Ether during design [44], the one-time cost for deploying the contract was 3.5\$ and that for performing a transaction was 0.80\$. The cost is justified. The cost is for performing the transaction over public Ethereum network. If a private blockchain network is designed specifically for the crowdsourcing, this cost can be mitigated. The cost associated in a private blockchain will be that of maintaining the nodes and transactions won't involve cost. The time taken in seconds for the transactions to get committed is also presented. The incentive provided to the crowdsources may be calculated as transaction cost incurred plus an incentive amount.

#### ***4.5 Implementation and Deployment in Real World***

We briefly discuss and consolidate the technologies in terms of hardware and software that would be required to implement and deploy the framework proposed in real world.

- It is proposed that the events can be reported by mobile applications and vehicles. The mobile crowdsourcing application can be designed for both Android and iOS and made available to users. The reporting and data collection from vehicle side can take place via ELM327 OBD 2 scanner integrated onto vehicles [45]. All vehicles manufactured since 2011 comes with an OBD port. The OBD port and dash camera in vehicle may be integrated to Car OS or can be paired with user mobile application to communicate event reports.
- For blockchain implementation, we may prefer permissioned blockchain mechanisms like hyperledger fabric [46] or IOTA Tangle [47] over public networks like Ethereum. Permissioned blockchain application can be custom designed for the application. IPFS service can be used as proposed in Sect. 4.2.
- A frontend website for authorities to interact with blockchain can be created using Web3.js. [48] discusses how frontend applications can be integrated with blockchain.
- The authorities may download and store video-, image- and sensor-based evidences that are labelled with type of incident and location. This forms the data set for training the machine learning model to classify incidents based on type. High-end systems or cloud computing facilities may be used to train the deep learning models based on computer vision.
- For automatic detection of incidents at the edge by devices like CCTV and edge infrastructures, the trained machine learning model needs to be deployed onto the edge device. Edge computing devices and platforms like NVIDIA Jetson Nano [49], Qualcomm vision intelligence platform integrated with system on chip (SoCs) [50] can provide the IoT-enabled traffic infrastructure devices the required processing and intelligence capability.

### **5 Conclusion and Future Work**

A conceptual framework based on spatial mobile crowdsourcing was proposed to report traffic related incidents over blockchain using IPFS as decentralized storage mechanism. The results presented are part of the ongoing work by the authors in designing an efficient framework for secure event detection and evidence management over blockchain. In contrast to traditional traffic event reporting mechanisms, the events reported via the framework by citizens plays a dual role. Firstly, it helps the authorities with investigation of incidents and penalizing the rule violators, hence ensuring traffic law and order. Secondly, the data collected via mobile crowdsourcing also helps towards enabling automatic detection of traffic events by infrastructures

and vehicles in future. The data gathered are specific to traffic event scenarios and act as a source of labelled data set for training machine learning models to be deployed at the edge. These models can be deployed onto the infrastructures like CCTVs and RSUs to detect traffic incidents and events automatically. The framework proposed ensures traffic event reporting and detection model for the present and the future. The framework also discussed how the evidence-related files could be stored and managed over blockchain and IPFS, ensuring trust and providing incentives for data providers via smart contracts.

The framework proposed has got challenges and scope for further extension in the following aspects.

- Development and integration of a proper Web interface for authorities to interact and view data from blockchain and IPFS without having to bother about backend technology.
- With the provision to report incidents anonymously, it is a challenge to allocate and disperse incentives to the information provider. Better incentive calculation algorithms can be designed for crowdsourcing.
- Efficient deep learning models can be designed to automatically label the media files obtained for each traffic event category and integrate it as part of continuous learning model.
- Efficient computer vision-enabled edge computing machine learning models capable of automatic event detection can be developed.

## References

1. World Health Organisation Report: <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries> (2021). Accessed Oct. 2021
2. Tong, Y., Chen, L., Shahabi, C.: Spatial crowdsourcing. Proceedings of the VLDB Endowment 10 (2017). <https://doi.org/10.14778/3137765.3137827>
3. Oommen Philip, A., Rak, S.: A vision of connected and intelligent transportation systems. Int. J. Civ. Eng. Technol. **9**, 873–882 (2018)
4. Ijjina, E.P., Chand, D., Gupta, S., Goutham, K.: Computer vision-based accident detection in traffic surveillance. In: 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE (2019)
5. Hochstetler, J., Padidela, R., Chen, Q., et al.: Embedded deep learning for vehicular edge computing. In: 2018 IEEE/ACM Symposium on Edge Computing (SEC). IEEE (2018)
6. Franklin, R.J., Mohana, M.: Traffic signal violation detection using artificial intelligence and deep learning. In: 2020 5th International Conference on Communication and Electronics Systems (ICCES). IEEE (2020)
7. Sharma, R., Sungsheetha, A.: An efficient dimension reduction based fusion of CNN and SVM model for detection of abnormal incident in video surveillance. J. Soft Comput. Paradigm **3** (2021). <https://doi.org/10.36548/jscp.2021.2.001>
8. Manoharan, S.: AN improved safety algorithm for artificial intelligence enabled processors in self driving cars. J. Artif. Intell. Capsule Netw. (2019). <https://doi.org/10.36548/jaicn.2019.2.005>
9. Bestak, R.: Intelligent traffic control device model using ad hoc network. J. Inf. Technol. Digit. World **01** (2019). <https://doi.org/10.36548/jitdw.2019.2.002>

10. Chowdhury, M.J.M., Colman, A., Kabir, M.A., et al.: Blockchain versus database: a critical analysis. In: 2018 17th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/12th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE). IEEE (2018)
11. Hamrouni, A., Ghazzai, H., Frikha, M., Massoud, Y.: A spatial mobile crowdsourcing framework for event reporting. *IEEE Trans. Comput. Soc. Syst.* **7** (2020). <https://doi.org/10.1109/TCSS.2020.2967585>
12. Economic Times Report
13. Chen, J.I.Z., Hengjinda, P.: Enhanced dragonfly algorithm based K-Medoid clustering model for VANET. *J. ISMAC* **3** (2021). <https://doi.org/10.36548/jismac.2021.1.005>
14. Bhatia, T.K., Ramachandran, R.K., Doss, R., Pan, L.: A comprehensive review on the vehicular ad-hoc networks. In: 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). IEEE (2020)
15. Kaiwartya, O., Abdullah, A.H., Cao, Y., et al.: Internet of vehicles: motivation, layered architecture, network model, challenges, and future aspects. *IEEE Access* **4** (2016). <https://doi.org/10.1109/ACCESS.2016.2603219>
16. Sewalkar, P., Seitz, J.: Vehicle-to-pedestrian communication for vulnerable road users: survey, design considerations, and challenges. *Sensors* **19** (2019). <https://doi.org/10.3390/s19020358>
17. Li, Y.: An Overview of the DSRC/WAVE Technology (2012)
18. Wang, P., Di, B., Zhang, H., et al.: Cellular V2X communications in unlicensed spectrum: harmonious coexistence with VANET in 5G systems. *IEEE Trans. Wirel. Commun.* **17** (2018). <https://doi.org/10.1109/TWC.2018.2839183>
19. Casteigts, A., Nayak, A., Stojmenovic, I.: Communication protocols for vehicular ad hoc networks. *Wirel. Commun. Mob. Comput.* **11** (2011). <https://doi.org/10.1002/wcm.879>
20. Paul, R., Sebastian, N., Yadukrishnan, P.S., Vinod, P.: Study on data transmission using Li-Fi in vehicle to vehicle anti-collision system (2021)
21. Vondrick, C., Ramanan, D., Patterson, D.: Efficiently scaling up video annotation with crowdsourced marketplaces (2010)
22. Lin, Y., Li, R.: Real-time traffic accidents post-impact prediction: based on crowdsourcing data. *Accid. Anal. Prev.* **145** (2020). <https://doi.org/10.1016/j.aap.2020.105696>
23. Ma, Y., Sun, Y., Lei, Y., et al.: A survey of blockchain technology on security, privacy, and trust in crowdsourcing services. *World Wide Web* **23**: (2020). <https://doi.org/10.1007/s11280-019-00735-4>
24. Mihelj, J., Zhang, Y., Kos, A., Sedlar, U.: Crowdsourced traffic event detection and source reputation assessment using smart contracts. *Sensors* **19** (2019). <https://doi.org/10.3390/s19153267>
25. Lu, Y., Tang, Q., Wang, G.: ZebraLancer: private and anonymous crowdsourcing system atop open blockchain. In: 2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS). IEEE (2018)
26. Cebe, M., Erdin, E., Akkaya, K., et al.: Block4Forensic: an integrated lightweight blockchain framework for forensics applications of connected vehicles. *IEEE Commun. Mag.* **56** (2018). <https://doi.org/10.1109/MCOM.2018.1800137>
27. Philip, A.O., Saravanan, R.A.K.: Secure incident and evidence management framework (SIEMF) for internet of vehicles using deep learning and blockchain. *Open Comput. Sci.* **10**, 408–421 (2020). <https://doi.org/10.1515/comp-2019-0022>
28. Abhay, P.A., Jishnu, N.V., Meenakshi, K.T., et al.: Auto block IoT: a forensics framework for connected vehicles. *J. Phys. Conf. Ser.* **1911** (2021). <https://doi.org/10.1088/1742-6596/1911/1/012002>
29. Oham, C., Jurdak, R., Kanhere, S.S., et al.: B-FICA: BlockChain based framework for auto-insurance claim and adjudication. In: 2018 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData). IEEE. (2018)
30. Tang, W., Zhang, K., Ren, J., et al.: Privacy-preserving task recommendation with win-win incentives for mobile crowdsourcing. *Inf. Sci.* **527** (2020). <https://doi.org/10.1016/j.ins.2019.02.011>

31. Tong, Y., Zhou, Z., Zeng, Y., et al.: Spatial crowdsourcing: a survey. VLDB J. **29** (2020). <https://doi.org/10.1007/s00778-019-00568-7>
32. Zahed Benisi, N., Aminian, M., Javadi, B.: Blockchain-based decentralized storage networks: a survey. J. Netw. Comput. Appl. **162**. <https://doi.org/10.1016/j.jnca.2020.102656>
33. IPFS: <https://ipfs.io> (2021). Accessed Oct. 2021
34. Li, C., Qu, X., Guo, Y.: TFCrowd: a blockchain-based crowdsourcing framework with enhanced trustworthiness and fairness. EURASIP J. Wirel. Commun. Netw. (2021). <https://doi.org/10.1186/s13638-021-02040-z>
35. Nirmali, B., Wickramasinghe, S., Munasinghe, T., et al.: Vehicular data acquisition and analytics system for real-time driver behavior monitoring and anomaly detection. In: 2017 IEEE International Conference on Industrial and Information Systems (ICIIS). IEEE (2017)
36. Feng, W., Yan, Z., Zhang, H., et al. A survey on security, privacy, and trust in mobile crowdsourcing. IEEE Internet of Things J. **5** (2018). <https://doi.org/10.1109/JIOT.2017.2765699>
37. Wang, D., Chen, X., Wu, H., et al.: A blockchain-based vehicle-trust management framework under a crowdsourcing environment. In: 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). IEEE (2020)
38. Haidar, F., Kaiser, A., Lonc, B.: On the performance evaluation of vehicular PKI protocol for V2X communications security. In: 2017 IEEE 86th Vehicular Technology Conference (VTC-Fall). IEEE (2017)
39. Figma Tool: <https://www.figma.com> (2021). Accessed Oct. 2021
40. Android Studio. In <https://developer.android.com/studio> (2021). Accessed Oct. 2021
41. Infura: <https://infura.io> (2021). Accessed Oct. 2021
42. Ropsten Test Network. <https://ropsten.etherscan.io> (2021). Accessed Oct. 2021
43. Ethereum. In <https://ethereum.org/en/> (2021). Accessed Oct. 2021
44. Ethereum Cost Calculator. In: <https://ethereumprice.org/calculator/>
45. Amarasinghe, M., Kotegoda, S., Arachchi, AL., et al.: Cloud-based driver monitoring and vehicle diagnostic with OBD2 telematics. In: 2015 Fifteenth International Conference on Advances in ICT for Emerging Regions (ICTer). IEEE (2015)
46. Hyperledger: <https://www.hyperledger.org> (2021). Accessed Oct. 2021
47. IOTA Tangle. <https://www.iota.org> (2021). Accessed Oct. 2021
48. Web3: <https://web3js.readthedocs.io/en/v1.52/> (2021). Accessed Oct. 2021
49. Jetson Nano Development Kit. <https://developer.nvidia.com/embedded/jetson-nano-developer-kit> (2021). Accessed Oct. 2021
50. Qualcomm Vision Intelligence Platform. <https://developer.qualcomm.com/hardware/vertical-platforms/vision-intelligence-platform> (2021). Accessed Oct. 2021

# A Flexible Protocol for a Robust Hospitals Network Based on IoT



Salma Rattal, Kamal Ghoumid, and El Miloud Ar-Reyouchi

**Abstract** The Internet of things (IoT) requires reliable connectivity with a great network capacity and free packet loss. This article offers a flexible protocol based on network coding (PBNC) to control the healthcare data transmission of the patients in hospital by using IoT, recovering and correcting lost and wrong packets, respectively. We propose a network coding (NC)-based method for an IoT situation where one radio modem (RM) must interact with many remote terminal units (RTUs) via a wireless channel. We examine the advantages and drawbacks of acknowledgment (ACK). In addition, we compare and assess the performance of the suggested PBNC with a state-of-the-art method (SoAM). The results indicate that the PBNC significantly improves data transmission quality in the network hospital, reducing the number of transmission and retransmission by 95%, recovering lost packets, and improving network capacity by 75%.

**Keywords** Internet of thing devices · Network coding · ACK · Wireless communication protocol

## 1 Introduction

Both reducing retransmission and improving the total network capacity are a process that seems straightforward, but in fact, it may be a tremendously tough and rather complicated challenge. NC [1] is a basic combination method that may offer numerous potential benefits to the IoT network. As shown by Alabady et al. [2], it improves network capacity, optimizes wireless communication systems [3], remarkably reduces the number of retransmissions for the efficient detection and correction of erroneous packets, and provides many advantages to the network. Random linear

---

S. Rattal · K. Ghoumid  
ENSAO, Mohammed I University, Oujda, Morocco

E. M. Ar-Reyouchi (✉)  
Faculty of Sciences, Abdelmalek Essaâdi University, Tetouan, Morocco  
e-mail: [e.arreyouchi@m.ieice.org](mailto:e.arreyouchi@m.ieice.org)

network coding (RLNC) [4] is a sophisticated NC approach that allows network nodes to create coded packets over a constrained field from input source data in a random linear combination. RLNC can decode and decrease the amount of retransmission needed at the destination. In our context, the most common method to repair and restore lost packets is to use adaptable and robust medical network coding protocols to recover erroneous or lost packets easily. In [5], the authors assess the efficacy of error correction techniques, Automatic Retransmission reQuest (ARQ), for data transmission reliability in transport protocols but do not examine and determine their performance. They only take into account a small number of permitted retransmissions. In the IoT, some codes may encourage protocol flexibility and favor energy-efficient, reliable and accurate data on the present condition, like in [6, 7]. Different limitations produce varying modulation rates on transmitted packets. Higher modulation rates result in lower receiver sensitivity and a smaller coverage range due to faster packet speeds. In this regard, work [8] investigates the impact of packet size on bitrate, forwarding time, and modulation rate. The product of cost and packet delay [9], the polling cycle [10], and the dependability of IoT communication protocols [11] are all parameters that need to be improved in the IoT. The performance of the medical network is influenced by several factors, including error correction techniques, average packet size, and modulation rate. We demonstrate in this paper that NC can importantly decrease the number of retransmissions, recover lost packets, and improve IoT applications. We propose a coding-based retransmission network for recovering lost packets, minimizing retransmission, and greatly improving network capacity [12]. In our case, each packet is retransmitted with a new format before being sent to the other network. This paper proposes PBNC for recovering lost packets based on RLNC. It solves packet loss recovery from the RTUs to monitor medical information, improving and reforming wireless communication for IoT devices. The main contribution of this research is to reduce the number of retransmissions required to rectify possible mistakes, thus increasing device efficiency. Therefore, our contribution is to protect a hospital network against transmission errors over noisy channels, transmission errors, and data loss. These contributions can also improve the network capacity by using a robust protocol that considers the factors related to error corrections. The short introduction is the first of seven parts of this article. The error-correcting codes are discussed in Sect. 2. The issue description and a solution are presented in Sect. 3. In Sect. 4, the system model is presented. The procedure is described in Sect. 5. In Sect. 6, the results are provided and discussed. Lastly, in Sect. 7, the conclusions are presented.

## 2 The Principal Used Error Correction and Proposed PBNC Purpose

There are two main techniques for recovering faulty packets in any wireless medical network: forward error correction (FEC) and ARQ.

## 2.1 FEC Technique

The FEC method enables the IoT [13] to offer a great foundation for ensuring a stable wireless connection in a wireless network connection. FEC is widely used in Rayleigh fading channels in broadcasting and simplex mode, and it is utilized for identifying and repairing mistakes without a reverse channel. It is the most widely used, but it is also widely utilized in IoT and wireless communication, where reversible connections are permitted. The FEC may be used to recover the lost bit. FEC, on the other hand, is not the sole method needed for the error control system. It often works in conjunction with another system, such as ARQ, as described in [14]. Small mistakes are corrected without retransmission, while major errors are corrected using a retransmission service. Convolutional codes and block codes are two of the most important types of FEC. Low network bandwidth is a disadvantage of FEC and ACK [15].

## 2.2 ARQ Scheme

The ARQ protocol retransmits lost and missing packets using acknowledgment (ACK) [5] and negative acknowledgment (NAK or NACK) [16]. This technique is utilized in our protocol for error detection and correction. However, it is a linear combination of packets that consists of portions of the original packets. It is important to note that the FEC works by injecting controlled redundancy into the original message before transmission. This redundancy is used instead of retransmission to restore any lost, dropped, or missing packets at the receiver. As a consequence, the IoT is using the fundamental principles of these two protocols. In this paper, lost packet, we consider that lost packets are not recovered by FEC but by ARQ. Both ARQ and FEC techniques are intended to be the primary methods for detecting and correcting errors in wireless data communication. However, this post considers only the FEC method for packet loss recovery.

## 2.3 The Proposed PBNC Purpose

Several authors use a novel method incorporating ACK to reduce the error and loss impacts in wireless Internet applications. They offer an overview of the IoT, focusing on additional protocols and application problems. When reception conditions deteriorate, numerous retransmissions are required to complete the data transfer. They cause a substantial increase in latency as well as a reduction in network capacity. The three basic metrics that may enhance IoT network performance are recovering packet loss, increasing capacity, and reducing retransmission numbers. In all situations, NC has shown more than double performance gains in all transmission scenarios. The

proposed PBNC enables IoT devices to offer dependable service for mission-critical applications such as SCADA [17] and telemetry for health and medical information provided from the body and various other smart critical uses in IoT. PBNC is also used to quickly assess IoT device network performance, namely the proper reception of data packets by IoT devices. Evaluating the basic parameters discussed in this article aims to improve network performance's objectivity and systematism. They also provide the suggested PBNC protocols a quick and simple picture of the IoT device network, such as the number of transmissions and retransmissions, network capacity, average packets size, and modulation type with their rates.

### 3 Issue and Solution Statement

The challenge of generating an appropriate method for packet loss repair in IoT is presented in this section.

#### 3.1 Issue Statement

FEC, in concept, enables IoT devices to identify and fix mistakes with retransmission, thus increasing data transmission efficiency. However, since it transmits the code to the IoT device regardless of whether it is accurate, it necessitates a lengthy error correction code, which adds unnecessary cost in low error situations. FEC substantially reduces packet loss; nevertheless, the latency and throughput become disproportionately lengthy and poor. Furthermore, FEC coding has a poor coding rate, implying ineffective channel exploitation; furthermore, it cannot cancel or reduce jitter except for out-of-order packets, which are frequent in the Internet network. The gain may come at the cost of network capacity or the size of the average packets to send a single packet from RTUs to the base station as a radio modem. The FEC protocol has a much higher bandwidth cost. Because the duplicated information in the transmission process according to a predefined algorithm also puts a greater computational load on the receiving device. ARQ necessitates using a reverse channel to transmit ACKs/NAKs; this results in low transfer speeds due to poor signal conditions and delay fluctuations due to the retransmitted data. The proposed PBNC compares the effect of ACK when recovering lost packets with ARQ.

#### 3.2 Proposed Solution

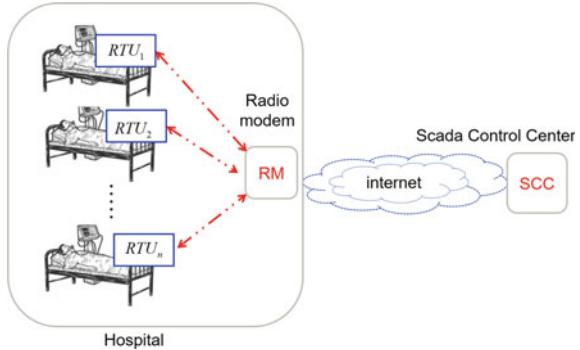
The proposed approach decreases the number of the transmission while recovering lost and dropped packets using RNC. It has the potential to enhance network performance significantly. However, particularly in practice, the proposed PBNC seems

to be quite straightforward and traditional. But, exploring the suggested RNC-based protocol in the IoT is both feasible and profitable. Its performance improves the point where it outperforms RNC's functional exploitation. The RLNC technique may offer various practical and feasible benefits to the wireless network, such as reducing the number of retransmissions, identifying and repairing missing and lost packets, lowering latency, and improving the network capacity. The suggested technique protects medical information against transmission problems, while the average packet is delivered to the destination. In the case of fixed-length packets, the body or data of the packet may be padded with additional data to make it the correct length. With PBNC, the ARQ does not provide enough time to ensure that a packet is really lost and not just out of order.

## 4 A System Model for the Medicine Application

We consider the scenario in which the  $RTU_i$  wishes to send medical packets to  $RM$ . The RTUs, in other terms, serve as the primary source of medical information in this study, and that each RTU may only send one packet sequentially at a time. Assume the following monitoring system architecture: IoT RTUs are wirelessly connected and interacted with the RM through the proposed PBNC, as illustrated in Fig. 1. This monitoring system model aims to offer medical data correctly received by the SCADA control center (SCC). This architecture is required when several patients provide certain kinds of medical information through sensors and medical devices. Consider a hospital consists of  $n$  patients. The medical devices of each patient are connected to RTUs labeled  $RTU_1, RTU_2, RTU_3, \dots$ , and  $RTU_n$ . Each RTU receives  $k$  packets from patient medical devices and then transmits them to  $RM$  as shown in Fig. 1. SCC software provides remote web-based access to real-time medical information via RTUs using a web browser and wireless networks. The  $RTU_i$  permanency collects medical information from each patient through medical devices. When a packet request arrives at a  $RTU_i$  hospital, they directly transmit their data to the RM (the intermediate node between the patient and SCC). Since the RM is based on modern technology, it can then deliver Ethernet interface speed, or serial (COM) interface baud rate [bps], which vary between 2400 kbps until 115,200 kbps. The  $RTU_i$  that receives the request is the coordinating point since it controls its life cycle and assembles the answer. This medium may be a node solely devoted to organizing requests or one of the hospital's data RTUs. This medium may be a node solely devoted to organizing requests or one of the hospital's data nodes. Assume that every  $RTU_i$  in the hospital transmits one packet, which is  $p_1, p_2, p_3, \dots, p_n$ . In a single time step, an RM from  $RTU_1, RTU_2, RTU_3, \dots$ , and  $RTU_n$ . In principle, the research objective describes the medical data collection that exchanges data with the free error between the RTU and patient and then sends them to the SCC via the Internet using NC. The receiving must acknowledge each packet sent from RTU using the extremely short service packet (ACK) to signal that it has successfully received it. If no ACK is reached, RM will use the retries option to resend the packet: The

**Fig. 1** Hospital with patients wirelessly connected to SCC via RM using Internet



acknowledgment/retransmission mechanism is built into the protocol and operates independently of any higher-level retries. The BPNC transmission buffer is in charge of data that is waiting to be sent. Its size is determined by the number of records (queue length) and the overall storage space (queue size) required. Records are stored in a queue, which becomes full when the queue length or the queue size is achieved. When the queue is filled, new arriving packets are not accepted.

## 5 Proposed PBNC Description

### 5.1 PBNC Description

Consider that the topology in Fig. 1 is well respected. The  $RTU_1, RTU_2, RTU_3, \dots$ , and  $RTU_n$  wish to disseminate  $n$  packets  $p_1, p_2, p_3, \dots, p_n$ , arriving from patient medical devices. The  $RTU_1, RTU_2, RTU_3, \dots$ , and  $RTU_n$  are in the communication range of the  $RM$  and might transmit the packets  $p_1, p_2, p_3, \dots, p_n$ . The  $RTU_1$  transmits the packet  $p_1$ , but only the IoT  $RM$  receives it.

Afterward, the  $RTU_2$  transmits the packet  $p_1$ , but also at this time, only the IoT  $RM$  correctly receives  $p_2$ . Then, the  $RTU_3$  transmits the packet  $p_3$ . This process continues until  $n$  packets are sent, and so on until the packet's transmission number is  $n$ , and finally  $p_n$  is correctly received by  $RM$ . Thus, IoT  $RM$  receives a different packet. However, improper propagation circumstances often need an extra number of transmissions to properly transfer the data, resulting in a substantial increase in latency and a significant reduction in throughput. Conventional ARQ protocols [18], which use the SoAM with one acknowledgment for each packet, render traditional schemes ineffective for lost packet recovery and energy saving between IoT  $RM$  and mobile IoT  $RTU$ .

In addition, as we all know, the retransmission system detects and recovers missing packets. In the typical situation shown in Fig. 1, the  $RTU_1, RTU_2, RTU_3, \dots$ , and  $RTU_n$  retransmit packets  $(p_1^1, p_2^1, \dots, p_n^1)$ ,  $(p_1^2, p_2^2, \dots, p_n^2)$ ,  $\dots$ ,  $(p_1^k, p_2^k, \dots, p_n^k)$ . If

these retransmissions are successful, IoT RM will receive the lost packet ( $p_1^1, p_2^1, \dots, p_n^1$ ), ( $p_1^2, p_2^2, \dots, p_n^2$ ) so forth until ( $p_1^k, p_2^k, \dots, p_n^k$ ). Hence, for *IoTRM* to successfully receive  $n$  packets, we require a total of  $n$  transmissions and  $(k - 1) \times n$  retransmissions, which implies that we need  $k \times n$  transmissions and retransmissions. The redundancy provided by FEC enables them to identify a limited number of mistakes that may occur anywhere in the message packets and, in many cases, correct these errors without the need for retransmission procedures. Because FEC transmits the code to the receiver, regardless of whether it is accurate or not, it requires a lengthy and strong error correction code and the repair of error patterns. As a result, FEC reduces packet loss while increasing latency and capacity. Instead of retransmitting duplicate packets:

$$\sum_{i=1}^n \sum_{j=1}^k p_i^j \quad (1)$$

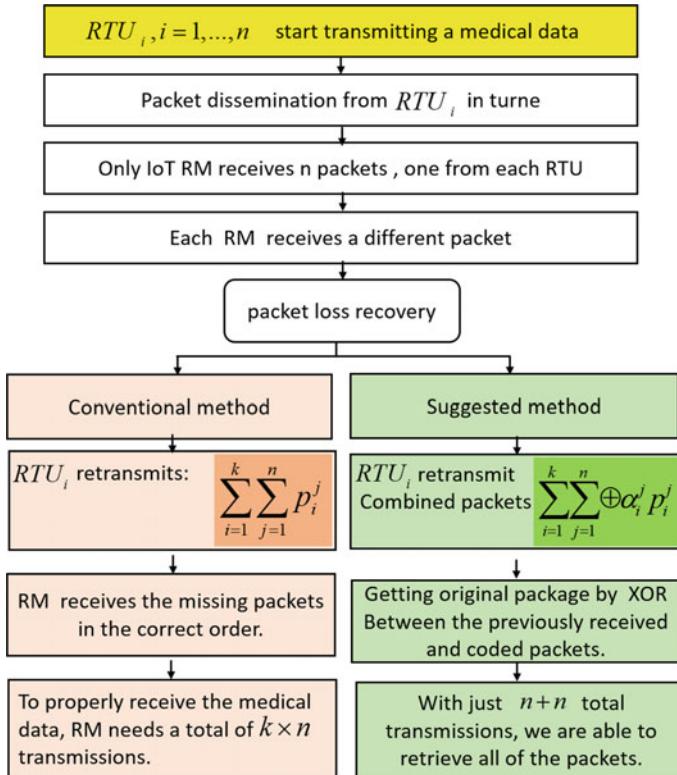
the  $RTU_i$ ,  $i = 1, 2, \dots, n$  in the proposed protocol, retransmits the coded packet in the following format:

$$\sum_{i=1}^n \sum_{j=1}^k \oplus \alpha_i^j p_i^j \quad (2)$$

where  $\alpha_i^j$  are finite field coefficients corresponding to the  $j$ -th packet of the  $i$ -th patient, with  $j = 1, 2, \dots, k$ , and  $i = 1, 2, \dots, n$ . The field in this paper is used by the network.

As a result, we need a total of  $n$  transmissions and retransmissions for IoT RM to successfully receive  $n$  packets, which means we need  $n + n$  transmissions and retransmissions. After receiving this packet, we utilize the XOR operator to combine the previously received packets with the coded packet. This operation allows us to extract the original packets. We find that as the number of packets  $k$  increases, the suggested protocol becomes more robust and efficient than the alternatives. The suggested protocol's operating stages are shown in Fig. 2.

The proposed protocol initializes its step with the star transmitting medical data from  $RTU_i$ ,  $i = 1, 2, \dots, n$ . The dissemination of the packet took turns transmitting  $RTU_1$ ,  $RTU_2, \dots$ , and  $RTU_n$  then back to  $RTU_1$ . All transmitted packets are received only by RM, which can get one packet from each RTU. The conventional packet loss recovery method using repetition but without NC needs a total  $k \times n$  transmission and retransmission. In contrast, with the suggested method, we need only  $n + n$  transmission and retransmission.



**Fig. 2** Operational phases of the protocol are proposed

## 5.2 Simulation Model Parameters

The RM and RTU are separated in the system architecture (Fig. 1) by variable Kbit/s connections (ranging from 150 to 160 Kbit/s). The packet size changes again (packet size goes from 0 to 1500 bytes), and each connection may create a propagation and processing delay of 10  $\mu$ s when used consecutively. Assume that the RTUs directly begin forwarding after receiving the packet's concluding bit and that the queues are empty. The simulation parameters are analytically expressed in Table 1.

As in the case of IoT wireless communications, we predict that packet loss may happen. In this respect, retransmission is started if a message is not delivered properly. This process is continued until the transmission is successful. We are presuming that subsequent broadcasts are self-contained. MATLAB is utilized to optimize a suggested technique and define the PBNC objective function in this work. The rice distribution is a suitable choice since the direct routes utilized between the RTUS and RM (i.e., line-of-sight, LOS) are often accessible.

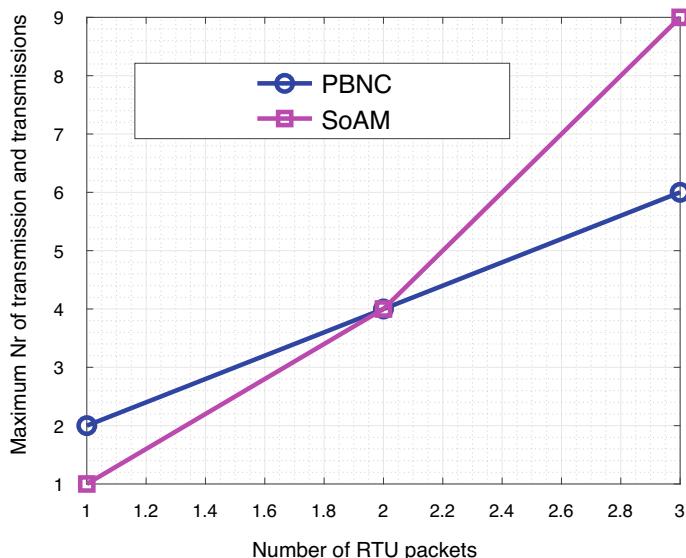
**Table 1** Values of the simulation parameters

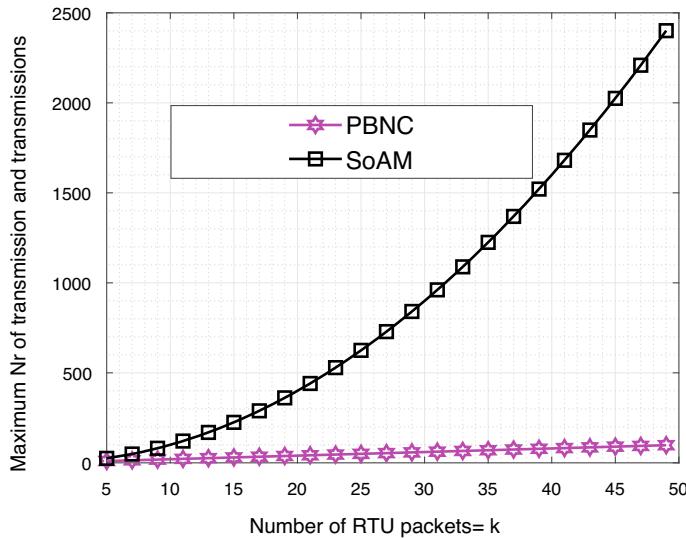
Parameters	Condition selected
Maximum number of retransmissions	400
Average message data length	500 bytes
Kbit/s links	15–160 Kbit/s
Number of packets	50
ACK	On/off
Modulation rate	20 Kbit/s

## 6 Results and Analysis

According to the findings, the BPNC may substantially decrease the number of retransmissions and increase network capacity. These are the most basic characteristics used to improve wireless communication performance. Figure 3 addresses the limitations of the proposed BPNC analysis as a method for the values of  $k$  analysis

In addition, as shown in Fig. 3, our suggested protocol remains ineffective for a value of  $k$  less than or equal to two ( $k \leq 2$ ). Once the value of exceeds 2 ( $k \geq 2$ ), the efficiency and robustness of the proposed protocol increase as the number of packets coming from each RTU increases. Figure 4 shows how the number of retransmissions and transmissions has changed as the RTU packets have increased.

**Fig. 3** Limitations of BPNC protocol analysis



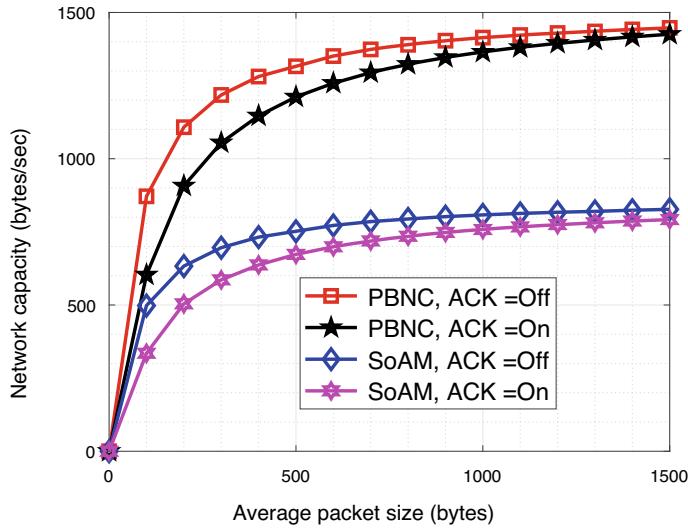
**Fig. 4** Evolution of the number of retransmission and transmission according to the increase in packets from RTU

**Table 2** Reduction in the number of transmissions and retransmissions

Number of packets from each RTU	19	29	39	49
BPNC (%)	89.47	93.1	94.87	95.92

Figure 4 clearly illustrates the robustness and efficiency of the proposed BPNC; the simulation study shows that the BPNC outperforms the SoAM in terms of a maximum number of transmission and retransmissions. The lost packets can be recovered using the SoAM, but in return, it necessitates many transmissions and retransmissions numbers. While for BPNC, the packets are recovered with only the minimum number of transmission and retransmission, reducing greatly this number. The percentage results of transmission and retransmission number reduction versus a varied number of packets are provided in Table 2.

The simulation findings indicate that the number of transmissions for 19, 29, 39, and 49 packets in a wireless hospital network drastically decreased by 89.47%, 93.1%, 94.87%, and 95.92%, respectively; therefore, the BPNC offers the best performance when the message packet increases. In Fig. 5, we compare the BPNC with ACK. The results without ACK show that the capacity increases when the average packet size increases. The capacity rises significantly from zero to 700 kbps for average packets size until 500 and then increases slowly from 700 to 820 kbps for an average packets size ranging from 500 to 1500 bytes.



**Fig. 5** Impact of ACK on network capacity with average packet size

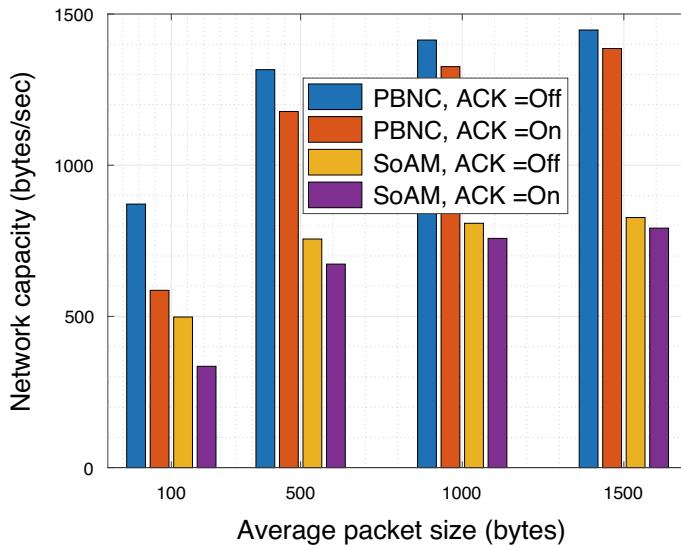
**Table 3** Network capacity for different average packets size

Average packets size (bytes)	ACK	400	800	1200
Proposed protocol (BPNC)	Off	1281	1389.5	1429.8
	On	1114.7	1286.2	1356.2
SoAM	Off	732	794	817
	On	637	735	775

It can be observed that network capacity without any correction is better than the performance of ACK. Furthermore, consequently, ACK significantly reduces the network capacity of unrecoverable packet loss. In addition, Fig. 5 illustrates the impressive improvement in network capacity of the proposed BPNC. Table 3 summarizes the comparison results of network capacity (kbps) for different average packets size (bytes) with and without ACK using a low fixed modulation rate (21 kbyte/s).

Comparison results in Table 3 prove that the proposed protocol outperforms SoAM by 75% with or without ACK. We note that the network capacity progressively decreases when compared to the presence of ACK. For BPNC, Fig. 6, the ACK has 5% less than SoAM.

The simulation results derived from Fig. 6 indicate that the PBNC is a powerful and flexible protocol for improving the performance of wireless medical networks by the cooperative RM using average packets size equal to 400, 800, and 1200



**Fig. 6** Comparison between the PBNC and the SoAM using ACK

bytes. Average packet size can play a fundamental role in varying network capacity, especially when the transmitted packets come from patients (remote) to RM (center). In this phase, this factor can significantly affect network capacity.

## 7 Conclusion

A BPNC for identifying and recovering missing packets is presented in this article. The comparison results indicate that the suggested network outperforms the SoAM, demonstrating great efficiency in the medical IoT network. This protocol may also enhance the performance of wireless network IoT devices, boosting network capacity and reducing the number of retransmissions than most current error-correcting methods. In the future, we want to expand the scope of our BPNC in the field of agriculture.

## References

1. Ahlswede, R., Cai, N., Li, S.-Y.R., Yeung, R.W.: Network information flow. *IEEE Trans. Inf. Theory* **46**(4), 1204–1216 (2000)
2. Alabady, S.A., Salleh, M.F.M., Al-Turjman, F.: LCPC error correction code for IoT applications. *Sustain. Cities Soc.* **42**, 663–673 (2018)

3. Hammouti, M., Ar-reyouchi, E.M., Ghoumid, K.: Power quality command and control systems in wireless renewable energy networks. In: International Renewable and Sustainable Energy Conference (IRSEC), pp. 763–769. IEEE, Marrakech (2016). <https://doi.org/10.1109/IRSEC.2016.7983989>
4. Ho, T., et al.: A random linear network coding approach to multicast. *IEEE Trans. Inf. Theory* **52**, 4413–4430 (2006)
5. Kotuliaková, K., Šimlašťíková, D., J. Polec, J.: Analysis of ARQ schemes. *Telecommun. Syst.* **52**(3), 1677–1682 (2011)
6. Faheem, M., Butt, R.A., Raza, B., Ashraf, M.W., Ngadi, M.A., Gungor, V.: Energy efficient and reliable data gathering using internet of software-defined mobile sinks for wsns-based smart grid applications. *Comput. Standards Interfaces* **66**, 103341 (2019)
7. Macher, G., Diwold, K., Veledar, O., Armengaud, E., Römer, K.: The quest for infrastructures and engineering methods enabling highly dynamic autonomous systems. In: Walker, A., O'Connor, R., Messnarz, R. (eds.) *Systems, Software and Services Process Improvement. EuroSPI 2019. Communications in Computer and Information Science*, vol. 1060. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-28005-5\\_2](https://doi.org/10.1007/978-3-030-28005-5_2)
8. Chatei, Y., Ghoumid, K., Ar-reyouchi, E.M.: Narrow-band channel spacing frequencies metric in one-hop wireless mesh networks. In: 2017 2nd International Conference on Communication and Electronics Systems (ICCES), pp. 296–301 (2017)
9. Ghasempour, A., Moon, T.K.: Optimizing the number of collectors in machine-to-machine advanced metering infrastructure architecture for internet of things-based smart grid. In: 2016 IEEE Green Technologies Conference (GreenTech), pp. 51–55 (2016)
10. Ar-Reyouchi, E.M., Maslouhi, I., Ghoumid, K.: A new fast polling algorithm in wireless mesh network for narrowband Internet of Things. *Telecommun. Syst.* **74**, 405–410 (2020). <https://doi.org/10.1007/s11235-020-00671-z>
11. Al-Sarawi, S., Anbar, M., Alieyan, K., Alzubaidi, M.: Internet of Things (IoT) communication protocols: review. In: 2017 8th International Conference on Information Technology (ICIT), pp. 685–690 (2017) (Amman 2017)
12. Ar-Reyouchi, E.M., Lamrani, Y., Benchaib, I., Ghoumid, K., Rattal, S.: The total network capacity of wireless mesh networks for IoT applications. *Int. J. Interact. Mobile Technol. (IJIM)* **14**(8), 61–75 (2020). <https://doi.org/10.3991/ijim.v14i08.12697>
13. Chen, D., et al.: Interleaved FEC/ARQ coding for QoS multicast over the internet. *Can. J. Electr. Comput. Eng.* **29**(3), 159–166 (2004)
14. Ar-Reyouchi, E.M., Chatei, Y., Ghoumid, K., Lichioui, A.: The powerful combined effect of forward error correction and automatic repeat request to improve the reliability in the wireless communications. In: International Conference on Computational Science and Computational Intelligence (CSCI), pp. 691–696. (2015). <https://doi.org/10.1109/CSCI.2015.39>
15. Ar-Reyouchi, E.M.: Analysis of the new method for assessing the total network capacity of wireless mesh networks for IoT devices. *J. Internet Technol.* **21**(4), 1025–1035 (2020). <https://doi.org/10.3966/160792642020072104012>
16. Lin, S., Daniel Costello, D.J.: *Error Control Coding: Fundamentals and Applications*. Chapter 15, Prentice-Hall (1983)
17. Rezai, A., Keshavarzi, P., Moravej, Z.: Key management issue in SCADA networks: a review. *Eng. Sci. Technol. Int. J.* **20**(1), 354–363 (2017)
18. Maslouhi, I., Ar-Reyouchi, E.M., Ghoumid, K., Baibai, K.: Analysis of end-to-end packet delay for Internet of Things in wireless communications. *Int. J. Adv. Comput. Sci. Appl.* **9**(9), 338–343 (2018)

# Toward Data Visualization and Data Forecasting with COVID-19 Vaccination Statistics



Vaishnavi Kulkarni, Jay Kulkarni, and Anurag Kolhe

**Abstract** The world runs on data. Various organizations, businesses, and institutions utilize and generate data. This information is a valuable commodity if availed of in the right way. Big data can be large and incomprehensible on its own, but when analyzed computationally, it can be a powerful tool for revealing patterns and trends, forecasting future values of certain data parameters as well as providing clarity about the metrics in the data. Data visualization and forecasting using such data are fields that have applications in every sector—from information technology, to education, to healthcare. Since the world was hit by the debilitating COVID-19 pandemic in 2019, life has become a blur of statistics—daily new case counts, daily deaths and recoveries, number of people vaccinated, etc. Such data are of paramount importance to everyone affected by the pandemic, and presenting it in a way that is easily understandable to a layperson and using it to glean insights into the spread and curb of the disease as well as the efficacy of the vaccines is necessary. This paper takes COVID vaccination statistics as a use case for the fields of data visualization and data forecasting. It elucidates the methodology and benefits of both interactive visualizations of vaccination data and forecasting future trends in vaccine and case metrics based on data over time.

**Keywords** Data visualization · Data forecasting · COVID vaccinations · Polynomial regression · Machine learning · Data science

## 1 Introduction

With the exponential growth of the Internet, tremendous volumes of data are getting generated each day. According to a Forbes article [1], 2.5 quintillion bytes of data

---

V. Kulkarni (✉) · J. Kulkarni  
Pune, India  
e-mail: [vaishnavikulkarni@hotmail.com](mailto:vaishnavikulkarni@hotmail.com)

A. Kolhe  
Akola, India

are created each day and this pace is increasing every day. This data are related to professional fields like medicine, technology, science, communication, engineering, etc., along with other fields like social media, news, sports, etc. It is generally difficult to extract useful information from these heaps of large data. Hence, to make sense of these large volumes of data present, data visualization comes into the picture. Data visualization is the graphical representation of any given information. It gives us a clear picture of what the data are representing by using visual elements like graphs, maps, charts, etc. Properly organizing all the data based on some parameters into graphs or charts helps us to comprehend the meaning of it. Further, it proves beneficial for deriving trends and detecting anomalies in the available information.

Today, when we talk about the tremendous volumes of data, big data are the word that comes first into the picture. Big data refer to the flood of structured and unstructured digital data that are gathered by organizations day in and day out. It can be characterized by the 5 versus of the big data—volume, velocity, variety, veracity, and value. Volume is the volume of produced data. Velocity is the pace of generating and moving data. Variety denotes the diverse type of data collected from varied sources. Veracity is the credibility of the collected data. Finally, value refers to the importance to the data to company's business.

The traditional storage system cannot be used to store big data. Large storage space is required to store such gigantic data. Similarly, powerful processing systems, analytical tools, databases for quick access of information, etc., are essential when dealing with big data. This is where cloud computing comes into the picture. Renowned cloud services such as Azure, AWS, and GCP provide the user with servers, databases, storage, networking, analytics, etc. The cloud infrastructure allows for real-time processing of big data. Data could be efficiently and quickly stored in the storage facilities of cloud services. Similarly, data could be retrieved and looked upon quickly and further could also be interpreted in real-time. Data are fragmented and stored securely on these cloud platforms. This fragmentation process is, however, dispersed and scattered, thus lacking proper order. This increases the time in information collection. In [2], hybridized historical aware algorithm (HHAR) is used to minimize the dispersed and scattered packages.

In the world of big data and data science, data visualization is highly significant. Machine learning and deep learning models require a huge amount of data for training and testing purposes. Without initially learning from the immense data, it will not perform as expected and give the required output. Initially, before building any model, we need to analyze the available data over which we will be using different algorithms. This analysis of data is complex and is not possible manually. Data visualization tools can help solve this quandary and help us with the computational analysis of data to present it in a way that is easily understood by the human eye. Real-world data, though complex, may not be outputted in 2- or 3-dimensional spaces. [3] Hence, dimensionality reduction techniques like PCA, TSNE, LDA, etc., could be used to reduce higher dimensional data into 2D or 3D. Before building the models, visualizing this data into graphs may help us to reveal some hidden patterns, clusters that could be formed within the data, etc. We can also figure out whether the data are

linearly separable, or if it overlaps too much, etc. This initial analysis is thus crucial for selecting a particular ML model for a given set of data.

Data forecasting is another area that can exploit ML techniques with the availability of tremendous data. In data forecasting, ML models are trained using past data to analyze the future trends of upcoming data. This is helpful in areas like sales marketing, climate changes, etc. In sales for example, with the help of historical data, predictions like whether a stock price will fall or rise can be done. The more accurately an ML model is trained, the more accurate its forecasting. [4] Forecasting can be classified as time series forecasting and time series classification. In time series forecasting, techniques are used for predicting future values using an ML algorithm. Whereas in time series classification, by looking at past data, techniques are used to classify an item into a particular class.

In 2019, the entire world was hit by the coronavirus pandemic which affected many people, resulting in high mortality rates. It also crippled the economies of many countries. Vaccines have now been administered globally. However, the efficacy of the vaccines can only be decided by analyzing the past data. In this paper, we will be analyzing the data globally for the COVID-19 pandemic. Visualizing this data in the form of charts and graphs helps reveal how inoculation against the disease is proceeding worldwide. Furthermore, we have used the polynomial regression technique to forecast the trend of the number of new COVID cases based on the number of vaccinations.

## 2 Related Work

For United States, the authors in [5] have proposed a model to accurately forecast the COVID-19 cases and deaths.

The authors did not use a traditional single model to predict the required trend by keeping the parameters fixed. They instead used a “last-fold partitioning” which is a general learner. This gave them the best parameters for their model, the best combination of features, and the forecasting model with best history-length.

Supervised learning ML models are used in [6] to predict an unknown input instance. Learning methods use regression techniques and classification algorithms for predictive models’ development. Four regression models are used to study the proponents of COVID-19. This helps in the forecasting of factors like recently contracted cases, the mortality rate as well as the count of cured cases over a time frame of almost a fortnight.

In [7], a new prediction model is proposed by the combination support vector regression (SVR) model with conventional random vector functional link (RVFL) model to improve the prediction capabilities of COVID-19 cases. RVFL network is hybridized with 1D discrete wavelet transform, and a wavelet coupled RVFL (WCRVFL) network is proposed.

The study in [8] aims to do a time series forecasting to predict the COVID-19 patients’ rise, recovery, and death in India based on the daily data obtained from

the Indian Government. It uses a statistical model called autoregressive integrated moving average (ARIMA) which has proven effective in short-term forecasting in many other diseases. ARIMA models aim to describe the autocorrelations in the data. ARIMA is also combined with an exponential smoothing model which is based on a description of the trend and seasonality in the data.

In [9], the authors rank different ML classification algorithms with the help of the COVID-19 World Vaccination Progress dataset. Four classification algorithms are considered—decision tree, K-nearest neighbor, random tree, and Naïve Bayes. For this comparison, an open-source Java platform WEKA is used. WEKA contains a series of ML algorithms that enable researchers to analyze their data for patterns, trends, etc. After running the algorithms over a dataset consisting of 6745 instances and more than 14 attributes, findings are recorded. The results show that the decision tree classifier's percentage of correctly classified instances was highest among the four. And that for Naïve Bayes was lowest. Accordingly, the root mean square error (RMSE) is lowest for the decision tree classifier, thus making it the best classification algorithm among above 4. Similarly, for Naïve Bayes, the RMSE value was highest.

In [10], the authors predict the death rates as well as survival rates of COVID-19 patients. Supervised machine learning is used to achieve this. These rates are forecasted by considering the effects of chronic diseases, features unique to particular groups of people belonging to similar age-groups, ethnicities, etc., as well as the initial data from clinical trials. COVID-19 samples from the King Fahad University Hospital, Saudi Arabia were used as the key dataset for this study. Patient records were classified into 2 classes—survived and deceased. Dataset is classified on various features like body temperature, shortness of breath, chronic disease like diabetes, etc. As the mortality rate was low, the deceased class contains fewer values and thus causes class imbalance. To resolve this, the synthetic minority oversampling technique (SMOTE) was used. The authors employed three classification algorithms—logistic regression (LR), random forest (RF), and extreme gradient boosting (XGB). For parameter optimization, the grid technique was used. From the results, the authors concluded that the random forest algorithm was the most effective. It performed much more efficiently compared to the other classifiers using the top 20 features with SMOTE data. On the other hand, logistic regression gave the least accurate results.

In [11], authors have used logistic regression for forecasting the probability of recovery for patients with debilitating COVID symptoms. They have used data of 183 patients from Tongji Hospital, Wuhan. Four variables—lymphocyte count, age, d-dimer, and high-sensitivity C-reactive protein were selected and used to fit the logistic regression model. The areas under the receiver operating characteristic curves (AUROCs) in the logistic regression model were 0.895. Other models have also been considered in this paper but the logistic model ultimately decided upon due to it being minimalistic yet effective. For prediction of death of patients, the AUROC of the external validation set, its sensitivity, and its specificity all had values close to 0.8.

To forecast COVID-19 reproduction rate, research in [12] aims at the performance evaluation of various non-linear regression techniques, such as gradient

boosting, KNN, XGBOOST, SVR, and random forest regressor. It also highlights the importance of hyperparameters tuning and feature selection. For the feature selection, methods such as gradient boosting, GBOOST, and random forest are applied. Depending upon the score of feature importance, the top 7 features that affect the rate of reproduction are recognized. Further, four different experiments are conducted with the presence and absence of feature selection. Across all the experiments, RMSE, mean absolute error (MAE), and relative absolute error (RAE) were used to measure the reproduction rate. In all the experiments performed, KNN approach obtained altogether the best values for RAE, RMSE, and MAE. It was succeeded by the XGBOOST and random forest. Further, tuning of hyperparameter with all the features demonstrated a low prediction error rate and the best performance for these parameters was shown by random forest.

In [13], to evaluate the metrics of ML models in analyzing disease infection, an epidemiology Mexico dataset of COVID-19 cases is used. Different supervised ML algorithms are considered for this purpose. Eight clinical features and two demographic features are considered here. 1 is encoded for positive and 0 for negative. The correlation coefficient analysis is carried out to reveal each independent and dependent feature relationship. Here, also the performance of the models is again evaluated based on accuracy, sensitivity, and specificity. The findings show that among all the models, the highest 94.99% accuracy is achieved by the decision tree. In terms of sensitivity, SVM is better than the rest with a value of 93.34%, and Naïve Bayes achieves the highest specificity of 94.3%. Also, “age” is extracted as the most significant dependent feature by the decision tree.

Authors in [14] predict the number of daily confirmed cases of coronavirus after vaccination using 2 statistical models and a deep learning model—the autoregressive integrated moving average (ARIMA), the generalized autoregressive conditional heteroscedasticity (GARCH), and the stacked long short-term memory deep neural network (LSTM DNN). Dataset of WHO which is obtained from GitHub is used here. Every 3 models are applied to the dataset. The optimal hyperparameters for LSTM—that is the count of LSTM cells and blocks of cell—are obtained by a comprehensive search. When considered along the parameters of RMSE and MAE, the results of experiments based on these models show that LSTM DNN functions with the highest accuracy. Performances of ARIMA and GARCH depend on datasets.

Medical images could also be leveraged for data forecasting. Efficient image retrieval is crucial in such cases. In [15], a framework using adaptive state transition Kalman filtering technique is used to improve retrieval rate. It achieved a success rate of 96.2% of retrieval rate in image retrieval. In [16], classification of COVID-19 is done with the help of chest X-ray images of patients. Authors have used CNN with histogram-oriented gradients (HOG) as a methodology for feature extraction. The proposed CNN architecture consists of 5 layers. Max pooling layer has been used to reduce the space size with rectified linear unit ReLU as an activation function. The final flatten array gives categories to determine whether the result is COVID-19 positive or negative or pneumonia. For testing purposes, Cohen’s dataset with 400 positive COVID-19 X-ray images is used, and an accuracy of 93% was observed.

### 3 Data Visualization—COVID-19 Vaccinations

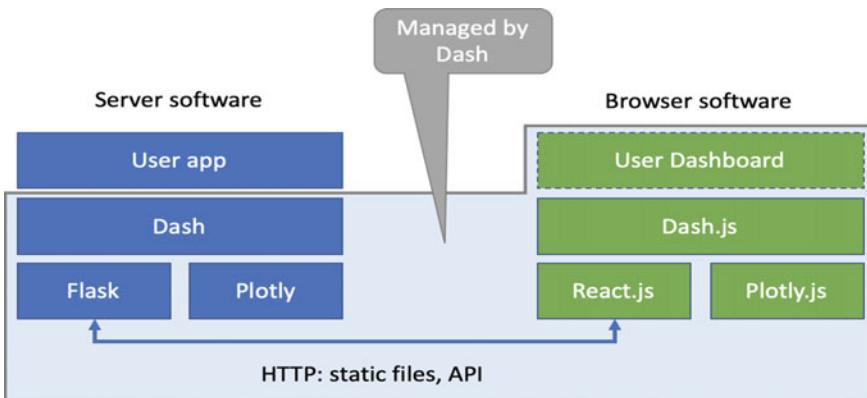
#### 3.1 Data Visualization Tools and Techniques

Businesses and organizations produce huge amounts of data every day. This data can be a powerful tool if leveraged correctly. For a person who is unfamiliar with data analytics, large datasets tend to be quite abstruse. Visualization techniques are employed to make datasets meaningful—some commonly preferred ones are bar graphs, pie charts, scatter plots, line graphs, histograms, Gantt charts, map visualizations, candlestick diagrams, etc.

Data visualization tools make it possible to create such charts and make big data easy to interpret and extract useful information from. Tools ranging from basic to advanced such as MS Excel, Google Charts, Tableau, Zoho analytics, and Datawrapper are available for generating relevant data visualizations [17]. Another important tool available is Dash. Dash is a Python framework that is useful for creating a more personalized, interactive dashboard application. Flask, Plotly.js, and React.js form the basis of this framework. Dash applications are comprised of two components, such as the layout and the callbacks. As Dash application development is done using Python, Dash is an ideal framework for tools that integrate visualization with forecasting and analytics.

#### 3.2 Use Case: COVID-19 Vaccination Dashboard

**Application Architecture:** We have built an interactive dashboard application that combines visualization as well as forecasting of COVID-19 vaccination statistics. The application is built using the Python Dash framework as shown in Fig. 1. With



**Fig. 1** Dash application architecture [9]

Plotly Dash, all of the user code for the dashboard is in Python. Dash is built on top of Flask (a micro-Web framework in Python) and serves the code over it using HTTP which is then deployed over a Web server like Nginx. Dash generates the necessary JavaScript code. The browser Web application is created and updated with the Web API generated by Dash [18]. Plotly is the library employed by the Dash framework to create interactive data visualizations.

**Datasets:** The application uses the vaccination datasets provided by our world in data. [19] This dataset uses official vaccination statistics sourced globally from governments and health ministries. The datasets are updated daily based on the latest information received from official sources, which enables our application to be updated daily with the latest statistics. Our dashboard employs two different datasets—one of global vaccination metrics and one of state-wise vaccinations in the United States of America. The latter replies on daily updates sourced from the United States Centers for Disease Control and Prevention.

The global dataset comprises of data with the following headers: location, iso\_code, date, total\_vaccinations, people\_vaccinated, people\_fully\_vaccinated, total\_boosters, daily\_vaccinations\_raw, daily\_vaccinations, total\_vaccinations\_per\_hundred, people\_vaccinated\_per\_hundred, people\_fully\_vaccinated\_per\_hundred, total\_boosters\_per\_hundred, and daily\_vaccinations\_per\_million. The US state dataset has the following data columns—date, location, total\_vaccinations, total\_distributed, people\_vaccinated, people\_fully\_vaccinated\_per\_hundred, total\_vaccinations\_per\_hundred, people\_fully\_vaccinated, people\_vaccinated\_per\_hundred, distributed\_per\_hundred, daily\_vaccinations\_raw, daily\_vaccinations, daily\_vaccinations\_per\_million, and share\_doses\_used.

Population estimates for per-capita statistics in the datasets are based on the United Nations World Population Prospects. Income groups are based on the World Bank classification.

**Data preprocessing:** The data used for the visualizations are read from the source datasets and stored as DataFrames. A DataFrame is a data structure of pandas, a powerful library for Python programming that facilitates data analysis and manipulation. DataFrames are two-dimensional, size-mutable structures containing tabular data (possibly of different types). [20] They contain labeled rows and columns. From the two DataFrames (of global and US state data), two dictionaries are created. A dictionary (dict) is a Python data structure with a key-value pair. The DataFrames are filtered by country or state, respectively, and a dict of DataFrames is formed with each DataFrame in the dict having its corresponding country or state as its key. These dicts are sorted based on the people\_fully\_vaccinated metric for the latest date in each DataFrame in descending order—i.e., the country or state with the highest number of fully-vaccinated people is first. Further, sorting and preprocessing are done based on the data required for each type of visualization—based on date, daily doses administered, etc.

**Visualizations:** The first section of the dashboard displays the global vaccination metrics using different graphics. Figure 2 shows the top of the dashboard with the different tabs as well as the total number of vaccinated people worldwide. This lets the user get an idea of the state of vaccinations on a global scale at a glance.

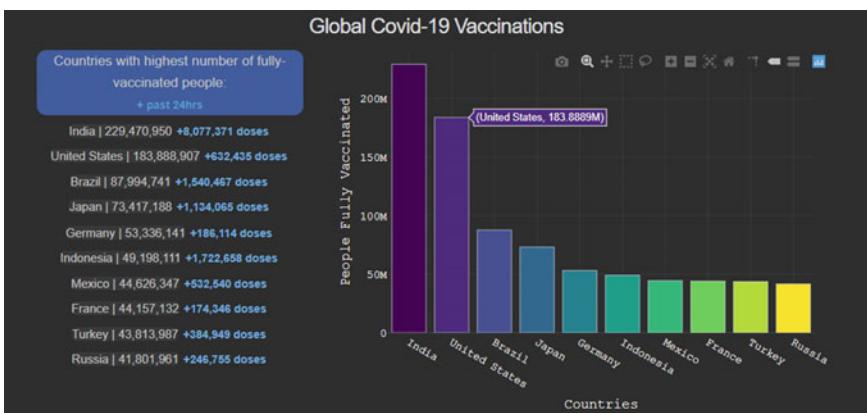


**Fig. 2** Dashboard vaccination statistics tab

The top 10 countries with the highest number of fully-vaccinated people are displayed next, with the latest updated count as well as the change in the count of doses administered since the last update (Fig. 3). This graphic enables the user to track the progress of the most highly-vaccinated countries on a daily basis and provides information about total vaccinations as well as the number of doses administered per day.

A line graph showing the daily dose count in these countries since December (based on the “date” and the “daily\_vaccinations” columns in the dataset) is the next graphic (Fig. 4). This shows the trend of vaccinations in these top ten countries, and the user can observe how daily vaccination rates have evolved over time—whether they are increasing, whether they peaked and then dropped, or whether they have been constant. This visualization can be interacted with by selecting a single country to display (Fig. 5), which enables the user to observe the vaccination trend for a particular country more closely.

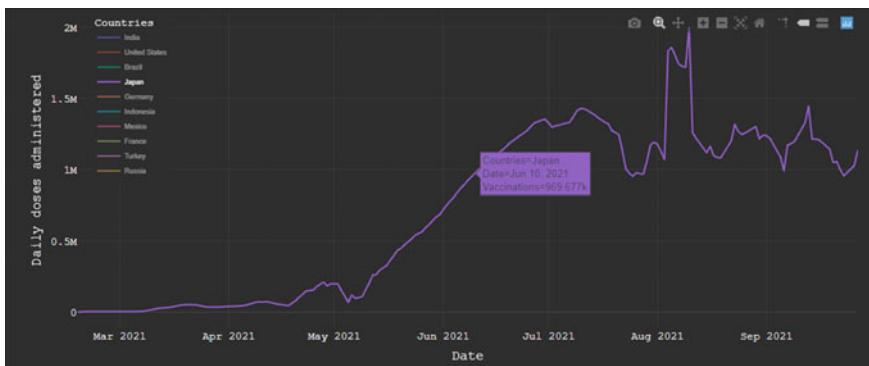
The second section of the dashboard is comprised of state-wise information for the United States of America (USA). As of November 2021, the USA has both the highest number of cases as well as the highest count of active cases globally [21].



**Fig. 3** Top 10 countries based on number of fully-vaccinated people



**Fig. 4** Daily vaccinations in top 10 countries



**Fig. 5** Single country selected

Thus, visualizations of the vaccinations administered in the states of the USA are useful for tracking which states are implementing vaccination measures well and which states are lacking in their efforts. This section of the dashboard thus consists of an infographic of the 20 states with the highest count of fully-vaccinated people (Fig. 6). An interactive scatter plot of the daily doses administered in a state over time is also displayed, based on the state selected from the table (Fig. 7 and 8). This scatter plot allows users to view and understand the progress of vaccination drives in the selected state over time, and to glean whether vaccination efforts are still being undertaken rigorously or whether they have stagnated.

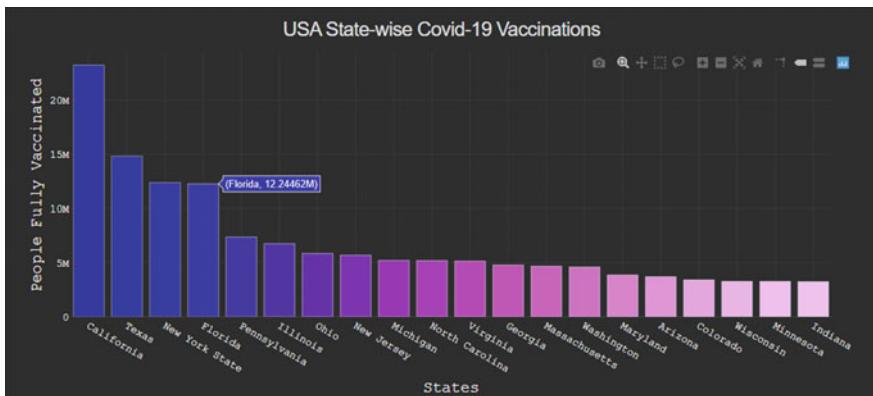


Fig. 6 Top 20 states based on number of fully-vaccinated people

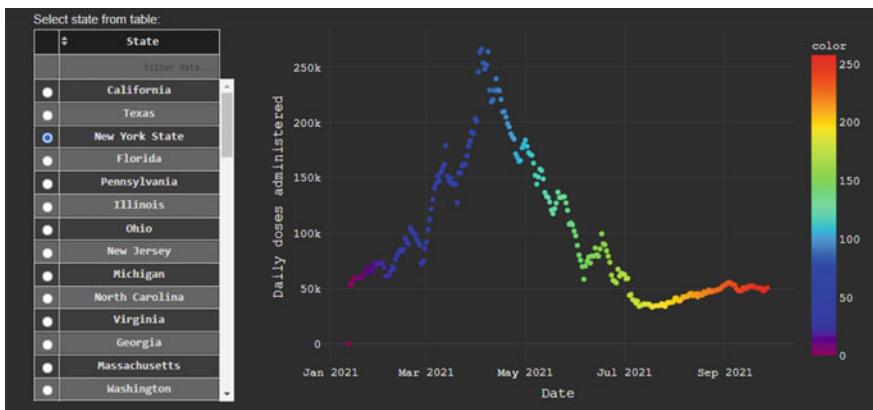


Fig. 7 The state of New York selected

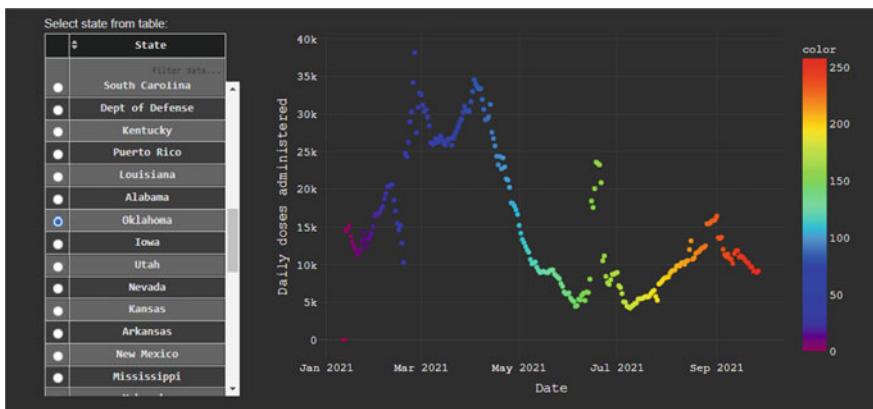


Fig. 8 The state of Oklahoma selected

### 3.3 Advantages and Future Scope

Data visualization is a powerful tool for data analysis and for building statistical models. Taking the use case of COVID vaccinations which we have detailed here, the spread of misinformation about the recently developed vaccines is rampant. People are hesitant to get vaccinated due to the unknown nature of the vaccines. In such a situation, clear and easily comprehensible visualizations of complex and opaque vaccination datasets can help give people an accurate idea of the situation of vaccination drives all over the world. Additionally, such a tool that is kept up-to-date with the latest information can be advantageous to healthcare providers for streamlining their logistics and operations. They can ensure that they are not short-staffed or falling short in supply of the vaccine by making an informed estimate with the help of easy-to-read, user-friendly visualizations.

The dashboard also integrates data forecasting of new COVID-19 cases using the vaccination data, which will be explained in the next section.

## 4 Data Forecasting—COVID Vaccinations Per Hundred Versus New Cases

### 4.1 Motivation

We attempted to model the relationship between the “people vaccinated per hundred” and “number of new cases.” We chose to analyze the relationship between these 2 factors as determining their relationship can prove to be extremely useful to forecast the trend in the number of new cases upon increasing the number of inoculations.

### 4.2 Algorithm

Various techniques such as regression and correlation are used to analyze the relationship between 2 and more variables. We chose to use the polynomial regression technique as it modeled the best relationship among all the techniques surveyed. According to [22], polynomial regression is a regression technique in which the relationship between the independent variable  $x$  and the dependent variable  $y$  is modeled as an  $n$ th degree polynomial in  $x$ . The equation can be modeled as follows wherein  $a_0, a_1, a_2, \dots, a_n$  are the coefficients of the regression equation.

$$y = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + \dots + a_nx^n \quad (1)$$

In our case, the independent variable  $x$  was “people\_vaccinated\_per\_hundred,” and the dependent variable  $y$  was “number\_of\_new\_cases.”

### 4.3 Implementation Details

**Dataset:** After analyzing numerous datasets, we chose to use the dataset made available for public use provided by [23]. Ritchie et al. [23] is one of the most comprehensive COVID datasets and contains extensive data which can be used for various kinds of research. The columns “people\_vaccinated\_per\_hundred” and “new\_cases” were used for the algorithm.

**Data Preprocessing:** The dataset was filtered, and data for India were collated. The data used were after the month of May as the number of vaccinations before May were not significant enough to model the relationship. The dataset contained incomplete data at numerous instances for both the columns mentioned above. The incomplete data were estimated using pandas interpolate method [24]. The interpolation technique used was “polynomial” of order 5.

**Polynomial Regression:** We used scikit-learn [25] for performing polynomial regression in Python. Scikit-learn has a linear regression module that can be used to train a polynomial regression function. PolynomialFeatures are a tool in scikit-learn which computes the desired degrees or powers of feature variables. It transformed the given feature variable into a matrix containing the degrees of the variables.

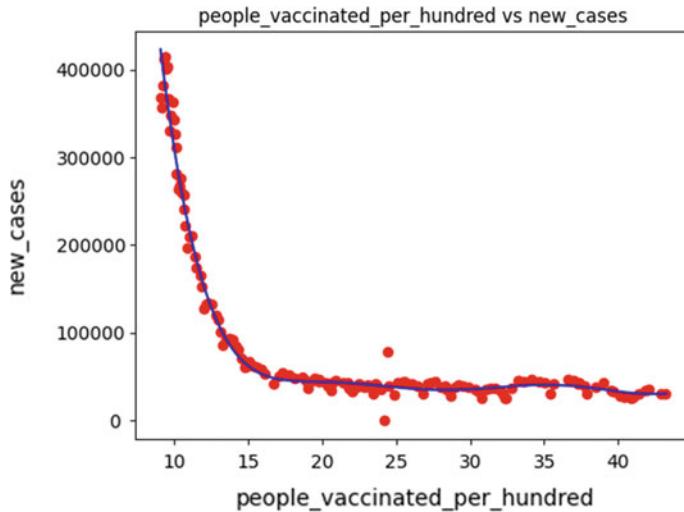
The model was trained using a random split of 80% training to 20% test data. Using the fit function, the polynomial features were fit into the polynomial regression model, and the equation of the regression was obtained. The degree of the polynomial regression was chosen to be 8 after extensive experimentations. The following is the equation of the regression where the variables are as defined in Sect. 4.2:

$$\begin{aligned} y = & -487053.151x - 4912.572x^2 + 5008.596x^3 \\ & - 414.69x^4 + 16.45x^5 - 0.355x^6 + 0.00403x^7 - 0.0000188x^8 \end{aligned} \quad (2)$$

Future values for the number of new cases can be predicted by entering the value of the people vaccinated per hundred in the above equation. A scatterplot of the regression function was plotted using matplotlib [26] as shown in Fig. 9.

### 4.4 Results

Using the predict method, the values for the test input data were predicted. There are various methods to evaluate regression models such as relative absolute error, range of prediction, mean absolute error, relative squared error, and coefficient of determination ( $R^2$ ). We chose to use the coefficient of determination denoted popularly as  $R^2$ , which is the amount of variation in the dependent variable [27] which can be predicted from the independent variable. The  $R^2$  score provides a measure of how well the variances of two variables are dependent on each other. E.g., if the  $R^2$  score of the regression is 0.50, then roughly 50 percent of the variation can be explained by the inputs.  $R^2$  was calculated between the predicted and the actual values for the



**Fig. 9** People\_vaccinated\_per\_hundred versus new\_cases

input data. It usually ranges from 0 to 1 with 1 being the highest possible score. The  $R^2$  score is calculated as follows:

$$R^2 = 1 - (\text{SS}_{\text{res}}/\text{SS}_{\text{tot}})$$

where  $\text{SS}_{\text{res}}$  is the sum of squares of residuals, also called the residual sum of squares:

$$\text{SS}_{\text{res}} = \sum (y_i - f_i)^2$$

$\text{SS}_{\text{tot}}$  is the total sum of squares (proportional to the variance of the data):

$$\text{SS}_{\text{tot}} = \sum (y_i - \bar{y})^2$$

$\bar{y}$  is mean of the observed data:

$$\bar{y} = 1/n \left( \sum y_i \right)$$

The  $R^2$  score of our algorithm was 0.9878. This shows how closely the variance of the independent variable—people vaccinated per 100 explains the variance of the dependent variable—no of cases.

Training	Test	$R^2$ score
80% of input data	20% of input data	0.9878

#### 4.5 Vantages and Future Scope

The above-mentioned forecasting model using polynomial regression can certainly prove to be beneficial in a lot of ways. The model can be used by the authorities to forecast the cases at a future point in time. This can assist them to manage the existing resources more efficiently and allow for better planning and allocation of resources. The forecasting model will also help the vaccine manufacturers to predict the efficacy of their vaccines.

The model can be further enhanced by expanding the dataset size and entering actual values in place of the missing values in the dataset. Other regression models such as Lasso and Ridge could also be employed to forecast the relation between the number of vaccinations and the number of cases. Various alternative outlier detection methods could also be used to remove the outliers present during the second wave.

### 5 Conclusions

Our data forecasting model predicts the future trend of new cases by taking the past data into consideration and is integrated with a visualization dashboard. This visualization of the data along with forecasting is highly useful to medical professionals. They can easily see the current and past trends as displayed by the dashboard by filtering through various parameters. Further, data visualization combined with data forecasting ensures that the efforts being taken behind vaccination drives are being channeled in the right direction. An optimistic trend resulting in fewer deaths after vaccinations makes a strong point for the government to encourage their citizens to get vaccinated. Vice-a-versa, if the trend is not as positive as expected, then the efficacy of the vaccine could be questioned, thus alerting a particular country to look for an alternative measure. Hence, relevant measures, steps, and decisions could be taken by the medical professionals as well as the government officials by utilizing this data visualization and forecasting.

## References

1. How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read, <https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=1503adeb60ba>. Last accessed 15 Sep 2021
2. Pandian, A.P., Simys, S.: Effective fragmentation minimization by cloud enabled back up storage. *J. Ubiquitous Comput. Commun. Technol. (UCCT)* **2**(01), 1–9 (2020)
3. Data Visualization using Python for Machine Learning and Data science, <https://towardsdatascience.com/data-visualization-for-machine-learning-and-data-science-a45178970be7>, last accessed 2021/09/16
4. ML time-series forecasting the right way, <https://towardsdatascience.com/ml-time-series-for-forecasting-the-right-way-cbf3678845ff>. Last accessed 19 Sep 2021
5. Ramazi, P., Haratian, A., Meghdadi, M., et al.: Accurate long-range forecasting of COVID-19 mortality in the USA. *Sci Rep* **11**, 13822 (2021). <https://doi.org/10.1038/s41598-021-91365-2>
6. Rustam, F., et al.: COVID-19 future forecasting using supervised machine learning models. *IEEE Access* **8**, 101489–101499 (2020). <https://doi.org/10.1109/ACCESS.2020.2997311>
7. Hazarika, B.B., Gupta, D.: Modelling and forecasting of COVID-19 spread using wavelet-coupled random vector functional link networks. *Appl. Soft Comput.* **96**, 106626–106626 (2020)
8. Darapaneni, N., Reddy, D., Paduri, A. R., Acharya, P., Nithin, H. S.: “Forecasting of COVID-19 in India using ARIMA model.” In: 2020 11th IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), pp. 0894–0899 (2020). <https://doi.org/10.1109/UEMCON51285.2020.9298045>
9. Abdulkareem, N. M., Abdulazeez, A. M., Zeebaree, D. Q., Hasan, D. A.: COVID-19 world vaccination progress using machine learning classification algorithms. *Qubahan Acad. J.*, **1**(2), 100–105 (2021). <https://doi.org/10.48161/qaj.v1n2a53>
10. Aljameel, S. S., Khan, I.U., Aslam, N., Aljabri, M., Alsulmi, E. S.: “Machine learning-based model to predict the disease severity and outcome in COVID-19 patients.” *Sci. Program.*, **2021**, 10 (2021) Article ID 5587188. <https://doi.org/10.1155/2021/5587188>
11. Hu, C., Liu, Z., Jiang, Y., Shi, O., Zhang, X., Xu, K., Suo, C., Wang, Q., Song, Y., Yu, K., Mao, X., Wu, X., Wu, M., Shi, T., Jiang, W., Mu, L., Tully, D.C., Xu, L., Jin, L., Li, S., Tao, X., Zhang, T., Chen, X.: Early prediction of mortality risk among patients with severe COVID-19, using machine learning. *Int. J. Epidemiol.* **49**(6), 1918–1929 (2020). <https://doi.org/10.1093/ije/dyaa171>
12. Jayakumar, K., et al.: Performance evaluation of regression models for the prediction of the COVID-19 reproduction rate. *Frontiers Public Health* **9**, 729795 (2021). <https://doi.org/10.3389/fpubh.2021.729795>
13. Muhammad, L.J., Algehyne, E.A., Usman, S.S., et al.: Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. *SN Comput. Sci.* **2**, 11 (2021). <https://doi.org/10.1007/s42979-020-00394-7>
14. Kim, M.: Prediction of COVID-19 confirmed cases after vaccination: based on statistical and deep learning models. *Sci. Med. J.* **3**(2), 153–165 (2021)
15. Dhaya, R.: Analysis of adaptive image retrieval by transition kalman filter approach based on intensity parameter. *J. Innovative Image Process. (JIIP)* **3**(01), 7–20 (2021)
16. Chen, J.-Z.: Design of accurate classification of COVID-19 disease in X-ray images using deep learning approach. *J. ISMAC* **3**(02), 132–148 (2021)
17. Stobierski, T.: “Top data visualization tools for business professionals.” 12 January 2021. [Online]. Accessed 6 Sep 2021
18. Plotly Dash or React.js + Plotly.js? A side-by-side comparison, <https://towardsdatascience.com/plotly-dash-or-react-js-plotly-js-b491b3615512>. Last accessed 10 Sep 2021
19. Mathieu, E., Ritchie, H., Ortiz-Ospina, E., et al.: A global database of COVID-19 vaccinations. *Nat. Hum. Behav.* (2021)

20. pandas.DataFrame, <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html>. Last accessed 11 Sep 2021
21. WHO Coronavirus (COVID-19) Dashboard, <https://covid19.who.int/>. Last accessed 01 Nov 2021
22. Polynomial Regression, [https://en.wikipedia.org/wiki/Polynomial\\_regression](https://en.wikipedia.org/wiki/Polynomial_regression). Last accessed 15 Sep 2021
23. Ritchie, H., Mathieu, E., Rodés-Guirao, L., Appel, C., Giattino, C., Ortiz-Ospina, E., Hasell, J., Macdonald, B., Beltekian, D., Roser, M.: “Coronavirus pandemic (COVID-19)”. In: Published online at OurWorldInData.org. Retrieved from: ‘<https://ourworldindata.org/coronavirus>’ [Online Resource]
24. Pandas, <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.interpolate.html>. Last accessed 16 Sep 2021
25. Scikit-learn, <https://scikit-learn.org/stable/>. Last accessed 16 Sep 2021
26. Matplotlib, <https://matplotlib.org/>. Last accessed 17 Sep 2021
27. Coefficient of determination, [https://en.wikipedia.org/wiki/Coefficient\\_of\\_determination](https://en.wikipedia.org/wiki/Coefficient_of_determination). Last accessed 19 Sep 2021

# Network Physical Layout-Based Reliable Routing in Vehicular Ad Hoc Networks



S. Padmakala, A. Akilandeswari, G. Gugapriya, and Himanshu Shekhar

**Abstract** The main paramount of vehicle navigation is faster travel times and reduced collisions. So, the reliable navigation of vehicles is developed to consider both with Reliable Navigation Protocol in Vehicular Ad Hoc Networks (RNPVANETS). It calculates the asymptotic region within a road segment capacity and the number of vehicles as a function of distance. Thus, a reliability metrics calculated as a function of distance provides ingress within a road segment avoiding congestion. Further, associating the spatial–temporal coordinates of vehicle density and speeding it meticulously has been done to validate simulations of RNPVANETS using Network Simulator-2.

**Keywords** Reliability index · Load balancing rate

## 1 Introduction

Cooperative intelligent transport system process of identifying vehicle and delivering content awareness message has been studied with kinematic data. Two scenarios were used namely: autonomous or human-driven vehicles where kinematic data has been used as a standard comparison using a predefined threshold for driving. The work signifies the needs of synchronization within the time domain on theoretical bases

---

S. Padmakala (✉)

Department of CSE, St. Joseph Institute of Technology, Chennai, Tamil Nadu, India  
e-mail: [drspadmakala@gmail.com](mailto:drspadmakala@gmail.com)

A. Akilandeswari

Department of ECE, Saveetha School of Engineering, SIMATS, Chennai, Tamil Nadu, India  
e-mail: [akilandeswaria.sse@saveetha.com](mailto:akilandeswaria.sse@saveetha.com)

G. Gugapriya

School of Electronics Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India  
e-mail: [gugapriya.g@vit.ac.in](mailto:gugapriya.g@vit.ac.in)

H. Shekhar

Department of ECE, Hindustan Institute of Technology and Science, Chennai, Tamil Nadu, India

while accelerating and decelerating [6]. MAC protocol for considering communication pattern is discussed with spatial relationship. The one-dimensional network physical layout provides a statistical approach for geometrical estimation of transmission opportunity as well its outage probability [7]. In [8], the vehicle information propagation source and its location along with the vehicle density have been stated. The decreasing factor of information propagation source with its transmission range had been made stable with suitable relay conditions. In [13], non-homogenous poisson point process has been developed for one-dimensional and two-dimensional model. The signal-to-noise interference threshold adapts as per the wireless physical interface and modulation scheme, reducing computational complexity. Traffic stream model in [18] states vehicle speed is analogous to the hydrodynamics and related to velocity, density and flow.

The traffic navigation of vehicle from source to destination path along a road segment provides present and prior locations. This location coordinates of independent vehicles at appropriate time stamp resolves congestion. However, updating vehicle coordinates has to be synchronized among other vehicles via road side unit or by vehicle itself. This work proposes models for estimating vehicle density speed and collision avoidance using discrete event networking and mobility models. Simulations of vehicle movement via network physical layout and its probability of collision are also updated simultaneously. Thus, the medium access capability of the protocol is derived in terms of load balancing and access delay for vehicle movement.

Section 2 deals with the related study of VANETs. Section 3 deals with proposed system modules. Section 4 deals with simulation results of VANETs. Section 5 concludes the overall work.

## 2 Related Works

The relationship between speed and density had been quantified as speed of vehicle is inversely proportional to vehicle density [10]. The work in [11] stated when there is a transition from a free flow state into a congested state, which modeled using lower asymptotic and upper asymptotic using logistic models.

Deep neural network (DNN) has been used in VANETs for routing along congested path enforcing lower latency among vehicles. The combination of Internet of Things and BAT algorithm provides traffic updates to DNN so as to reduce the network wide mobility management [1]. In [2], stability of vehicles is attained via clustering in accordance with the speed to achieve safety. Further, TDMA provides synchronization of frames governing reliable navigation [2]. Reinforcement learning has been used in [3] to avoid local optimum which occurs in topology coordinates. Thus, it provides transfer ability to change its area, thus experiencing better updates in routing vehicles. In [4], multi-hop cooperative model has been discussed to analyse system capacity and response time via Markovian model. The model provides proper data dissemination between vehicles and between road side units for appropriate communication.

In [5], bipartite matching algorithm has been used to predict traffic identity based on either vehicle to vehicle or vehicle to infrastructure basis. The service delay has also been reduced in this scheme by using artificial neural network enabled with software-defined radio [5]. The beacon rate and channel busy ratio have been discussed where lack of situation awareness results in safety measures. Thus, attaining the congestion state of individual vehicle which consolidates to overall congestion is eliminated by Markov decision process [9]. Quality of service-based VANETs with clustering and stability period have been discussed. The cluster heads which determine the stability period are taken considering the distance and velocity metrics along with available bandwidth [12]. Mobility-aware MAC has been proposed in [14]; it states that the TDMA scheme is difficult to incorporate as the varying velocities will lead to merging collisions. So, establishing the geographical connections at topological intersections and assigning time slots in multi-lane scenarios reduces collision. In [15], the influence of clock synchronization has been studied with priority levels using MAC protocol. The process of clustering is followed at application level to assign priority for collisions between vehicles [15]. The influence of cluster stability with mobility of vehicle affects frequent change in topology studied. Consideration with time is to leave a cluster based on road length and distance covered in a time epoch and its direction. Additionally, influence of relay node which coordinates among multiple is selected to reduce the overhead in message transfer [16]. To conserve network resources, a network coding scheme has been discussed with “adaptive quantum logic”. The quantum logic provides bidirectional communication among nodes by checking non-empty packets and providing inter-relay communication. However, the average data rates are needed prior in relay to reduce bandwidth constraints [17].

## 2.1 *Problem Description*

The main attributes associated with the topology are congestion and collision associated with the road length. The impact of kinematic data and its vehicle density are also metrics which requires delay-insensitive communication for optimal route interpretation. Wireless communication routing protocol aids to provide optimal route if vehicles are synchronized to one another and the road side unit.

### 3 Proposed System

#### 3.1 Reliable Navigation Protocol in Vehicular Ad Hoc Networks (RNPVANETS)

Second-order reliability index ( $\beta$ ) formula is given by Eq. 1 where  $\mu_M$  denotes the mean and standard deviation is denoted using  $\sigma_M$ .

$$\beta = \frac{\mu_M}{\sigma_M} \quad (1)$$

Safe navigation is determined by vehicle density and is given as in Eq. 2. The variable “A” denotes the road segment capacity and “B” number of vehicles. If the value of M is above zero, the vehicle navigation is in safe state; only congestion is to be considered. If the value of M is less than zero, it is said to be in state of failure. Similarly, if “M” is equal to zero, it is in limit state.

$$M = A - B \quad (2)$$

The limit state is being interpreted before road segment gets congested. So, Eq. 3 relates to road segment capacity and Eq. 4 relates to number of vehicles.

$$A' = \frac{A - \mu_A}{\sigma_A} \quad (3)$$

$$B' = \frac{B - \mu_B}{\sigma_B} \quad (4)$$

Equations (3) and (4) can be rewritten as (5) and (6).

$$A = \sigma_A A' + \mu_A \quad (5)$$

$$B = \sigma_B B' + \mu_B \quad (6)$$

Substitute the values of (5) and (6) in Eq. 2. The limit state equation in terms of reduced variable is given in Eq. 7.

$$M = (\sigma_A A' + \mu_A) - (\sigma_B B' + \mu_B) \quad (7)$$

This scenario where capacity is greater than demand is given by Eq. (8), and it is in limit state; then it can be rewritten as in Eq. (9).

$$\sigma_B B' - \sigma_A A' = \mu_A - \mu_B \quad (8)$$

$$\frac{B'}{\left(\frac{\mu_A - \mu_B}{\sigma_B}\right)} + \frac{A'}{-\left(\frac{\mu_A - \mu_B}{\sigma_A}\right)} = 1 \quad (9)$$

Reliability ( $d$ ) is a measure given by distance of failure line to origin. The origin of the lane, in this case, is denoted using the Roadside unit and coordinate geometry formula as in Eq. 10.

$$d = \frac{\mu_A - \mu_B}{\sqrt{\sigma_A^2 + \sigma_B^2}} \quad (10)$$

### 3.2 RNPVANETS Working

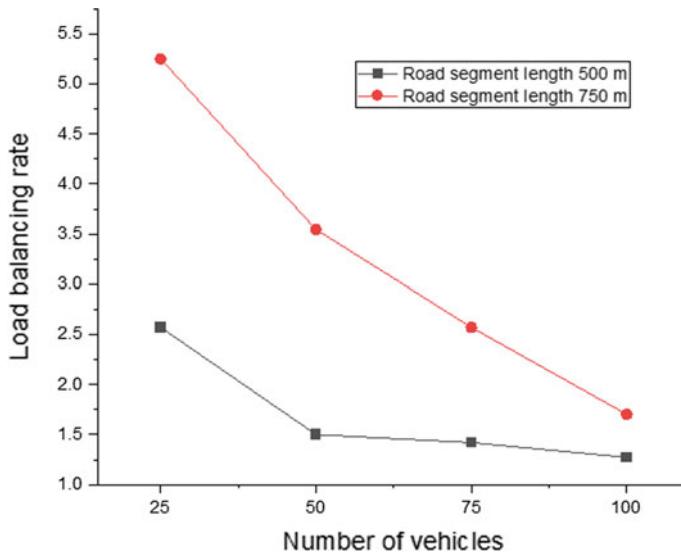
The topological configuration for vehicle navigation is assumed such that unidirectional transfer of vehicle appears from source to destination. In every time epoch, the vehicle density at ingress and egress junctions has been calculated based on the limit capacity of road segment and the number of vehicles. Thus, if the asymptotic region of a road segment is reached, then the vehicles are being rerouted to avoid congestion. Thus, more number of vehicles is simultaneously allowed considering the capacity.

## 4 Results and Discussion

Simulation of NS-2 is done with VANETMobiSim using a parser to provide syntactic roles [19]. The pivotal use of mobility using the VANETMobiSim module provides “intelligent driver module” at intersection based on function and speed of vehicle [20]. The parameters used for simulations and its values are shown in Table 1. The

**Table 1** Simulation parameters and values used for RNPVANETS

Parameters	Values
Number of vehicles	25, 50, 75, 100,
Vehicle speed	12–16 m/s
Terrain	2000 m × 2000 m
Channel	Wireless
Traffic type	CBR
Packet size	1 KB
Total simulation duration	500 s



**Fig. 1** Number of vehicles versus load balancing rate

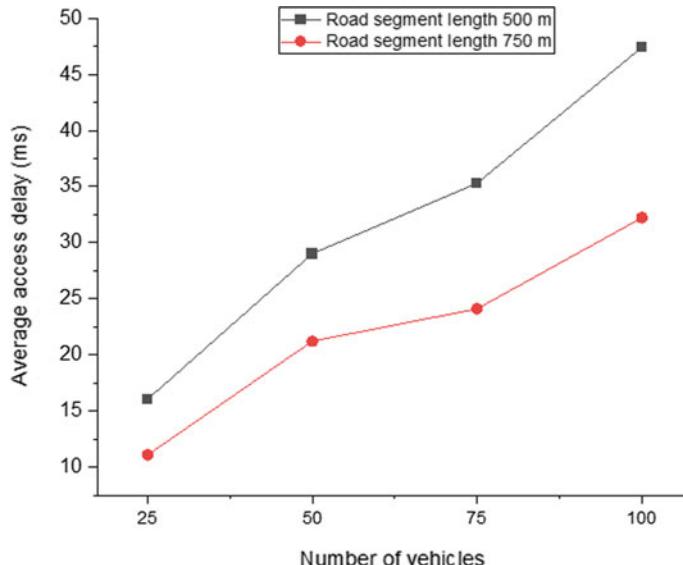
topology has one mandatory RSU positioned in the entrance of lane and one at the end of the lane and is perfectly synchronized.

Figure 1 shows the evaluation of RNPVANETS across two lanes of different lengths. The ability of the protocol to handle vehicle traffic with increase in vehicle density is shown. Load balancing rate is calculated at the RSU positioned during the ingress of road segment with the rate of vehicles allowed divided by that of denied vehicles.

Figure 2 shows the evaluation of RNPVANETS across two lanes of different lengths for delay tolerance values. The efficiency of the protocol to validate the networking information in accommodating vehicles in accessing the lane has been provided using average end to end delay. Since the vehicle density increases, the delay in processing request also increases, but the protocol balances the traffic associating RSU for ingress traffic.

## 5 Conclusion

Wireless data transfer between vehicles and road side unit infrastructure has been derived as a function of distance, associating it with reliability for movement. Load balancing function provides route reliability and determines the flow of vehicles. Evaluating the performance of protocol in scenario of two lanes with load balancing and average access delay with reliability index has been found to be superior. Future



**Fig. 2** Number of vehicles versus average access delay

work will deal with finite reaction time of drivers in terms of deceleration or no-reaction of vehicle in realistic environment.

## References

1. Kannan, S., Dhiman, G., Natarajan, Y., Sharma, A., Mohanty, S.N., Soni, M., Gheisari, M.: Ubiquitous vehicular ad-hoc network computing using deep neural network with iot-based bat agents for traffic management. *Electronics* **10**(7), 785 (2021)
2. Chiluveru, R., Gupta, N., Teles, A.S.: Distribution of safety messages using mobility-aware multi-hop clustering in vehicular Ad Hoc network. *Future Internet* **13**(7), 169 (2021)
3. Zhao, L., Bi, Z., Lin, M., Hawbani, A., Shi, J., Guan, Y.: An intelligent fuzzy-based routing scheme for software-defined vehicular networks. *Comput Netw* **187**, 107837 (2021)
4. Ravi, B., Thangaraj, J.: Stochastic traffic flow modeling for multi-hop cooperative data dissemination in VANETs. *Phys. Commun.* **46**, 101290 (2021)
5. Tang, Y., Cheng, N., Wu, W., Wang, M., Dai, Y., Shen, X.: Delay-minimization routing for heterogeneous VANETs with machine learning based mobility prediction. *IEEE Trans. Veh. Technol.* **68**(4), 3967–3979 (2019)
6. Lyamin, N., Vinel, A., Jonsson, M., Bellalta, B.: Cooperative awareness in VANETs: on ETSI EN 302 637-2 performance. *IEEE Trans. Veh. Technol.* **67**(1), 17–28 (2017)
7. Liu, W., He, X., Huang, Z., Ji, Y.: Transmission capacity characterization in VANETs with enhanced distributed channel access. *Electronics* **8**(3), 340 (2019)
8. Huang, R., Wu, J., Long, C., Zhu, Y., Li, B., Lin, Y.B.: SPRCA: distributed multisource information propagation in multichannel VANETs. *IEEE Trans. Veh. Technol.* **66**(12), 11306–11316 (2017)

9. Aznar-, J., Garcia-Sanchez, A.J., Egea, E., Garcia-, J.: MDPRP: a Q-learning approach for the joint control of beaconing rate and transmission power in VANETs. *IEEE Access* **9**, 10166–10178 (2021)
10. Greenshields, B. D., Bibbins, J. R., Channing, W. S., Miller, H. H.: A study of traffic capacity. In: Highway research board proceedings. vol. 1935, National Research Council (USA), Highway Research Board (1935)
11. Wang, H., Li, J., Chen, Q.Y., Ni, D.: Logistic modeling of the equilibrium speed-density relationship. *Transp. Res. Part A: Policy Practice* **45**(6), 554–566 (2011)
12. Fatemidokht, H., Rafsanjani, M.K.: QMM-VANET: an efficient clustering algorithm based on QoS and monitoring of malicious vehicles in vehicular ad hoc networks. *J. Syst. Softw.* **165**, 110561 (2020)
13. Zhao, J., Wang, Y., Lu, H., Li, Z., Ma, X.: Interference-based QoS and capacity analysis of VANETs for safety applications. *IEEE Trans. Veh. Technol.* **70**(3), 2448–2464 (2021)
14. Lyu, F., Zhu, H., Zhou, H., Qian, L., Xu, W., Li, M., Shen, X.: MoMAC: mobility-aware and collision-avoidance MAC for safety applications in VANETs. *IEEE Trans. Veh. Technol.* **67**(11), 10590–10602 (2018)
15. Abbas, G., Abbas, Z.H., Haider, S., Baker, T., Boudjit, S., Muhammad, F.: PDMAC: a priority-based enhanced TDMA protocol for warning message dissemination in VANETs. *Sensors* **20**(1), 45 (2020)
16. Ullah, S., Abbas, G., Waqas, M., Abbas, Z.H., Tu, S., Hameed, I.A.: EEMDS: an effective emergency message dissemination scheme for urban VANETs. *Sensors* **21**(5), 1588 (2021)
17. Hammond, O.A., Kahar, M.N.M., Hammond, W.A., Hasan, R.A., Mohammed, M.A., Yoob, A.A., Sutikno, T.: An effective transmit packet coding with trust-based relay nodes in VANETs. *Bull. Electric. Eng. Inf.* **9**(2), 685–697 (2020)
18. Fiore, M.: Mobility models in inter-vehicle communications literature. *Politecnico di Torino* **147** (2006)
19. Härrí, J., Filali, F., Bonnet, C., Fiore, M.: VanetMobiSim: generating realistic mobility patterns for VANETs. In: Proceedings of the 3rd international workshop on Vehicular ad hoc networks., pp. 96–97 (2006)
20. Härrí, J., Fiore, M., Filali, F., Bonnet, C.: Vehicular mobility simulation with VanetMobiSim. *SIMULATION* **87**(4), 275–300 (2011)

# A Hybrid Split and Merge (HSM) Technique for Rapid Video Compression in Cloud Environment



R. Hannah Lalitha, D. Weslin, D. Abisha, and V. R. Prakash

**Abstract** Media files require huge resources and long hours for telecasting, as they involve large amount of data which need to be compressed. The rise of cloud computing, data intensive applications become attractive for a good public that does not require the same resources to due to large scale distribution. In this paper, an effort has been made to design and implement the performance of digital video to MPEG4 transcoding within the cloud environment. A hybrid split and merge (HSM) technique for compressing video in cloud is proposed. A long duration sport video is taken to validate the technique. The video compression involves hybridizing the merits of object-based and block-based methods. The moving objects are detected and coded using object-based approach, whilst the moving file segmentation and compression are implemented with the help of block-based approach. The audio stream and the video chunks are scattered between the nodes to be compressed all together for fast processing. The suggested technique uses the cloud resources optimally for attaining low cost, time, and improved peak signal to noise ratio (PSNR).

**Keywords** Data compression · Cloud environment · Video frames · Motion estimation · Encoding time · Peak signal to noise ratio

---

R. H. Lalitha (✉)

Department of Electrical and Electronics Engineering, B.S Abdur Rahman Crescent Institute of Science and Technology, Chennai, Tamil Nadu, India

D. Weslin

Department of Information and Technology, Mohammed Sathak A.J College of Engineering, Chennai, Tamil Nadu, India

D. Abisha

Department of Computer Science and Engineering, National Engineering College, Kovilpatti, Tuticorin, Tamil Nadu, India

e-mail: [abisha\\_cse@nec.edu.in](mailto:abisha_cse@nec.edu.in)

V. R. Prakash

Department of ECE, Hindustan Institute of Technology and Science, Chennai, Tamil Nadu, India

## 1 Introduction

Data compression is a technique that decreases the data size and the data gets stored in fewer bits which results in less data space storage, usage of resources or transmission capacity. The video compression states the problem in bandwidth minimization and storage of video content. Encompassing cloud has been used in various sectors such E-health wherein users are provided access to master, slave, and cloud works [1]. Visibility application in images the feature learning and distributed learning discussed cloud as a suitable means of prediction in multiple views [2]. Temporal filter where the earlier approaches which provides consistency across the approaches of noisy images by vector matching [3]. Objective quality metrics such as peak signal to noise ratio (PSNR) is one of the main considerations which have been considered for video processing. Finding the object in region of interest and merging the attributes is one of the approaches which have been discussed in this work using cloud and video processing techniques. Section 2 deals with brief review of digital video compression and cloud environment. Section 3 deals with hybrid split and merge techniques for video access. Section 4 describes MATLAB results of HSM. Section 5 concludes overall work.

## 2 Existing System

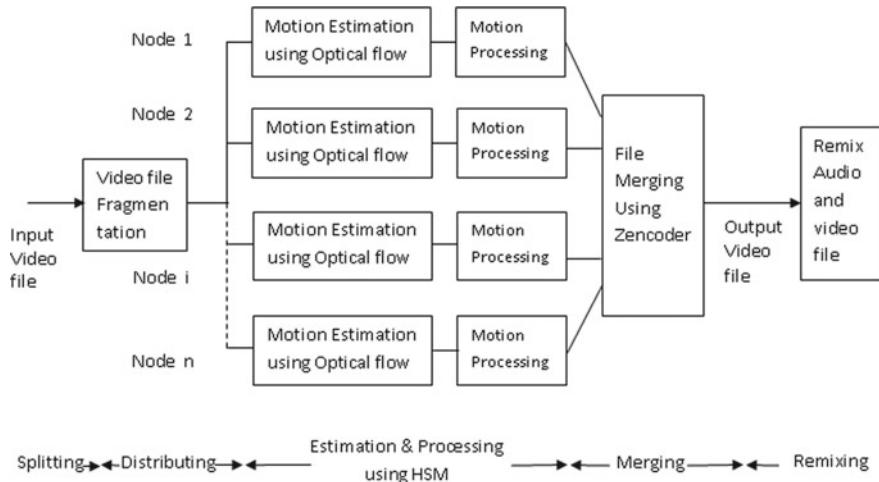
Detector working on decoded frames has been one of the approaches to enable speed and accuracy in video processing [4]. A network transparent approach has been discussed for working in pixel level and compression domains. The platform provides object detection without the addition of excess modules [5]. Motion vector estimation has been done in MVmed with Python package which does not used a fixed interval between fames [6].

## 3 Proposed System

The block diagram of proposed system from video file segmentation to Zencoder is shown via Fig. 1.

### 3.1 Split Step

The splitting of video file and the distribution of encoding process hikes the efficiency of encoding in the cloud environment. The video file is extracted from audio file initially. Then, the video is split into small files and can be processed on several



**Fig. 1** The proposed HSM technique for video compression in cloud environment

machines parallel as explained in Fig. 3. This decreases the total encoding time. At first, the video is splitted from audio, since the video encoding process is more complicated than audio. The audio is not necessary to be fragmented as the fragmentation does not affects the performance. If audio and video are compressed together, the synchronization is affected. To elude the problems of delays, the audio and video are separated. The video file is fragmented to chunks with frames which is unchanging with respect to runtime. The size of chunk which was produced with less than 250 frames, when we have video of 29.97 fps.

### 3.2 Process Step

After, fragmentation of video is the compression process. The initial frame is divided into  $8 \times 8$  blocks, DCT and quantization is applied to each block. The Huffman coding is applied to encode the initial frame [7]. The successive frames are coded with the optical flow technique. The foreground object is segmented from the video file using the binary generated by technique called optical flow. This technique is more efficient than the other moving object estimation methods. In this method, very small object can be easily detected. In addition, the foreground aperture and ghosting problems are avoided, and hence, the compression ratio can be improved [8–11].

### 3.3 Estimation of Optical Flow

The optical flow within the frames is estimated, the noises are eliminated, and the object shapes are improved. The apparent motion of object in the frames is represented by optical flow method. It can be projected on the image sequences having small displacements and smooth motion fields.

$$P(a, b, t) = P(a + \delta a, b + \delta b, t + \delta t) \quad (1)$$

$P$  is the original image frame.

$P(a, b, t)$  is the intensity value of pixel located at the position  $(a, b)$  in the  $t$ th frame, and the intensity values are moved to the new position  $(a + \delta a, b + \delta b, t + \delta t)$  in  $(t + \delta t)$ th frame and become  $P(a + \delta a, b + \delta b, t + \delta t)$ .

$(\delta a, \delta b,)$  and  $\delta t$  are the spatial and temporal coordinate difference between the frames.

$$P(a, b, t) = P(a, b, t) + \frac{\partial p}{\partial a} \delta a + \frac{\partial p}{\partial t} \delta t \quad (2)$$

$$P_a \delta a + P_b \delta b + P_t \delta t = 0 \quad (3)$$

$$P_a u + P_b v + P_t = 0 \quad (4)$$

$P_a$ ,  $P_b$  and  $P_t$  are derivatives of  $P$  with respect to  $a$ ,  $b$ , and  $t$ , respectively.

$u = \frac{\delta a}{\delta t}$ , the velocity components in the horizontal directions.

$v = \frac{\delta b}{\delta t}$ , the velocity components in the vertical directions.

To find the velocity components  $(u, v)$ ,

$$u = u_{av} - P_a \frac{N}{D} \quad (5)$$

$$v = v_{av} - P_b \frac{N}{D} \quad (6)$$

$$N = P_a u_{av} + P_b v_{av} + P_t \quad (7)$$

$$D = \beta^2 + P_a^2 + P_b^2 \quad (8)$$

where  $u_{av}$  and  $v_{av}$  denote mean of neighbouring pixels of  $u$  and  $v$ , respectively.

$B$  is used to regularize the noise ratio signal.

The components of  $u$  are used to find the motion flow of video. Optical flow is a well organized and effective technique to segment the motion in the dataset such as video. This method finds the motion with the help of optical flow velocity components ( $u, v$ ).

### 3.4 Normalization

The magnitude of optical flow value of the pixel at position  $(a, b)$  is estimated as

$$Pl(a, b) = \sqrt{u(a, b)^2 + v(a, b)^2} \quad (9)$$

The magnitude is in the range of (0–255) to represent 2D grey scale frame. The following normalization is applied to stretch the values of  $Pl$  linearly in the range of [0–255]:

$$I(a, b) = \text{int}\left[\frac{Pl(a, b)}{Pl_{\max} - Pl_{\min}} X 25\right] \quad (10)$$

$Pl_{\max}$  = Maximum value of  $Pl$

$Pl_{\min}$  = Minimum value of  $Pl$

### 3.5 Otsu's Approach

The best threshold value is obtained using

$$\text{threshold} = \text{Arg}(\max(\delta^2)^2) \quad (11)$$

where  $\delta$  is the inter class variance between two classes of frames with different intensity ranges.

### 3.6 Binarization

To separate the foreground object, grey scale is converted to binary using Otsu's adaptive threshold.

$$P_g(a, b) = \begin{cases} 0 & \text{if } I(a, b) \leq \text{threshold} \\ 1 & \text{otherwise} \end{cases} \quad (12)$$

The threshold value is calculated from the interclass variance between the two frames [12].

### 3.7 Motion Estimation

To enhance the effectiveness of video coding, a method called block matching-based motion estimation is deployed. The content of block is computed with respect to contents of small different previous macro block. In block matching method, the frames are classified into overlapping and non-overlapping blocks of pixels [13–15]. Each block is estimated by matching with the reference frame in the search space. To find the best matching, mean absolute difference (MAD) is computed by

$$\text{MAD} = \frac{1}{w_m h_m} \sum_{i=0}^{w-1} \sum_{j=0}^{h-1} |c_{i,j}^{a,b} - R_{i,j}^{a+da,b+db}| \quad (13)$$

where

$w_m$ , the width and height of macro block.

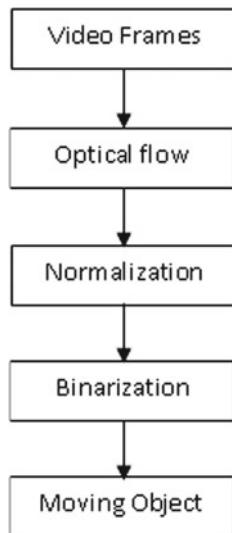
$h_m$ , the height of macro block.

$c_{i,j}^{a,b}$ ,  $R_{i,j}^{a+da,b+db}$  denote pixel value in macro block of the current and reference frame, respectively, at positions  $(a, b)$  and  $(a + da, b + db)$ .

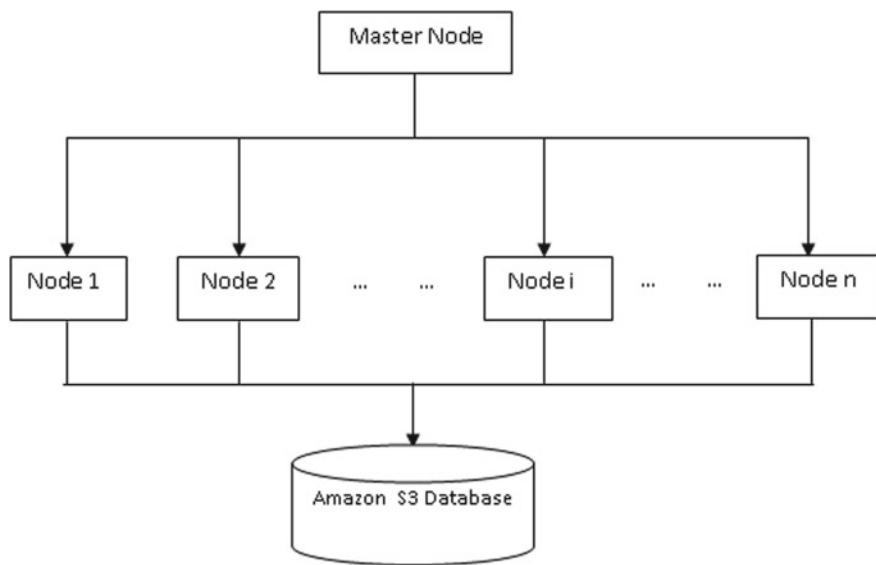
$(da, db)$  represents motion vector. Figure 2 explains the procedure for moving object segmentation.

#### 3.7.1 The Merge Step

Initially, the fragments of the video file are joined. There are several open-source tools for working with high definition audio and video operation in cloud. The merging of video files is done by Zencoder, which can give the output video that can be viewed through Internet or mobile device. The output is generated by remixing the audio stream with the video.



**Fig. 2** The flow diagram explaining moving object segmentation



**Fig. 3** Cloud environment for HSM technique

### **3.8 Cloud Environment for HSM Technique**

A public cloud is employed in which the resources can be utilized on request and payment. The video chunks are distributed between the nodes for processing. These nodes can be added or removed based on the request, by a master node to reduce the operation cost. Even the master node can be added or removed to reduce the cost. The public cloud environment for HSM technique is depicted in Fig. 3. The Amazon Web services (AWSs) are the platform for testing, and the proposed technique can also be implemented on other cloud services. The database in AWS is Amazon S3. According to the request and payment, the bandwidth links are limited.

## **4 Results and Discussion**

The proposed compression technique's performance is measured with the help of the long video dataset which was taken from UGC dataset at 30 frames/s rate. Three videos such as sport video, monkey video, and dog video were used for implementation. The size of sport video is  $530 \times 298$  pixels, monkey video and dog video are  $297 \times 246$  pixels. All the experiments are executed on Intel (R) core (TM) i7-4770cpu@3.40 GHz processor with 4 GB RAM and on MATLAB R2013 environment. AWS is utilized, and the compression was performed with H.24 standard. 80 nodes are implemented, and the fragment size is 800 frames.

Figure 4 shows the comparison between the output video file of the proposed HSM video compression technique in cloud environment and the conventional compression techniques for the sport, monkey, and dog video files (Fig. 5).

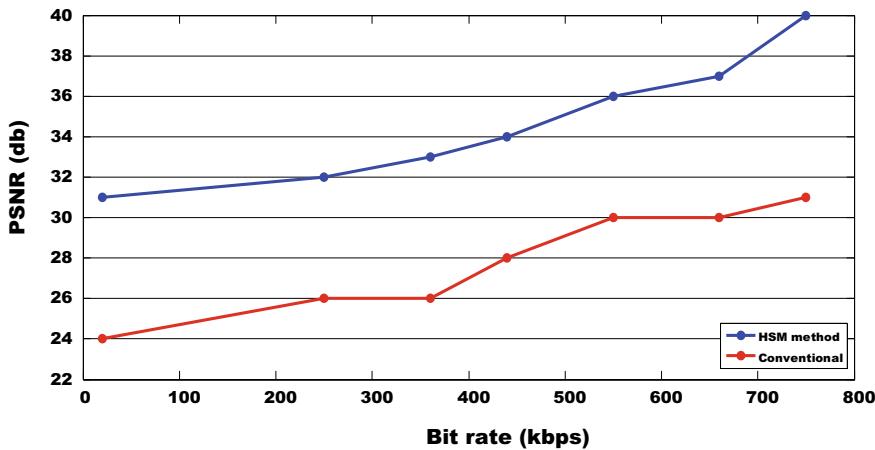
The proposed HSM technique was tested with videos of different duration, and the total time for encoding was computed in Table 1. From Fig. 6, it is seen that the time for encoding using the conventional method increases linearly with increase in duration of video input. The encoding time is almost constant for short duration videos in case of HSM compression method. For the videos of small duration, there is reduction in 12% in time taken by conventional methods. For the longer videos like match relays, the encoding time is reduced eminently. The increase in the number of nodes in the cloud reduces the cloud encoding time further. From the Table 1, it is evident that for a 120 min video, the encoding time is limited to 3 min which will take 480 min by the conventional technique. The total time for encoding is limited to 3 min, and several video files for advertisement, breaking news, TV programmes, movies, and matches are considered. The encoding time for video which has duration of 2 h is reduced from 8 h to 3 min public cloud environment. The available resources in cloud are enormous so that the video of very long duration of 1000 h can be done in few minutes using HSM technique in the cloud, which would take 100 h using traditional methods.



**Fig. 4** Comparison between the output video file of the proposed HSM and the conventional compression techniques for the sport, monkey, and dog video files

## 5 Conclusion

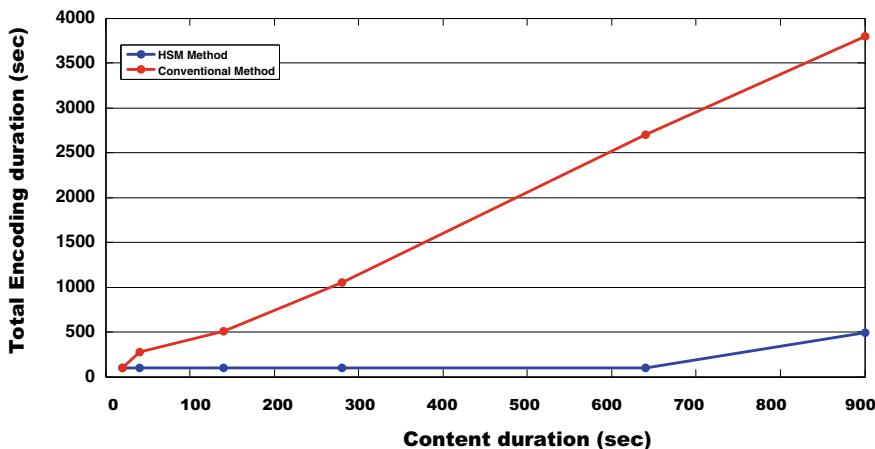
The huge measures needed for information processed video handing-off. This information ought to be compacted for productive transmission. The proposed HSM method crossbreeds the benefits of both substances-based and block-based procedures for video pressure. To precisely recognize and code the forefront objects by object-based coding approach, we have utilized the effective optical stream-based movement division and Freeman chain code strategies, individually. The pressure is done in a cloud climate which permits utilization of public foundation. The assets are utilized on request, to improve adaptability and works with simple preparing of enormous datasets. The superior quality long game video was tried for effective pressure as far as PSNR and encoding times. The proposed HSM guarantees expanded PSNR



**Fig. 5** PSNR-based performance evaluation for sport video

**Table 1** Comparison of the proposed HSM and conventional techniques in terms of encoding time

Video duration (minutes)	Encoding time (minutes)	
	Conventional technique	HSM technique
1	4	3
6	20	3
30	120	3
120	480	3



**Fig. 6** Graph showing variation of encoding duration with content duration for the proposed HSM and conventional methods

when contrasted with ordinary strategies. Further, the encoding times are especially decreased for the long recordings like film and match transfers.

## References

1. Memos, V.A., Psannis, K.E., Goudos, S.K., Kyriazakos, S.: An enhanced and secure cloud infrastructure for e-health data transmission. *Wireless Pers. Commun.* **117**(1), 109–127 (2021)
2. Song, M., Han, X., Liu, X.F., Li, Q.: Visibility estimation via deep label distribution learning in cloud environment. *J. Cloud Comput.* **10**(1), 1–14 (2021)
3. Moura, R. C., Hemerly, E. M., da Cunha, A. M.: Temporal motion vector filter for fast object detection on compressed video. *J. Commun. Inf. Syst.*, **29**(1) (2014)
4. Alvar, S. R., Bajić, I. V.: MV-YOLO: Motion vector-aided tracking by semantic object detection. In: 2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP), pp. 1–5, IEEE (2018)
5. Nohara, M., Nishi, H.: Video object detection method using single-frame detection and motion vector tracking. In: 2020 IEEE 18th International Conference on Industrial Informatics (INDIN). Vol. 1, pp. 119–125, IEEE (2020)
6. Bommes, L., Lin, X., Zhou, J.: MVmed: fast multi-object tracking in the compressed domain. In: 2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA), pp. 1419–1424. IEEE (2020)
7. Liu, Y.K., Z̄ alik, B.: “An efficient chain code with huffmancode”. *Patt. Recogn.*, **38**(4), 553–557 (2005)
8. Belloulata, K., Belalia, A., Zhu, S.: Object-based stereo videocompression using fractals and shape-adaptive DCT. *AEU-Int. J. Electron Commun.* **68**(7), 687–697 (2014)
9. Li, Y., Tao, X., Lu, J.: “Hybrid model-and-object-based realtimeconversational video coding”. *Signal Process. Image Commun.*, **35**, 9–19 (2015)
10. Talluri, R., Oehler, K., Barmon, T., Courtney, J.D., Das, A., Liao, J.: A robust, scalable, object-based video compression technique for very low bit-rate coding. *IEEE Trans. Circuits Syst. Video Technol.* **7**(1), 221–233 (1997)
11. Zhu, Z., Wang, Y., Jiang, G.: On multi-view video segmentationfor object-based coding. *Digital Signal Process.* **22**(6), 954–960 (2012)
12. Sengar, S. S., Mukhopadhyay, S.: “Motion segmentation-based surveillance video compression using adaptive particle swarm optimization.” *Neural Comput. Appl.*, Springer (2019)
13. Cuevas, E., Zaldivar, D., Cisneros, M., Sossa, H., Osuna, V.: Block matching algorithm for motion estimation based on ArtificialBee Colony (ABC). *Appl. Soft Comput.* **13**(6), 3047–3059 (2013)
14. Guo, X., Jiang, G., Cui, Z., Tao, P.: Homography-based blockmotion estimation for video coding of PTZ cameras. *J. Visual Commun. Image Represent.* **39**, 164–171 (2016)
15. Gallant, M., Cote, G., Kossentini, F.: An efficient computation-constrained block-based motion estimation algorithm for lowbit rate video coding. *IEEE Trans. Image Process.* **8**(12), 1816–1823 (1999)

# A New Intrusion Detection and Prevention System Using a Hybrid Deep Neural Network in Cloud Environment



**Subalakshmi Mani, Bose Sundan, Anitha Thangasamy, and Logeswari Govindaraj**

**Abstract** Cloud computing has become an innovative technology, with distributed on-demand services; it has an attractive target for potential cyber-attacks by intruders. Intrusion detection systems (IDS) and intrusion prevention systems (IPS) are the most commonly used mechanisms to detect and prevent large-scale network traffic and any type of online attacks. In this paper, a novel framework for a deep long short-term memory (LSTM)-based intrusion detection system is proposed to detect network traffic flow patterns as either malicious or normal in a cloud environment. The proposed IPS prevents malicious attacks received from IDS by increasing the detection rate of malicious attacks and reducing computational time. The experimental results for the overall performance of the intrusion detection and prevention system were evaluated with 99% accuracy, precision, recall, and F-score. The evaluation results prove that the proposed system is suitable for effective attack detection and prevention in resource-constrained operational in cloud computing environments.

**Keywords** Intrusion detection system (IDS) · Intrusion prevention system (IPS) attacks · Security · Deep long short-term memory model · Cloud computing

## 1 Introduction

An intrusion detection system (IDS) is the most commonly used method to detect any specified type of attack. Many network-based attacks may particularly attack cloud security at the level of the network layer. It includes IP spoofing, port scanning, a man in the middle of the attack spoofing, denial-of-service attacks (DOS), and the distributed denial-of-service (DDOS) attacks. Generally, an intrusion detection system (IDS) is a software package that is responsible for detecting threats across

---

S. Mani · B. Sundan · A. Thangasamy (✉) · L. Govindaraj

Department of Computer Science and Engineering, College of Engineering, Guindy, Anna University, Chennai, India  
e-mail: [ani.astt18@gmail.com](mailto:ani.astt18@gmail.com)

B. Sundan  
e-mail: [sbs@annauniv.edu](mailto:sbs@annauniv.edu)

the network or system, while an intrusion prevention system (IPS) is the software responsible for stopping all the events. Nowadays, systems run both as intrusion detection and prevention systems. Among that, it can define some of the open-source software such as Snort, Suricata, and Bro.

Due to the increase in a newer form of attacks and malware type of detection was introduced. It was primarily introduced to detect newer attacks that were signature-based. The only problem with this type of detection approach might be suffering from false positives. Now from the above concepts, it clearly understands how an intrusion detection and prevention system are work. It gives precise information ranging from installation of machines such as victim, attacker, and snort to deploying it and performing attacks to see the working of snort to prevent the attacks that are detected. Despite the fact that the number of cloud projects has expanded rapidly in recent years, assuring the availability and security of project data, services, and resources remains a critical and demanding research subject. Attacks can deplete the cloud's resources, take the majority of its bandwidth, and devastate an entire cloud project in a short amount of time. Cloud computing is the preferred choice of every organization since it provides flexible and pay-per-use-based services to its users. However, security and privacy are a major hurdle in its success because of its open and distributed architecture that is vulnerable to intruders. The most popular approach for detecting attacks is an intrusion detection system (IDS). It examines the type, positioning, detection time, detection technique, information sources, and threats that existing cloud-based intrusion detection systems can detect.

A snort is an open-source tool used for intrusion detection and prevention systems. This paper discusses all the elements of snort [1, 2] that must prevent malicious attacks using snort to configure it as a full-fledged IPS. It also gives an overview of the functioning of snort by performing several attacks such as cross-site scripting and SQL injection attack. On the whole, it presents a complete step-by-step process of deploying a complete package of snort along with the configuration of both the attacker and victim.

The following three types of modes in snort [3] are mentioned below:

1. Sniffer mode: Sniffer mode examines each packet from the network that collects and shows it on the snort console.
2. Packet logger mode: To write all of the logs to disk, packet logger mode is utilized. It will capture all the incoming packets and store them in the database.
3. Intrusion detection mode: This mode will monitor and analyze all network traffic according to user-defined rules, and it prevents attacks whenever it come.

The main contribution of the work is as follows: A novel deep LSTM detection system is proposed with the preprocessing techniques to make the dataset more concise. Based on the analysis, the feature is extracted using the combined algorithms of PCA and LDA for dimensionality reduction and to avoid over-fitting.

1. The proposed deep LSTM-based IDS uses a grid search algorithm to choose the best hyperparameter weight, along with seep LSTM to remain a memory for a long time to detect network traffic flows as either malicious or normal in

- a cloud environment. The detected type of malicious attack is prevented using a proposed prevention system for increasing the detection rate and reducing computational time.
2. The overall performance of the integrated intrusion detection and prevention system is compared with the existing work, and it provides 99% accuracy, precision, recall, F-score, and malicious packets dropped at a high rate.

The other section of the paper is organized as follows. Section 2 describes the detection of any kind of attack with various deep learning techniques. Section 3 discusses the proposed deep LSTM intrusion detection and prevention system for the attacks. Section 4 is about the evaluation metrics, and the experimental results are discussed in Sect. 5.

## 2 Related Works

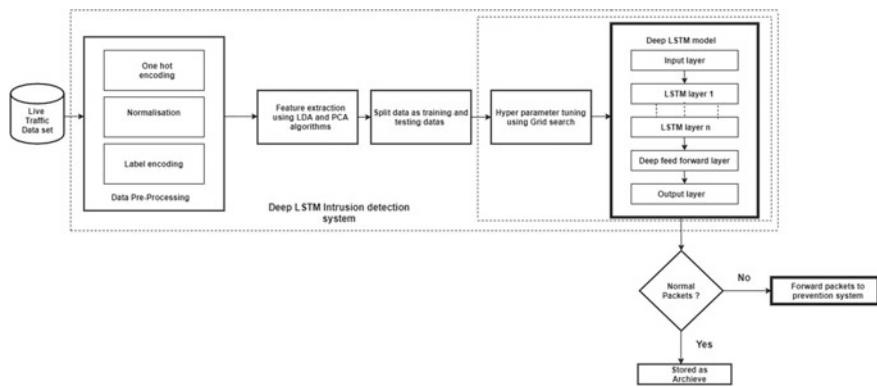
Millar et al. [4], in today's Internet, distributed denial-of-service (DDOS) attacks are one of the harmful threats that disrupt essential services. The most compromising challenge in DDOS detection is the volume of live traffic that is to be analyzed is coupled with the attack approaches. LUCID [4] presents a practical lightweight deep learning [3–6] detection system that uses CNN to sequentially classify live traffic flow as malicious or benign. Zengguang et al. [7] largely focus on edge computing and machine learning, and it gives the novelty of the weight update method, which is similar to the principle of distributed machine learning [8], and here federated learning has some more similarities over local data. Secondly, FL uses the process of model training. It deals with the effective measure of data privacy protection. The traditional protection methods are k-anonymity and diversification to specifically reduce noise in the data. In this experiment, CNN-based intrusion detection models were analyzed [9]. The self-organizing map algorithms that successfully updated community rules and learning rates proposed by Aneetha et al. [10] to govern the original data size as the original weight vector, as well as the static structure and random assignment of simple SOM weight vectors. The new technology is evaluated using performance metrics such as detection rate and false alarm rate. Aqeel et al. [11] the most common DDOS attacks with multiple comprised systems in clouds that act like a victim. From the traffic pattern, real-time attack alerts are obtained. It also demonstrates the alternative path to the origin source. So, it can block the path source of the attack. Gupta, Sharma et al. [2], an intrusion prevention system (IPS) is an IDS with the capability of blocking intrusions. It will drop the malicious packets, blocking the specific IP address or it will reset the connection with the system. An IPS acts faster on the malicious one whereas an IDS only gets a copy of the network traffic. Patel A et al. [12] propose the new IDPS system with alarm management and this prevention technique to investigate attacks in the cloud environment [13].

### 3 Proposed Methodology

This paper presents deep LSTM intrusion detection and prevention system that can be deployed in cloud environments. Our proposed system identifies and prohibits harmful activity TCP and UDP patterns of malicious attacks in a cloud environment. To identify cloud attacks, a novel deep LSTM intrusion detection and prevention system is proposed for network traffic in cloud environments. After receiving the packets from the proposed deep LSTM detection system the condition uses protocol and length attributes or features is used to check whether the received packets are normal or not. This section describes the deep LSTM IDS framework Fig. 1. The network traffic preprocessing method, feature extraction techniques, proposed intrusion detection system and prevention system.

#### 3.1 Dataset Description

The dataset consists of online traffic flows which are collected as a packet by a tool called Wireshark. Live traffic is collected using a virtual machine in which Ubuntu is a target address (carried by LOIC) and Windows 8.1 as a source address. Table 1 shows the classification of the train and test dataset. The dataset consists of the source port, destination port, protocol, time, length, etc., of 32 features and 1 feature



**Fig. 1** Proposed deep LSTM intrusion detection and prevention system

**Table 1** Train and test set classification of dataset

Dataset data type	Train set	Test set
Malicious	102,422	25,605
Normal	78,175	19,543
Total	180,597	45,148

as a label that shows the attack is either malicious or normal. The total parameters are above two lakh parameters. The features of the protocol contain 10 categories, whereas the label consists of two categories, i.e., malicious and normal, in which malicious counts up to 128,027, and normal counts to 97,718.

### **3.2 Data Preprocessing Techniques**

In the data preprocessing step, the unbalanced data are converted to balanced data using normalization, label encoding, and one-hot encoding. Algorithm 1 provides all the steps that govern the data preprocessing for deep LSTM IDS.

#### **3.2.1 Normalization**

Feature scaling is an essential preprocessing step for our proposed framework. Normalization scales each feature separately to a fixed range. Usually, the range [0,1] is used. To scale these features to this range, the minimum and maximum of each feature in the dataset must be calculated.

#### **3.2.2 Label Encoding**

As the dataset, features of protocol and label are in an unbalanced state to convert it into machine-readable form. It mainly involves converting each value in a column to a number.

#### **3.2.3 One-Hot Encoding**

One-hot encoding is applied to categorical data to convert it into a binary vector representation for use in many deep learning algorithms.

---

Algorithm 1: Data preprocessing

---

Input:	Unbalanced data
Output:	Balanced data
Step 1:	To convert the unbalanced data into a balanced one, it involves the following techniques: normalization, label encoding, and one-hot encoding
Step 2:	In normalization, it scales the feature values to the same specified range without distorting differences in the range of values. In normalization, selected standard scalar The value is normalized as follows, $Y = (x - \text{mean}) / (\text{standard\_deviation})$ Where $x$ is the original feature vector

(continued)

(continued)

---

Algorithm 1: Data preprocessing

---

Step 3: Label encoding can be used to transform non-numeric to numeric labels

---

Step 4: One-hot encoding changes the variable of binary representation as 0's and 1

---

### 3.3 Feature Extraction Methods

In dimensionality reduction algorithms, PCA and LDA are used to select the best features. As the dataset has too many features with too many attributes and to make our model more lightweight, the dataset is reduced with the combined algorithm of PCA and LDA. Combining attributes into a new reduced set of features. The feature extraction method involves.

1. Principal component analysis algorithm
2. Linear discriminant algorithm.

#### 3.3.1 Linear Discriminant Algorithm

LDA is a supervised classification technique that provides more classes to the feature set while also reducing its dimensionality. And the dimensionality reduction property also makes our model more accurate. It aims to make the distance between data points of the same class more compact, which is shown in Algorithm 2.

---

Algorithm 2: Linear discriminant analysis algorithm

---

Input: Preprocessed data

---

Output: Extracted Features

---

Step 1: Compute the mean vectors for the input features dataset

$$\text{Mean : } \bar{X} = \frac{1}{n} \sum X_i$$

Where n is the number of values and  $X_i$  is the sum of the values of each input  $X$

Step 2: Calculate the scatter matrices, within the class ( $S_w$ ) and between classes ( $S_b$ )Step 3: Find the linear discriminants by computing the eigenvalues  $S_w - 1, S_b$ Step 4: Select the eigenvectors  $W$  with the highest eigenvalues as the linear discriminants for the new feature set

Step 5: The new feature set obtained from the linear discriminants is then used to obtain the transformed input dataset

$$Y = X.W$$

where  $X$  is an  $n \times d$ -dimensional matrix representing the  $n$  samples, and  $y$  is the transformed  $n \times k$ -dimensional samples in the new subspace

---

### 3.3.2 Principal Component Analysis Algorithm

PCA uses a conversion method to convert the data into a lower-dimensional space while maximizing the data to analyze the covariance of each X and Y feature using the following Algorithm 3. It is a method for summarizing data. The resulting output features are the uncorrelated orthogonal basis set called principal components. The largest eigenvalues contain the most massive amounts of data.

---

Algorithm 3: Principal component analysis

---

Input:	Preprocessed data
Output:	Extracted Features
Step 1:	Compute mean vectors for the input features dataset ( $x_i$ ) Mean : $\bar{x} = \frac{1}{n} \sum x_i$ Where n = number of values and $x_i$ is the sum of values of each input x
Step 2:	Determine the scatter matrix – covariance matrix, two feature vectors $x_j$ and $x_k$ the covariance between them $\sigma_{jk}$ can be calculated using the following equation: $\sigma_{jk} = \sum_{i=1}^n (x_j - \mu_j)(x_k - \mu_k)$
Step 3:	Compute eigenvectors and eigenvalues to compute the principal components
Step 4:	Eigenvectors to be sorted in descending order W
Step 5:	Integrate the principal components onto the input features

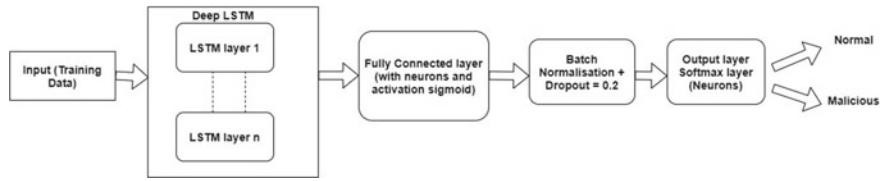
---

### 3.4 Splitting of Data

The random selection of 80% of the original data is split, and the testing set should be the remaining 20%. The parameters of the model of an optimization algorithm are adjusted during a training process. The test set must be separated from the training set. It is used to evaluate the accuracy of the trained model.

### 3.5 Hyperparameter Tuning Using Grid Search

The objective is to minimize the complexity of the time taken and the performance time of the deep LSTM model. To achieve this, a lightweight model is proposed. It has shared and reused parameters about the weight. This reduces the storage and memory requirements of our model. Hyperparameters are the setting parameters of the model that are not changed during the training process. The training parameters consist of batch size, number of epochs, learning rate, momentum, and dropout. To develop good LSTM models, the appropriate hyperparameters should be found. It is possible to determine the combinations of suitable hyperparameters. The grid search technique is used to find the best hyperparameters. Grid search will result in the most



**Fig. 2** Deep LSTM model for attack detection

‘accurate’ predictions. In automated research techniques, algorithms are used to find the best values. Steps involved in hyperparameter tuning using grid search algorithm are shown as Algorithm 4.

---

Algorithm 4: Hyperparameter tuning using grid search

---

Input:	Extracted features from dimensionality reduction algorithms
Output:	Fit the model
Step 1:	Briefly give the parameters that you use to train
Step 2:	Grid search along with the values that you wish to try (i.e., epochs = [10, 50, 100] Batch_size = [10, 50, 20] Dropout = [0.2, 0.3, 0.7] Learning rate = [0.001, 0.0002, 0.00001])
Step 3:	Fit the grid search

---

### 3.6 Deep LSTM IDS Model for Attack Detection

Deep LSTM model Fig. 2 is an advanced updation to the LSTM model that has one or more hidden layers. Each layer contains multiple memory cells that will hold information for a long duration. More hidden layers can be added to the multilayer perceptron neural network to make it look deeper. It shows the hyperparameter tuning with deep LSTM model detection is shown as algorithm 5.

---

Algorithm 5: Deep LSTM for attack detection

---

Input: Extracted features as input

---

Output: Classify the type of attack

---

Step 1: 80% of the training data is taken as input to the deep LSTM model

---

Step 2: As it is deep LSTM, this model contains more hidden layers and each layer contains memory cells

---

Step 3: The next layer is the fully connected layer with the activation function as sigmoid

---

Step 4: Batch normalization is done along with the dropout rate

---

(continued)

(continued)

---

**Algorithm 5: Deep LSTM for attack detection**

---

Step 5: The final layer is the output layer, which is the softmax layer that classifies malicious or benign

---

Step 6: Train the chosen deep LSTM model on the training set (80%)

---

Step 7: Validate the chosen deep LSTM model on the evaluation set (20%)

---

Step 8: Test the chosen deep LSTM model on the test set

---

Step 9: Repeat steps until the desired results are reached

---

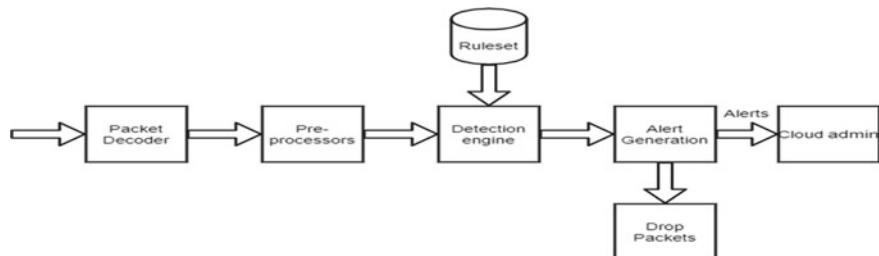
### 3.7 Proposed Intrusion Prevention System

The detection of malicious attacks employing deep learning algorithms is done in a deep LSTM intrusion detection system. To prevent detected malicious attacks, it employs proposed snort techniques to eradicate malicious attacks. Complete malicious attacks which are taken from the detection using the deep LSTM intrusion detection system are given to the input of the snort. Whenever the snort is activated, the port of TCP, UDP of attacks is identified concerning rules generated, using snort rule engine will identify and prevent the attacks.

The detected malicious attack is given as the input of the snort, which is manually generated or given as a dataset. The snort rules are generated. The snort with the given rules whenever the type of attack is seen, the snort will send the alarm of the attacks to the cloud administrator, and it will prevent accordingly. Figure 3 shows the proposed prevention system for attack prevention and the steps discussed below:

**Packet Decoder:** It is used to capture and analyze network traffic generated during packet capture. As packets are already captured and classified, the type of attack is classified from the detection phase.

**Preprocessors:** Packets are captured from the ports of TCP and UDP. It can handle the data over multiple packets. Here, snort uses preprocessors for pattern matching.



**Fig. 3** Proposed intrusion prevention system

**Detection Engine:** It will prevent any malicious system attacks. From the rule set, detection engine will eliminate it by blocking the attack.

**Ruleset:** The rule sets are created according to the ports of TCP and UDP. The type of attack is predicted from the detection system.

**Alert Generation:** Whenever the type of attack is detected, based on the ruleset alerts are generated. From the detection engine, alerts are generated to the cloud administrator for dropping the packets.

**Drop packets:** Followed by the alert generation from the cloud admin, the detection engine will detect and display the packets that are dropped finally eliminate them.

**Cloud admin:** From the generated alert, the corresponding port attacks are displayed on the cloud admin.

## 4 Evaluation Metrics

In our model, the most important performance is accuracy, precision, recall, and f-score of intrusion detection is used to calculate the performance of the proposed deep LSTM model. Also, we discussed TP, TN, FP, FN as the predicted equals the actual, if the actual is positive and the model predicted a positive value then it is TP, if the actual is negative and the model predicted a negative value then it is TN, if the actual is negative and the model predicts a positive, then it is FP if the actual value is positive and the model predicted is negative then it is FN.

**Accuracy:** Accuracy represents the total percentage of correctly classified samples of both normal and malicious as shown in Eq. (1).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

**Precision:** PPV represents the ratio between the correctly detected malicious, and all detected malicious samples are shown in Eq. (2).

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

**Recall:** Recall represents the percentage of malicious samples that are correctly classified as shown in Eq. (3).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

**F1-score:** It represents the percentage of samples that are falsely classified as malicious as shown in Eq. (4).

**Table 2** Confusion matrix for the two category experiments on the testing set

Predicted class	actual class	Malicious	Normal
Malicious	Malicious	19,394	11
Normal	Normal	25,744	0

$$F1 - \text{score} = \frac{FP}{FP + TN} \quad (4)$$

**Confusion Matrix:** The confusion matrix is a kind of error matrix. It visualizes the prediction for a classification task. Table 2 shows the confusion matrix for the two category experiments on the testing set with predicted class and actual class of testing data.

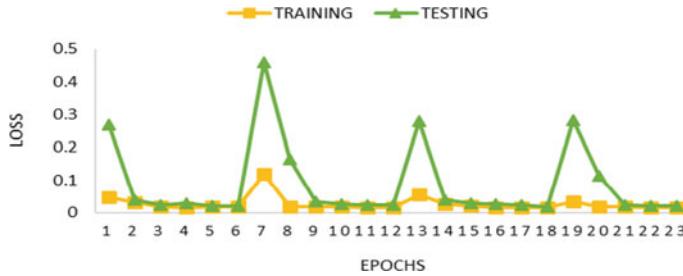
Hence, the motivation was to obtain high accuracy, precision, recall, f-score for the proposed deep LSTM IDS model. From the detected phase, it focuses on dropping packets using snort at a high rate.

## 5 Experimental Results

The experiments on lightweight deep neural networks were performed by TensorFlow and Keras and are used to implement the proposed deep LSTM intrusion detection and prevention system in the cloud environment, which is provided by IBM Watson studio. This experiment was conducted to detect TCP, UDP attacks. The dataset is collected from a tool called Wireshark, which consists of 32 features and parameters that are more than 2 lakhs, and is split into 80% of training data and 20% of test data. In the proposed deep LSTM model, which consists of more than 3 layers, 100 + neurons are used. The classification of this model involves the logistic activation function of binary classification. The detection of TCP, UDP attacks with deep learning algorithms is done and sent to a prevention system to prevent malicious packets using proposed prevention techniques to drop malicious packets with high accuracy. The accuracy of the training model with 2 epochs is shown in Figs. 4 and



**Fig. 4** Accuracy of training model of 23 epochs

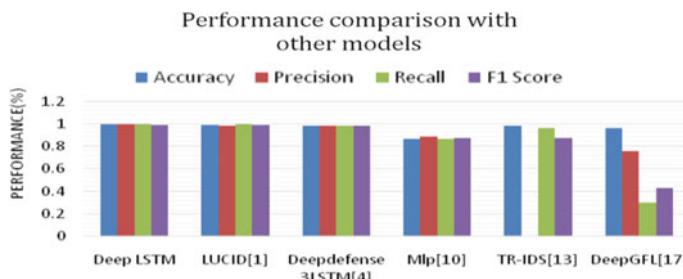


**Fig. 5** Loss of training model of 23 epochs

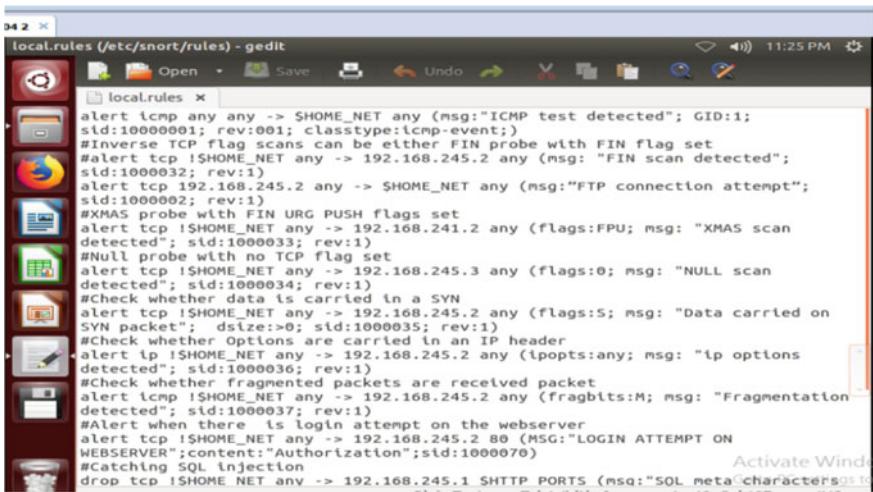
5 that shows the loss of the training model. Here we compare the existing model [4] with the state-of-the-art comparison of proposed deep LSTM IDS model attack detection. The comparisons were held with other models on metrics like ‘accuracy,’ ‘precision,’ ‘recall,’ and ‘F1 score’ as shown in Table 3, and the comparison chart is shown in Fig. 6. The local snort rules that are generated for TCP UDP attacks are shown in Fig. 7 using the rule engine for the prevention system to prevent attacks received from the detection system. Figure 8 shows the comparison chart of the proposed prevention system and compared with existing methods [3]. It shows the dropping of packets with a high accuracy rate.

**Table 3** Comparison results of the proposed deep LSTM IDS model

Model	Accuracy	Precision	Recall	F1-score
Proposed deep LSTM IDS model	0.9955	0.9926	0.9959	0.9897
Lucid [4]	0.9888	0.9827	0.9952	0.9889
Deep defense 3LSTM [5]	0.9841	0.9834	0.9847	0.9840
Mlp [14]	0.8634	0.8847	0.8625	0.8735
TR-IDS [15]	0.9809	0.0040	0.9593	0.8742
DeepGFL [15]	0.9624	0.7567	0.3024	0.4321



**Fig. 6** Performance comparison of the proposed deep LSTM IDS model

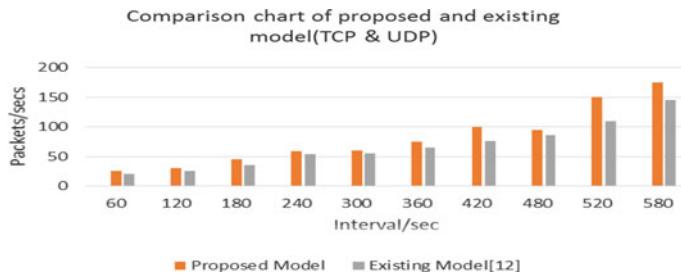


```

local.rules (/etc/snort/rules) - gedit
local.rules x
alert icmp any any -> $HOME_NET any (msg:"ICMP test detected"; GID:1;
sid:10000001; rev:001; classtype:icmp-event;
#Inverse TCP flag scans can be either FIN probe with FIN flag set
#alert tcp !$HOME_NET any -> 192.168.245.2 any (msg: "FIN scan detected";
sid:1000032; rev:1)
alert tcp 192.168.245.2 any -> $HOME_NET any (msg:"FTP connection attempt";
sid:1000002; rev:1)
#XMAS probe with FIN URG PUSH flags set
alert tcp !$HOME_NET any -> 192.168.241.2 any (flags:FPU; msg: "XMAS scan
detected"; sid:1000033; rev:1)
#Null probe with no TCP flag set
alert tcp !$HOME_NET any -> 192.168.245.3 any (flags:0; msg: "NULL scan
detected"; sid:1000034; rev:1)
#Check whether data is carried in a SYN
alert tcp !$HOME_NET any -> 192.168.245.2 any (flags:S; msg: "Data carried on
SYN packet"; dsiz:>0; sid:1000035; rev:1)
#Check whether Options are carried in an IP header
alert ip !$HOME_NET any -> 192.168.245.2 any (ipopts:any; msg: "ip options
detected"; sid:1000036; rev:1)
#Check whether fragmented packets are received packet
alert icmp !$HOME_NET any -> 192.168.245.2 any (fragbits:M; msg: "Fragmentation
detected"; sid:1000037; rev:1)
#Alert when there is login attempt on the webserver
alert tcp !$HOME_NET any -> 192.168.245.2 80 (MSG:"LOGIN ATTEMPT ON
WEBSERVER";content:"Authorization";sid:1000070)
#Catching SQL injection
drop tcp !$HOME_NET any -> 192.168.245.1 SHTTP_PORTS (msg:"SQL meta characters
detected");

```

**Fig. 7** Snort local rules for malicious attacks



**Fig. 8** Comparison chart of proposed prevention system [3]

## 6 Conclusion and Future Enhancement

Cloud computing in many sectors is becoming vast, as it uses to improve security in many aspects. In this paper, the incoming packets are classified to detect the behavior of the source whether the attack is malicious or normal. The results show that the novel deep LSTM-based intrusion detection system can accurately detect attacks. The detected attacks are passed to the proposed prevention mechanism using snort. The malicious packets are dropped with high accuracy. The overall proposed detection and prevention system have an accuracy of about 99% on the last iteration of best-chosen hyperparameter values under different attacks. The performance is validated and tested and is shown to be accurate. Thus, the proposed approach can efficiently improve accuracy using a security of data and will reduce the bandwidth usage and cut the over usage of resources. In the future, various types of DDoS attacks detection and prevention frameworks can be a study point along with deep learning algorithms.

## References

1. Gupta, A., Sharma, L.S.: Mitigation of DoS and port scan attacks using snort. *Int. J. Comput. Sci. Eng.* **7**, 248–258 (2019)
2. Patel, A., Taghavi, M., Bakhtiyari, K., Junior, J.C.: Review: an intrusion detection and prevention system in cloud computing: a systematic review. *J. Netw. Comput. Appl.* **36**, 25–41 (2013)
3. Min, E., Long, J., Liu, Q., Cui, J., Chen, W.: TR-IDS: anomaly-based intrusion detection through text-convolutional neural network random forest. *Security and Communication Networks* (2018)
4. Doriguzzi Corin, R., Milla, S., Scott Hayward, S., Martnez Del Rincon, J., Siracusa ICT, D., Fondazione Bruno Kessler.: LUCID: a practical, lightweight deep learning solution for DDoS attack detection. *IEEE Trans. Netw. Serv. Manage.*, **17**, 876–888 (2020)
5. Yuan, X., Li, C., Li, X.: Large-scale intelligent systems laboratory, deep defense: identifying DDoS attack via deep learning. *Smart Comp.* **1**, 1–8 (2017)
6. Saxena, R., Dey, S.: DDoS attack prevention using collaborative approach for cloud computing. *Cluster Comput.* **23**, 1329–1344 (2020)
7. Wu, K., Chen, Z., Li, W.: A novel intrusion detection model for a massive network using convolutional neural networks. *School Control Comput. Eng.* **6**, 50850–50857 (2018)
8. Kasongo, S.M., Sun, Y.: A deep learning method with filter based feature engineering for wireless intrusion detection system. *IEEE Access* **7**, 38597–38607 (2019)
9. Sahi, A., Lai, D., Li, Y., Dikyh, M.: An efficient DDoS TCP flood attack detection and prevention system in a cloud environment. *Inf. Comput. Sci.* **5**, 6036–6048 (2017)
10. Roopak, M., Yun Tian G., Chambers, J.: Deep learning models for cyber security in IoT networks. In: IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC) (2019)
11. Gupta, B., Badve, O.P.: Taxonomy of DoS and DDoS attacks and desirable defense mechanism in a cloud computing environment. *Neural Comput. Appl.* **28**, 3655–3682 (2017)
12. Yao, Y., Su, L., Lu, Z.: Deep GFL: deep feature learning via graph for attack detection on flow-based network traffic. In: Proc. of IEEE Military Communications Conference (MILCOM) (2018)
13. Aneetha, A.S., Bose, S.: The combined approach for anomaly detection using neural networks and clustering techniques. *Comput. Sci. Eng.* **2**, 37–46 (2012)
14. Subba, B., Biswas, S., Karmakar.: A neural network based system for intrusion detection and attack classification. In: Twenty Second National Conference on Communication (NCC) (2017)
15. Shah, S. A. R., Isaac, B.: Performance comparison of Intrusion detection systems and application of machine learning to snort system. *Future Gener. Comput. Syst.*, Elsevier, **80**, 157–170 (2018)

# Smart Energy Metre Based on IoT



K. Rubitha, J. Jecintha, K. Shifana Begum, and S. Suriya

**Abstract** Energy metre reading is a tiresome and a high-priced issue. The metre reader has to go and obtain the reading manually to provide the bill, which will later be entered in the software to computerize the billing and payment method. This paper proposes an innovative technology for energy metre reading by Internet of Things technology and web server beside the existing metres. An Internet of Things modem will be integrated with electronic energy metre to read the parameter such as voltage, current, frequency and uploaded on server or website. This device uses an infrared sensor to measure recent usage. The IR transmitter is situated in the spinning unit of the EB metre. To determine the number of rotations, the receiver photo diode must be installed in a secure location. With the number of rotations known, we can calculate current consumption. The Atmel will reduce the unit offered to a certain user after identifying the present consumption. The unit is a numerical value in this case. If the component is lowered to its minimal value, the user will be warned via an alarm and an LCD display. The user can use the EB section to request additional unit usage. The EB section can set the unit consumption limit in the Atmel controller using the IoT module. As a result, the recharge process is completed fast and with fewer manual interactions. The controllers utilized in prior articles did not have an ADC interface, necessitating the inclusion of additional peripherals. However, in this paper, an ATmega8 controller with an integrated ADC is used; hence, the overall speed can be increased. Our approach could be useful in a variety of applications, including industrial control, medical systems and access control.

**Keywords** Smart metre · IoT · Web page · Electric board · Arduino

---

K. Rubitha (✉) · J. Jecintha · K. Shifana Begum · S. Suriya  
Department of ECE, CARE College of Engineering, Trichy, Tamil Nadu, India  
e-mail: [rubitha.kannan@gmail.com](mailto:rubitha.kannan@gmail.com)

## 1 Introduction

### 1.1 *Introduction to IoT*

The Internet of Things (IoT) allows objects to be sensed or controlled remotely over existing network infrastructure, allowing for more direct integration of the physical world into computer-based systems and, as a result, improved efficiency, accuracy and economic benefit, as well as less human intervention. When IoT is combined with sensors and actuators, it is classified as a cyber-physical system, which includes smart grids, virtual power plants, smart homes and smart cities. Each object is uniquely identified by its embedded computing system, but it can communicate with other things on the Internet.

People also wish to use the Internet to communicate with non-living objects such as home appliances, furniture, stationery and clothing. People already have a variety of tools for interacting with living things, but the Internet of Things allows them to communicate with non-living objects in a more comfortable manner. The Internet of Things (IoT) is a collection of technologies such as ubiquitous, pervasive computing, ambient intelligence, sensors, actuators, communications technologies, Internet technologies and embedded systems, among others.

A vast network that supports IoT devices and applications is known as an IoT cloud. This comprises the infrastructure, servers and storage required for real-time processing and operations. For enterprises with limited resources, IoT clouds provide an efficient, flexible and scalable strategy for supplying the infrastructure and services required to enable IoT devices and applications. IoT clouds provide on-demand, cost-effective hyperscale, allowing businesses to take advantage of the IoT's huge potential without having to construct the underlying infrastructure and services from the ground up.

### 1.2 *Overview*

Electricity is the driving force behind any country's progress. Due to the increase in residential, commercial and industrial energy users around the world, it is now critical for utilities firms to create improved, ecologically friendly methods of evaluating utility consumption in order to produce and invoice an exact bill.

To computerize metre readings and upload them to a web server, the Internet of Things (IoT) model is used. Current project focuses on the connection and system element of the IoT in this epoch of smart city development. This project uses the Arduino Uno in the embedded system section to create and implement an energy consumption calculation based on the counting of calibration pulses. The ATmega8 is an 8-bit CMOS microcontroller based on the AVR RISC architecture that consumes very little power. The ATmega8 delivers throughputs approaching 1 MIPS per MHz by executing strong instructions in a single clock cycle, allowing the system designer

to balance power consumption and processing speed. The Internet of Things-based metre reading system is meant to continuously monitor the metre reading in the planned work. When a consumer fails to pay their monthly bill, the service provider can disconnect the power source. It also eliminates human involvement, provides accurate metre reading, and allows for prepayment.

This paper proposes a concept for a last-meter smart grid that is incorporated in a stage for the Internet of Things in this paper (IOT). It offers several advantages, like customer centricity, scalability and so on. Smart grid and smart home apps are seamlessly integrated. First, assume that the typical early adopter of a last-meter smart grid is also a smart home application user (dedicated to security, entertainment, home automation etc.). The stage must support both smart grid and other smart home applications to reduce duplication and enable possible synergy. It can gather data from a variety of sensor communication methods. The last-meter smart grid makes use of readily available communications.

## 2 Literature Survey

In this section, some recent works and studies are reviewed.

Devadhanishini et al. [1], smart power management energy consumption is a critical and difficult issue when using IoT. In big electric energy distribution systems, an automatic electrical energy metre is employed. The system functions as a smart power monitoring system thanks to the integration of the Arduino Wi-Fi and SMS. The data from a smart energy metre can be used to optimize and reduce power use. This system also contains a motion sensor, which will automatically turn OFF the power supply if there is no human in the house.

Mugunthan and Vijayakumar [2], the network's power usage, distribution, transmission and generating functions have all improved. The system's applications include power equipment installation surveillance and the application of dynamic scheduling for adjusting home use, electric vehicle parking and charging, management of power supply and demand, power supply equipment maintenance, failure and fault detection and so forth.

Subba Rao and Sri Vidya Garage [3], furthermore, developed a novel architecture for continually measuring and controlling electric metres in residential areas. Finally, they built a new technology that could be operated from afar. Using the GSM idea, this proposed architecture may transfer data to the server and send SMS. This framework was created with the help of an ARM processor controller and a few sensor sets.

Shaista Hassan Mir et al. [4], despite this, it created a sophisticated electric metre with the Arduino microcontroller and GSM technology. The electric bill is generated automatically by this smart metre, and it is sent to the consumer through GSM modem.

Patel et al. [5], however, developed and executed a new technology that eliminates the need for personal intervention in reading electric metres and creating energy

bills. The main advantage of this proposed system was the elimination of corruption in electric power usage and the generation of an electric bill. This system is built with a GSM method, an Arduino controller and an LDR sensor module with relay functionality. The LDR sensor is used in conjunction with the LED light on the metre box tool to communicate data to the microcontroller via the GSM module. This technology is also utilized to deliver SMS to the people who have expressed interest.

Mohammed Hosseiu et al. [6], design and implementation of smart metre using IoT, a study discussing the growth of IoT and digital technologies, were presented. The future energy grid must be built on a distributed architecture capable of dynamically absorbing various energy sources. The Internet of Things (IoT) can be used for a variety of smart grid applications, including power consumption, smart metres, electric power demand side management and many areas of energy production. The basic objective of Smart Energy Metering (SEM) is to collect information on energy consumption of household appliances, monitor environmental parameters and deliver the appropriate services to home users, as detailed in this paper.

Himanshu K Patel et al. [7] presented an Arduino-based smart energy metre that eliminates human participation in metre readings and bill creation, lowering the mistake rate that is common in India. The system has the ability to send an SMS to a user for an update on energy use, as well as the ability to generate a final bill and reload through SMS. A relay was used to disconnect the power supply on demand or owing to past due bills. For bidirectional communication, the system uses GSM.

Kumar et al. [8], according to this study, smart electric metres have the potential to improve energy efficiency. However, comparing the suggested system's installation to an existing traditional system was extremely challenging.

Al-Ali et al. [9], the LDR sensor was used to measure the LED blinking frequencies in this smart metre. The number of LED blinks is normally proportionate to the amount of normal power stored in the typical metre gadget. The data collected by the LED blinker will be saved on the web server for future use.

Geetha et al. [10], the went immediately to the users' location and recorded the metre readings in the existing system. The power bill was prepared and provided to the user based on the unit's value. The authors proposed a new system that uses an Arduino controller to automatically read metre readings. This solution allowed users to use the Wi-Fi module to check their energy use and bill amount. This technique will save human energy while also preventing machine repair.

Bibek Kanti Barman et al. [11], the proposed smart metre with IoT for effective energy use is critical for the development of the smart grid in the power system. As a result, appropriate power consumption monitoring and control are a top priority for the smart grid. The energy metre has a number of issues, one of which is the lack of full duplex communication. To address this issue, a smart energy metre based on the Internet of Things is proposed. The smart energy metre uses an ESP 8266 12E Wi-Fi module to manage and compute energy use and sends the data to the cloud, where the consumer or customer can view the results. As a result, the consumer's energy examination has gotten far more thorough.

### 3 Existing System

Home automation can be described as a system which controls all the home appliances using Zigbee and some other sensors. In home automation, it is emphasized that energy is preserved and security is maintained. Most systems exchange data and communicate with the help of Zigbee and GSM. For short distance communication, Zigbee module is only used. Every entity's electronic energy metre contains a GSM-based wireless communication component that allows for remote monitoring of electricity usage. The exchange of information is fast, secured and accurate with the help of wireless communication media.

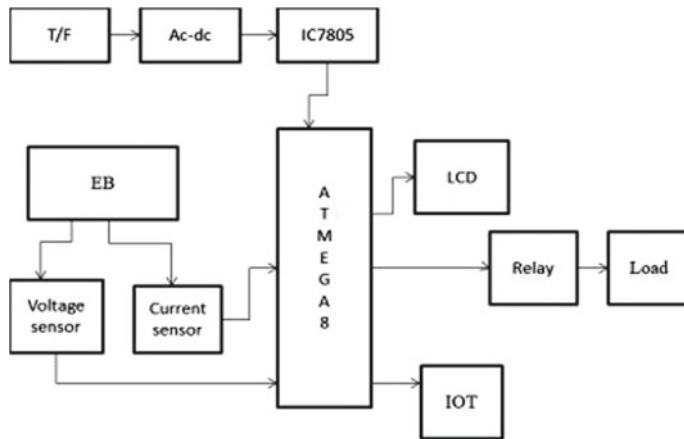
#### 3.1 *Disadvantages*

- It is short-range monitoring, low efficiency and has low data speed
- Cannot monitor every individual load in the power system
- If the load becomes heavier, the user might not get any knowledge of it. The user will not receive any alert messages
- Any expense will necessitate the use of manpower.

### 4 Proposed Model

This proposed model offers a detailed architecture and implementation of a “last-meter” smart grid in our work. The portion of the smart grid is incorporated in an Internet of Things (IOT) platform on a customer’s premises. Power theft monitoring is an essential study topic in the electric power system, as electricity theft prevention has become a major issue. Electricity theft is a long-term issue; nevertheless, each power supply department has made significant human and material investments, and the phenomena of resisting electricity theft are very important and not subsided, and the method of electricity theft is constantly improved. Electricity theft not only costs the power industry a lot of money, but it also jeopardizes the security and dependability of the main power supply.

In Fig. 1, microchip’s ATmega8 is a 28-pin, 8-bit AVR microcontroller with RISC architecture that is mostly utilized in embedded systems and industrial automation projects. Transformer, rectifier and voltage regulator make up the power supply. An electrical transformer is a device that converts alternating electricity from one voltage to another. An oscillating two-directional alternating current is converted into a single-directional direct current via a rectifier (Dc). The IC7805 is a 5 V voltage regulator that limits the output voltage to 5 V for a variety of input voltage ranges. The Arduino microprocessor receives input voltage. In load, there is a current sensor and a voltage sensor. The amount of volts is calculated and monitored using a voltage sensor. It has the ability to determine both AC and DC voltage. Current sensors, also



**Fig. 1** Functional block diagram

known as current transformers or CTs, are devices that use the magnetic field to detect and provide a proportional output to measure the current flowing through a wire that work with both AC and DC power. The output can be shown on an LCD, loaded via a relay, or sent over IoT. Relays are electromechanical switches that open and close circuits. The load here consumes electrical energy in the form of current and converts it into light. The ability of the Internet of Things is to transport data over a network without human -to-human or human-to-computer interaction.

#### 4.1 Advantages

- It uses less energy and is highly efficient
- The EB metre's voltage and current are measured using a voltage and current sensor
- User will be able to prevent energy waste and will be able to conserve energy
- Taking current readings does not necessitate the use of manpower
- To reduce the amount of electricity used
- Can quickly identify industries that use a lot of electricity.

## 5 Implementation of the Project

### 5.1 Software Implementation and Result

The proteus toolbox uses mixed mode SPICE circuit simulation, animated components and microprocessor models to make co-simulation of full microcontroller-based systems easier. Before constructing a real prototype, these concepts can be produced and tested. When it comes to mixed mode SPICE circuit simulation, the Proteus Design Suite (PDS) can assist by co-simulating both the high- and low-level microcontroller code. The PCB editing software included in the PDS is fully compatible with ISIS schematic capture.

#### Working Procedure of the ISIS Proteus

**Step 1:** Launch the ISIS professional application by double-clicking the icon on your desktop; the splash screen will appear.

**Step 2:** After that, a work area with interface buttons for circuit design will emerge. The third line in the room is a blue rectangle line; ensure that the entire circuit is created within the rectangular space.

**Step 3:** From the library, choose an apparatus. Select device/symbol from the menu bar library. Then, as illustrated below, a new window will appear, allowing you to pick the components in a different way. A tool bar is located on the left side of the workspace. Click the component mode button or choose from the library in that tool bar.

**Step 4:** Select all apparatus from the library and add them to the devices list. By clicking on the gadget and rotating it, you can adjust its angle. The component is then placed in the workspace after clicking in the workspace.

**Step 5:** Arrange all of the devices in the work area, then position the cursor at the component pin end and draw the connections with the pen symbol. Connect all of the components according to the circuit, as shown in the diagram below.

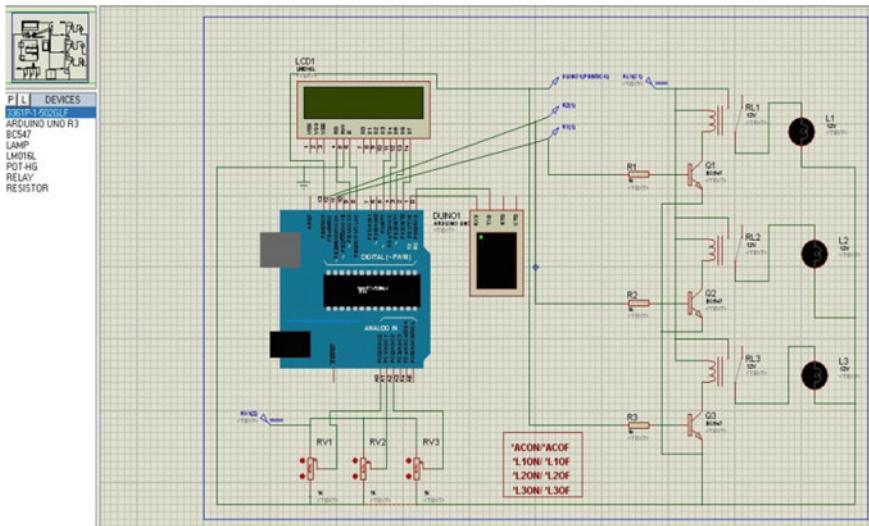
**Step 6:** If any changes to the component are required, place the mouse pointer over the right button and the option window will unfold.

**Step 7:** The components are wired together, or it can be connected with a wire label by attaching a generic wire label.

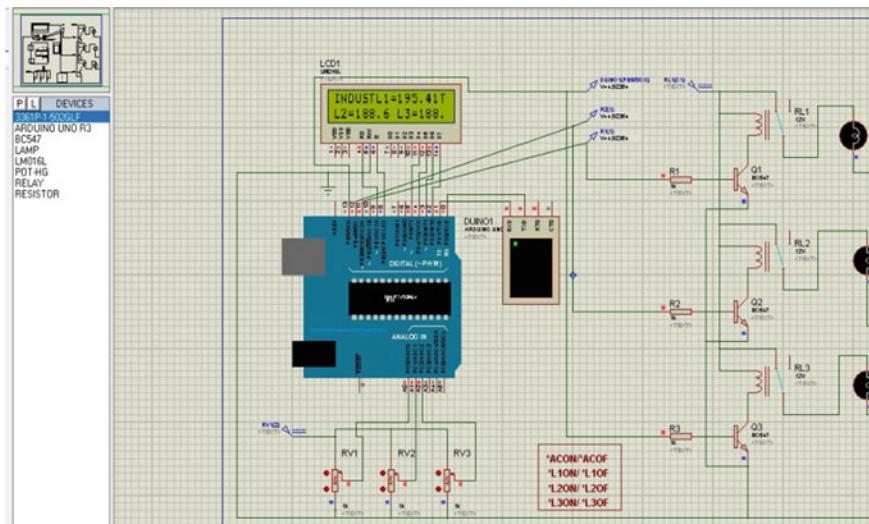
Figure 2 shows the schematic circuit diagram where the replacement of voltage sensor with variable resistors has taken place.

Figure 3 shows the output screen when all the three loads are ON, and the units consumed are displayed for each individual load. The virtual terminal screen is shown in Fig. 7.3, which shows the commands that are given to the users. Figure 4 shows the output screen of virtual terminal with command.

Figure 5 shows the output screen when all the 3 loads are ON, and the units consumed are displayed for each individual load. Based on this, the consumption is reduced individually.



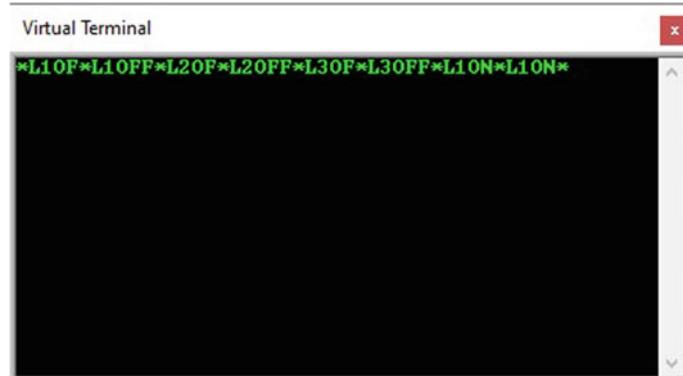
**Fig. 2** Circuit which shows simulation



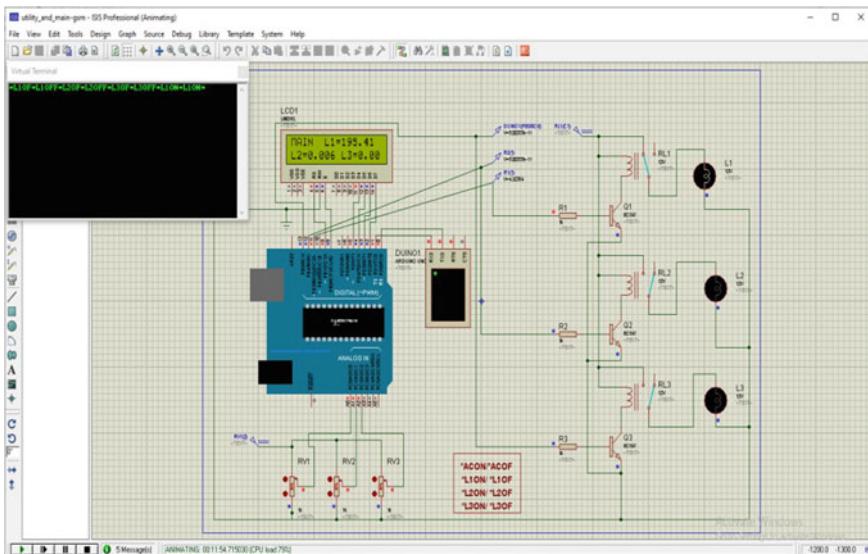
**Fig. 3** Circuit when all loads are ON

## 5.2 Hardware Implementation and Result

The proposed system is explained above. Figures 6 and 7 show the hardware connections of smart energy metre using IoT. In this, the power supply circuit consists of transformer to convert the 230 V into 5 V followed by bridge rectifier to minimize the

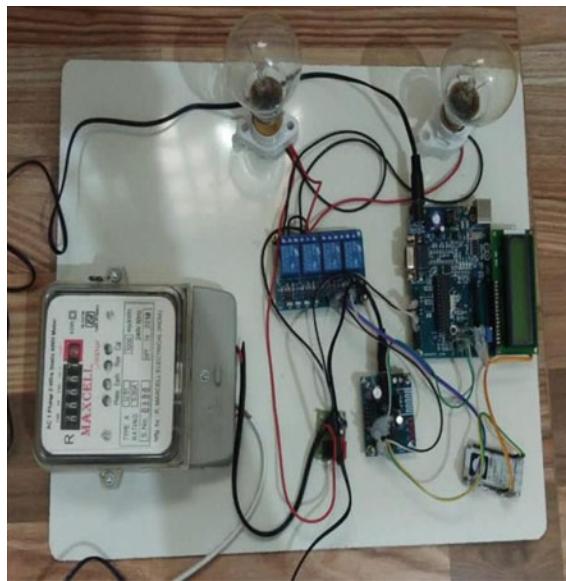


**Fig. 4** Virtual terminal with command

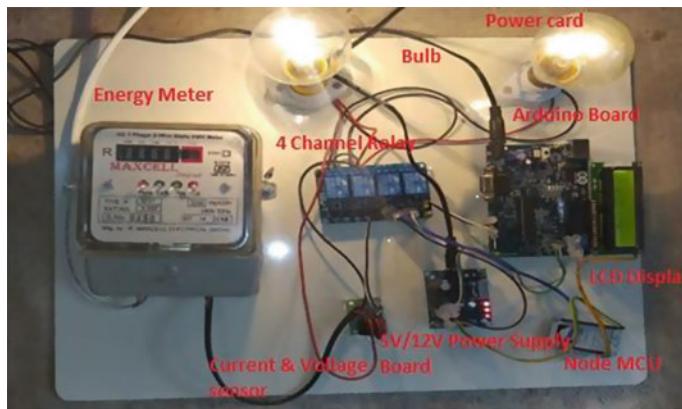


**Fig. 5** Circuit when load 1 is ON

AC current, and the capacitor is used to filter the present AC current (filtering circuit), resistors, led and switch. IC7805 to get the accurate 5 voltage in an output (voltage regulator that restricts the output voltage = 5v). The connection of ATmega328p microcontroller IC followed by relay, liquid crystal display is established. And, whenever the household appliances consume energy, the energy metre continuously reads the reading, and the consumed load is visible on the metre. The energy metre contains a constant blinking LED, and the Arduino microcontroller measures the units consumed based on the flashing of the LED on the energy metre. As per the blinking of LED when the LED blinks 3200 times, it will consider as one unit. The



**Fig. 6** Hardware connections without power supply



**Fig. 7** Hardware connections after power supply

consumed units, along with the time and the date of the consumed units (bill), are continuously displayed on web page using IoT. The consumed units will be displayed on LCD also.

**Step 1:** After completing connections, switch ON the power supply, you can see the load, switched ON.

**Step 2:** After that, monitor the parameters of these loads in our HTML web page. The parameters are monitored over the IoT.

Sno	Status	Date / Time
1	L1:9 L2:15	05-03-2021 08:38:08 AM
2	L1:0 L2:16	05-03-2021 08:45:43 AM
3	L1:0 L2:0	05-03-2021 08:46:40 AM
4	L1:4 L2:3	05-03-2021 08:47:52 AM
5	L1:4 L2:4	05-03-2021 08:49:08 AM
6	L1:4 L2:4	05-03-2021 08:50:25 AM

**Fig. 8** Image showing units consumed with date and time

**Step 3:** Open the Google Chrome and search my project final.in/demo\_home. The display page will be shown.

**Step 4:** Then enter the username as C17 and enter the password as admin and click on submit.

**Step 5:** This page will display the home page and display L1 and L2 switches. On clicking light1 ON, load1 is glowing.

On clicking light1 OFF, load1 is not glowing.

On clicking light2 ON, load2 is glowing.

On clicking light2 OFF, load2 is not glowing.

**Step 6:** On clicking the information at the top of the page, it shows the status of the unit consumed by each load separately, with date and time.

Figure 8 shows the units consumed by the home appliances with date and time.

## 6 Conclusion

Simple devices were connected to the Internet of Things, and the apparatuses were operated remotely over the web, demonstrating that home automation using the Internet of Things works well. The constructed framework not only monitors sensor data, such as temperature, light, gas and movement sensors, but also initiates a procedure in response to the requirement, such as turning ON the light when it gets dark. It also keeps the sensor parameters in a secure manner in the cloud (Gmail). This will assist the client in deciphering the state of several parameters in the home at any time and from any location.

## References

1. Devadhanishini, et al.: “Smart power monitoring using IoT”. In: 5th International Conference on Advanced Computing and Communication Systems (ICACCS) (2019)
2. Mugunthan, S., Vijayakumar, T.: Review on IoT based smart grid architecture implementations. *J. Electric. Eng. Autom.* **10**(1), 12–20 (2019)
3. Subba Rao, A., Garage, S. V.: “IOT Based smart energy meter billing monitoring and controlling the loads”. *Int. J. Innovative Technol. Explor. Eng. (IJITEE)*, ISSN: 2278–3075., **8**(4), 340–344 (2019)
4. Mir, S. H., Ashruf, S., Sameena, Bhat, Y., Beigh, N.: “Review on smart electric metering system based on GSM/IOT”. *Asian J. Electric. Sci.*, ISSN:2249–6297, **8**(1), 1–6 (2019)
5. Patel, H. K., Mody, T., Goyal, A.: “Arduino based smart energy meter using gsm”. In: 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU) (2019)
6. Yaghmaee, M. H.: “Design and implementation of an internet of things based smart energy metering”. In: 6th IEEE International Conference on Smart Energy Grid Engineering (2018)
7. Patel, H. K.: “arduino based smart energy meter”. In: 2nd Int'l Conf. on Electrical Engineering and Information and Communication Technology (ICEEICT) (2018)
8. Kumar, A., Thakur, S., Bhattacharjee, P.: “Real-time monitoring of AMR Enabled energy meter for AMI in smart city—an IoT application”. In: IEEE International Symposium on Smart Electronic Systems (iSES), pp. 219–222 (2018)
9. Al-Ali, A. R., Landolsi, T., Hassan, M. H., Ezzeddine, M., Abdelsalam, M., Baseet, M.: “An IoT—based smart utility meter”. In: 2018 2nd International Conference on Smart Grid and Smart Cities (ICSGSC) (2018)
10. Geetha, R., Abhishek, D., Rajalakshmi, G.: “Smart energy meter using IoT”. *Int. J. Recent Trends. Eng. Res. (IJRTER) Conf. Electron. Inf. Commun. Syst. (CELICS\_18)*, ISSN: 2455–1457, pp. 235–239 (2018)
11. Barman, B. K., et al.: “proposed paper smart meter using IoT”. In: Department of International Electronics And Electrical Engineering (IEEE) (2017)

# Author Index

## A

- Abdulhussain, Sadiq H., 657  
Abhay, P. A., 909  
Abhishek Rao, N., 493  
Abisha, D., 969  
Adilakshmi, T., 517  
Agarwal, Basant, 677  
Agrawal, Vidhi, 463  
Ajish, S., 791  
Akilandeswari, A., 961  
Akshayanjali, S. A., 83  
AlGhamdi, Sarah, 853  
Alghawas, Sadiqa, 853  
Alvi, A. S., 121  
Amirtharaj, Joyce R., 155  
Anil Kumar, K. S., 791  
Anithaashri, T. P., 215  
Anuradha, T., 551  
Aparna, C., 565  
Ariful Islam, Md., 479  
Ar-Reyouchi, El Miloud, 327, 931  
Aruna, Orchu, 67  
Ashok, Akshay, 93

## B

- Babakura, Abba, 421  
Babu, V. Suresh, 609  
Baiju, M. R., 609  
Benchib, Imane, 327  
Benjamin, Minu Inba Shanthini Watson,  
753  
Beula, R. Janefer, 699  
Bhat, Arpit, 295  
Bhatt, Pooja, 17

- Bodaghi, Amirhosein, 761  
Bodaghi, Hossein, 197  
Boucif, Mounia, 339  
Boyd Wesley, A., 617  
Bui, Aminu, 421  
Burdak, Ashish, 677

## C

- Caviedes, Juan C., 863  
Chakravarthy, N. S. Kalyan, 753  
Chandana, G., 505  
Charulatha, A. R., 383  
Chellasamy, Aarthy, 397  
Chethana, H. T., 349  
Chinchulkar, Chetan, 463

## D

- Daas, Mourad, 339  
Dabhade, Vaibhav, 121  
Dafni Rose, J., 827  
Dash, Manoranjan, 551  
Davila, Diego, 257  
Devaki, K., 49  
Devaraju, R., 223  
Devareddi, Ravibabu, 879  
Devi, B. Sudha, 739  
Dhanalakshmi, G., 39  
Dheeraj, G., 493  
Diaz, Azarquiel, 451  
Diller, Jonathan, 643

**E**

Elakiya, E., [307](#)

**F**

Felix Enigo, V. S., [837](#)  
 Fernando, Xavier, [633, 643](#)

**G**

Gamot, Amit, [463](#)  
 Gaur, Aakanksha, [1](#)  
 Gayathri, R., [1](#)  
 Ghoumid, Kamal, [327, 931](#)  
 Gnanadurai, Immanuel, [295](#)  
 Goddu, Jyothi, [317](#)  
 Goel, Lipika, [111](#)  
 González, Hernando, [451](#)  
 Gosaliya, Jaina S., [93](#)  
 Govinda, K., [493](#)  
 Govindaraj, Logeswari, [981](#)  
 Gugapriya, G., [961](#)  
 Gupta, Adarsh K., [93](#)  
 Gupta, Ruchika, [185](#)  
 Gupta, Satvik, [677](#)  
 Gupta, Sonam, [111](#)

**H**

Hadzhikoleva, Stanka, [575](#)  
 Hadzhikolev, Emil, [575](#)  
 Harjule, Priyanka, [677](#)  
 Hemavathi, S., [551](#)  
 Hoong, Yang, [285](#)  
 Hossain, Tanjil, [479](#)  
 Hung, Bui Thanh, [897](#)  
 Hussin, Siti Maherah, [657](#)

**I**

Illanko, Kandasamy, [633](#)  
 Isobe, Takashi, [593](#)  
 Iswarya, R., [837](#)

**J**

Jabasheela, L., [39](#)  
 Jaimes, Luis, [451](#)  
 Jecintha, J., [995](#)  
 Jitpattanakul, Anuchit, [531](#)

**K**

Kalaichelvi, T., [39](#)

Kalhor, Ahmad, [197](#)

Kanagaraj, K., [155](#)

Kanagaraj, R., [307](#)

Karanam, Santoshachandra Rao, [233](#)

Kavitha, T., [543](#)

Kayande, Shaunak, [463](#)

Khan, Alif Bin Rahman, [479](#)

Kirubakaran, M. K., [827](#)

Kolhe, Anurag, [945](#)

Kranthi Kumar, K., [317](#)

Kulkarni, Jay, [945](#)

Kulkarni, Vaishnavi, [945](#)

Kushal Kumar, S. G., [493](#)

**L**

Lakshmi Akshitha, Y., [505](#)  
 Lakshmi Divya, J. K., [837](#)  
 Lakshmi Kanthan, N., [809](#)  
 Lalitha, R. Hannah, [969](#)  
 Lasya Priya, K., [505](#)  
 Leena Jenifer, L., [49](#)  
 Lotlikar, Trupti, [295](#)

**M**

Maduranga, M. W. P., [29](#)  
 Mahesh, Gadiraju, [879](#)  
 Mallikarjun B.C., [437](#)  
 Manikandan, N. K., [505](#)  
 Mani, Subalakshmi, [981](#)  
 Manivannan, D., [505](#)  
 Manoj Sithara, J. P. D., [29](#)  
 Marali, Mounesh, [269](#)  
 Mekruksavanich, Sakorn, [531](#)  
 Meza, Carlos, [451](#)  
 Michael, Jee Joe, [753](#)  
 Misbha, D. S., [739](#)  
 Mishra, Sushil Kumar, [185](#)  
 MNSSVKR Gupta, V., [879](#)  
 Mohammadi, Seyed Omid, [197](#)  
 Mohammad, Nazeeruddin, [853](#)  
 Motwani, Kashish, [463](#)  
 Muñoz, Wilmar Yesid Campo, [863](#)  
 Murari, Thejovathi, [233](#)

**N**

Nagarathinam, Aishwarya, [397](#)  
 Nagavi, Trisiladevi C., [349](#)  
 Naidu, Allu Swamy, [367](#)  
 Naik, Archana, [1](#)  
 Nalajala, Paparao, [551](#)  
 Nandhini, S., [565](#)

Narendra, Rajashree, 137, 223  
 Navaneethakrishnan, R., 753  
 Niveditha, C. B., 83

**O**

Okada, Yoshihiro, 593  
 Omran, Alaa Hamza, 657

**P**

Padmakala, S., 961  
 Parikh, Swapnil M., 93  
 Patel, Dharmesh, 17  
 Patnala, Tulasi Radhika, 233  
 Pavithra, M., 543  
 Phalanetra, H. S., 437  
 Philip, Abin Oommen, 909  
 Pinto, Dion, 295  
 Polara, Vishal, 17  
 Prakash, V. R., 969  
 Praveen, Nida, 111

**R**

Raghu, D., 245  
 Rajkumar, N., 307  
 Ramaprabha, P. S., 39  
 Ramya Barathi, K., 155  
 Rao, Likki Venkata Krishna, 317  
 Rashini, H., 39  
 Rathod, Ketan, 17  
 Rattal, Salma, 327, 931  
 Ravikumar, M., 551  
 Reddy, Lakshmikiran, 269  
 Reddy, Reddigari Keerthi, 83  
 Resmi, R., 609  
 Revathy, G., 775  
 Rezania, Davar, 285  
 Roko, Abubakar, 421  
 Rubitha, K., 995  
 Rueda, Diego F., 863

**S**

Sadekur Rahman, Md., 479  
 Said, Dalila Mat, 657  
 Saidu, Ibrahim, 421  
 Saini, Shashikant, 1  
 Sakthi, U., 827  
 Saleema, J. S., 717  
 Sam, Dahlia, 827  
 Santosh Kumar, R., 223  
 Saoud, Bilal, 339

Sarasa-Cabezuelo, Antonio, 409

Saravanaguru, RA. K., 909

Sasikala, C., 775

Senthilrajan, A., 317

Sfiligoi, Igor, 257

Sharma, Amit, 67

Shekhar, Himanshu, 961

Shifana Begum, K., 995

Shiva Shankar, R., 879

Shobana, M., 39

Shyamala Devi, M., 493, 505

Sindhu Sai, A., 83

Singh, Ajeet, 367

Singh, Krishna Kumar, 171

Singh, Saurav, 677

Singh, Shivdeep, 677

Sivanandakumar, D., 775

Siva Prasad, B. V. V., 233

Siva Prasad, P., 317

Soleymanzadeh, Raha, 633

Sooda, Kavitha, 1

Sreedevi, B., 775

Sreenivasulu, G., 215

Srikanteswara, Ramya, 83

Srinivasan, K., 307

Srivastava, Priyanka, 171

Srivastava, Stuti, 1

Sucharitha, G., 233

Suhasini, A., 809

Sundan, Bose, 981

Sunitha, M., 517

Suriya, S., 995

Swaminathan, J. N., 753

Syam, Baddeti, 753

**T**

Tagare, Trupti Shripad, 137

Tentu, Appala Naidu, 367

Thangasamy, Anitha, 981

Theresa, W. Gracy, 39

Thomas, Aju Mathew, 269

Tigadi, Arun, 551

Tripathy, Hrudaya Kumar, 245

**U**

Unissa, Mehar, 517

**V**

Varghese, Jithy, 717

Venkatasubramanian, S., 809

Venkatasubramanian, Vaishnavi, 565

Venkatesan, K. G. S., 233

Vensila, C., 617

Victoria Priscilla, C., 383

Vijayalakshmi, G., 505

Vijaya, N., 775

Würthwein, Frank, 257

## Y

Yotov, Kostadin, 575

Yusuf, Mahmud Ahmad, 421

## W

Wankhede, Disha Sushant, 463

Wesley, A. Boyed, 699

Weslin, D., 969

## Z

Zhu, Jonathan J. H., 761